

Disclaimer

This document contains a preliminary version of a module of the Handbook on Methodology for Modern Business Statistics (Memobust). The contents of the module have been reviewed and accepted within the Memobust project. For this preliminary version, the integration with the rest of the handbook (cross-references to other modules, co-ordination of terminology, etc.) has not been thoroughly checked.



Statistics Netherlands

Process Development, IT, and Methodology
Methodology The Hague

*P.O.Box 24500
2490 HA Den Haag
The Netherlands*

Introduction to the Memobust handbook

Leon Willenborg, Sander Scholtus, Rob van de Laar

Remarks:

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

Project number:

Memobust-2

BPA number:

Date:

6 May 2013

INTRODUCTION TO THE MEMOBUST HANDBOOK

1. About the handbook

1.1 Introduction

This is the introduction to the *Handbook of Methodology for Modern Business Statistics* (henceforth: the Memobust handbook). This handbook was developed between January 2011 and March 2014 as the main output of an ESSnet project called Memobust. The Memobust project was primarily financed by Eurostat and it involved the national statistical institutes (NSIs) of initially eight (later: seven) European countries.

The Memobust handbook consists of a large number of modules describing themes and methods that are relevant to the design and production of business statistics (including trade statistics¹). This modular form was chosen for the handbook to facilitate its maintenance. More details on the form and structure of the handbook are given in Section 2 of this introduction.

1.2 Aim, intended readership, and scope

The Memobust handbook is intended to update the *Handbook on the Design and Implementation of Business Surveys* (Willeboordse, 1998). In fact, ‘update’ is somewhat of an understatement. ‘Rewrite’ is a more apt description since both the structure and the contents of the handbook have been profoundly changed.

The purpose of the Memobust handbook is to aid those working in the area of business statistics. As such, the intended readership of the handbook is rather diverse. The handbook is primarily aimed at professionals who are active in the area of business statistics at (national or international) statistical institutes, including business survey managers, statisticians, and methodologists. It may also appeal to researchers from academia who want to learn more about the techniques that are currently being applied to produce business survey data in practice. In particular, the handbook should be helpful to those who are new to (a particular area of) business statistics.

The prerequisites are modest. The technical level of many of the contributions has been deliberately kept low, with the aim of getting across the basic ideas behind a technique or methodology. For those who want to delve deeper into a particular topic, references are provided to more advanced, more detailed or more technical material. In principle, these references should be publicly accessible and written in English.

¹ Below we shall typically refer only to business statistics, but this does not necessarily imply that trade statistics are excluded.

In principle, the scope of the Memobust handbook is restricted to describing those methods that are currently in use in the production of business statistics within the European Statistical System, or that could potentially be used as such. In the former case it concerns methods that have been around for some time. In the latter case it concerns promising methods from recent research. Inevitably, the handbook also discusses some aspects that are not strictly methodological (e.g., related to process design and quality) and/or that are not restricted to the area of business statistics (e.g., some methods could also be used for person or household statistics). The Generic Statistical Business Process Model (GSBPM; see Vale, 2009) is used to structure the material in the handbook. The scope of the handbook extends to all phases of the GSBPM, with an emphasis on those phases with a strong methodological component (viz Collect, Process, and Analyse).

The Memobust handbook is *not* devoted to laws, regulations, conceptual definitions, etc., although they are referred to if appropriate. Neither is the handbook intended to be prescriptive concerning the use of methods for the production of business statistics. Rather, the merits and demerits of different methods are described and compared.

The title of this handbook includes the word ‘modern’. It should be stressed that this is not so much a statement of fact as an appeal to keep the handbook up to date. This can only be achieved if the handbook lives up to its expectations and is valued by its users.

1.3 Business statistics

As mentioned above, the Memobust handbook is devoted to business statistics. But what are business statistics? The aim of this section is to demarcate, to define an area, not to answer a philosophical question. More specifically we try to answer this question by contrasting business statistics to social (person and household) statistics.

Let us state beforehand that we believe that the demarcation line is not a sharp one and in some areas virtually non-existent. Nevertheless, there are differences between both areas. In Kloek (2011), three differences between business and social statistics are mentioned explicitly (complexity of units, skewness of distributions, type of variables), but several more are implicitly being stated as well.

In Table 1 we have made a comparison of business statistics, household statistics and person statistics on various aspects. Regarding most aspects, business statistics and person statistics can be thought of as being on different ends of the spectrum, with household statistics somewhere in between. It should be stressed that the differences are not always that extreme in practice.

These differences between business and person/household statistics result in different methodological requirements. We refer to Kloek (2011) for an overview of these differences by methodological topic.

Table 1. Comparison of business, household and person statistics

Characteristic	Business statistics	Household Statistics	Person Statistics
Complexity of statistical units	Large	Medium	Small
Demarcation of units	Difficult	Fairly complicated	Easy
Dynamics of units	Complex	Complex	Simple
Size variation of units	Large	Small	None
Skewness of distributions	Large	Mostly small	Small
Type of variables	Mainly numerical	Mainly categorical	Mainly categorical
Number of variables	Small	Large	Large
Population size	Small-medium	Medium-large	Large
International comparison	Hard	Hard	Hard

2. Form and structure

2.1 Form

The name ‘Memobust handbook’ is somewhat misleading. In fact, this handbook is not a traditional book like its precursor and other existing handbooks on business statistics, such as Cox et al. (1995). Instead of being a monolithic structure, it consists of a set of separate, but interconnected, electronic documents (PDF files), called modules. In addition to this core material, there are a few documents that serve as introductory, contextual or background material.

The main advantage of this modular form is that it allows continuous updating. These updates may include the modification of existing modules, the addition of new ones, and the deletion of obsolete ones. This updating can be done locally, affecting only a small part of the handbook, while leaving the bulk intact.

2.2 Topics and modules

The Memobust handbook is subdivided into topics. Each topic covers a specific part of the methodology of business statistics, for instance ‘Sample Selection’, ‘Data Collection’, and ‘Statistical Disclosure Control’. A full list of topics in the handbook is given in the table of contents.

Each topic in the handbook consists of at least one module. There are two types of modules: themes and methods. Roughly speaking, themes are less specific and more

verbal pieces that aim to discuss a common point in a general, non-technical way. They point out, for instance, what certain techniques have in common, why they are used, etc. Methods are more specific, and usually more technical in nature. Themes should be suited for a rather broad readership, whereas method modules are predominantly written for specialists, such as methodologists. Both types of contributions have a standardised format. They are written using templates, which have been especially designed for the handbook.

Most modules are full, in the sense that they attempt to cover a topic adequately. There are also some modules that are deliberately kept 'meager'. Such modules are 'place holders'. Their function is mainly to identify a particular (sub)topic and use it for the sake of referencing from other modules. Place holder modules may be detailed in due course. Or they may stay like this, as they are on the borderline of methodology and another area.

2.3 Navigating the handbook

There are several ways to access the information in the handbook. First of all, the electronic modules are stored on the Memobust website in a hierarchical structure. This structure should provide sufficient information for a reader to find modules on a particular subject.

Another option is to use the glossary which provides access to relevant modules on the basis of key words. The glossary also serves, of course, as a source of explanation for technical terms, concepts, vocabulary acronyms, etc. used in the handbook.

Finally, the modules in the handbook contain many cross-references to each other. This makes it possible to navigate within the handbook, without reverting to the glossary or the hierarchical structure.

3. The project team

Although there are only three authors of this introduction, it should be clear that the Memobust handbook is the result of a joint effort by many people. First of all, we should mention that the handbook would not have existed without the initiative and financial support of Eurostat and of the NSIs involved in the project.

Below, we have tried to list the names of all persons who have contributed in some way or other to the creation of the handbook over the course of the Memobust project. This includes writers and reviewers of modules, as well as persons involved in organisational activities. As can be seen, the list is quite long; we apologise if anyone has been left out by mistake.

At Eurostat: Jean-Marie Bolis, Daniel Defays, Wim Kloek, Jean-Marc Museux.

At Statistics Netherlands: Dirkjan Beukenhorst, Max Booleman, Bart Buelens, Astrea Camstra, Barry Coenen, Jacco Daalmans, Piet Daas, Arnout van Delden, Bram Duyx, Deirdre Giesen, Wim Hacking, Abby Israëls, Ronald Janssen, Edwin de Jonge, Paul Knottnerus, Sabine Krieg, Rob van de Laar, Mark van der Loo, Nino Mushkudiani, Peter van Nederpelt, Feysel Negash, Jeroen Pannekoek, Sander Scholtus, Marc Smeets, Ger Snijkers, Henk van de Velden, Harrie van der Ven, Piet Verbiest, Pieter Vlag, Leon Willenborg, Peter-Paul de Wolf.

At Statistics Sweden: Evalena Andersson, Marianne Ängsved, Stefan Berg, Suad Elezović, Eva Elvers, Johan Erikson, Lina Fjelkegård, Ann-Marie Flygare, Almira Hecimovic, Annica Isaksson, Annika Lindblom, Rickard Nilsson, Anders Norberg, Tiina Orusild, Fredrik Scheffer, Yingfu Xie.

At the Central Statistical Office of Poland: Grażyna Dehnel, Magdalena Homenko, Tomasz Józefowski, Grzegorz Grygiel, Tomasz Klimanek, Paweł Lańduch, Andrzej Młodak, Monika Natkowska, Marcin Szymkowiak.

At the Istituto Nazionale di Statistica (Italy): Cristina Casciano, Patrizia Cella, Anna Ciammola, Nicoletta Cibella, Michele D'Alò, Claudia De Vitiis, Loredana Di Consiglio, Marco Di Zio, Marcello D'Orazio, Stefano Falorsi, Andrea Fasulo, Maria Liria Ferraro, Anna Rita Giorgi, Roberto Gismondi, Ugo Guarnera, Roberto Iannaccone, Orietta Luzi, Susanna Mantegazza, Manuela Murgia, Stefania Macchia, Paolo Righi, Fabiana Rocci, Mauro Scanu, Fabrizio Solari, Tiziana Tuoto.

At Statistics Norway: Øyvind Langsrud, Magnar Lillegård, Tora Löfgren.

At the Hungarian Central Statistical Office: Ágnes Andics, Zoltán Csereháti, Ildikó Györki, András Herczeg, Beáta Horváth, Miklós Juhász, Orsolya Kocsis, Attila Lukács, Katalin Szép, László Telegdi, Zoltán Vereczkei, Judit Vigh.

At the Swiss Federal Statistical Office: Daniel Assoulin, Monika Ferster, Monique Graf, André Hüsler, Anne Massiani, Desislava Nedyalkova, Lionel Qualité, Paul-André Salamin.

*At the Hellenic Statistical Authority (Greece):*² Adamantia Georgostathi, Ioannis Nikolaidis, Vasiliki Spiliopoulou.

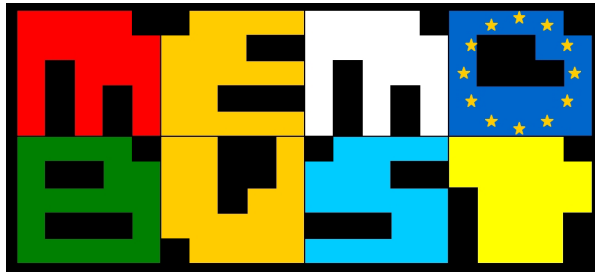
We want to express our gratitude to all colleagues involved in the project.

References

Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., and Kott, Ph.S. (eds.) (1995). *Business Survey Methods*. John Wiley & Sons, New York.

² The Hellenic Statistical Authority was involved only in the first half of the Memobust project, from January 2011 to June 2012.

- Kloek, W. (2011). What makes business statistics different? Paper presented at the European Establishment Statistics Workshop, Neuchâtel, 12–14 September 2011.
- Vale, S. (2009). Generic Statistical Business Process Model; Version 4.0 – April 2009. Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS).
- Willeboordse, A. (ed.) (1998). *Handbook on the Design and Implementation of Business Surveys*. Office for Official Publications of the European Communities, Luxembourg.



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Data Collection – Main Module

Contents

General section.....	3
1. Summary	3
2. General description.....	4
3. Design issues	7
4. Available software tools.....	7
5. Decision tree of methods	7
6. Glossary.....	7
7. References	7
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

Data collection is a “*systematic process of gathering data for official statistics*” (SDMX, 2009).

It is a very articulated process that develops itself along different steps of the survey process: from the design phase of the data collection methodology through the finalisation of the collected information (GSBPM, 2009), in order to collect data for statistical purposes by using many different techniques that can or cannot be assisted by computer and can or cannot need the support of interviewers (main ones: CAPI, CATI, WEB, PAPI, mail questionnaires and direct observation).

The choice of the technique to use depends on many factors (survey theme, timing of data delivery, difficulty in founding the information required, type of respondents involved, budget, etc.) and it is generally taken during the design phase of the process since the technique influences the way the data collection is carried out as well as the design of the survey questionnaire.

The use of mixed-mode, that is the combination of different data collection techniques for the same survey, can overcome those limitations that are specific of each technique and, if correctly designed, can reduce the unit non response rate.

A general trend among the NSIs is to gather the information they need by using administrative data in order to reduce respondent burden as well as costs. This is because NSIs can take the advantage of using already existing data, stored in public archives hold by other public organisations that have already performed a “data collection” phase, according to their needs and purposes that, anyway, might differ from the statistical ones. This trend is helped by the IT rapid developments in creating tools to facilitate the access to administrative data. Tools like these- the oldest EDI and the newest XBRL - represent another way of collecting data from public institutions as well as from enterprises, since they are based on the exchange of information among the data provider and the NSI on the base of a common and agreed structured data model.

Data collection process is not only a matter of interviewing techniques, but also of contact strategies as well as of monitoring activities: the first set of activities is necessary to get in touch with respondents and may vary according to the type of respondent unit (large or small enterprise, new enterprise, etc.). The second set of activities is important to keep under control the data collection while it is in progress and to take proper actions to improve or modify any factors that may badly interfere with data quality.

At the end of the data collection phase, information is ready to enter the next phase of the survey process, represented by the “*Phase 5.Process*” of the GSBPM, when data records are cleaned and prepared for the analysis. The way the following steps are faced and performed depends on how data collection is finalised since this depends on the mode(s) used to collect information.

2. General description

The data collection phase described in this module covers different sub-phases of the GSBPM that go from the design through the finalisation of data collection¹. In more detail, readers are guided through the following steps:

- design phase
- contact and reminders strategies
- preparation activities
- collection phase
- monitoring phase
- finalisation phase

These steps are deeply described in the theme modules that are linked to the present one. Specifically, they are:

- 1) “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method”;
- 2) “Data Collection – Design of Data Collection Part 2: Contact Strategies”;
- 3) “Data Collection – Mixed Mode Data Collection”;
- 4) “Data Collection – Techniques and Tools”;
- 5) “Data Collection – CATI Allocation”.

The first step of the data collection phase is the design of data collection methodology.

In this step researchers determine which are the most appropriate data collection method(s) and instrument(s)² as well as which is the most efficient contact strategy, where efficiency is in terms of many factors such as response rate, response burden, budget constraints, etc.

How to design the data collection methodology can be found in the theme modules mentioned above: the first one (“Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method”) describes which factors have to be considered when choosing data collection methods, advantages and disadvantages of each single possible mode to collect information and how these can be combined with budget and organisational constraints. The design of data collection methodology is in close connection with the questionnaire design phase, because the choice of the data collection technique is influenced by the questionnaire design and vice-versa: the various techniques allows for different interview lengths, different question formats, different question contents, different sets of checking rules, etc. This two-way influence is greater in mixed mode surveys where the questionnaire design has to take into account the presence of more techniques especially in the case when they are used concurrently.

¹ In the GSBPM these phases are labelled as: 2.3 *Design data collection methodology*, 3.1 *Build data collection instrument*, 4.2 *Set up collection*, 4.3 *Run collection* and 4.4 *Finalize collection*.

² Descriptions of the various steps are derived from GSBPM and are adapted to the aims of this topic.

The design of a mixed mode strategy is treated in the theme module “Data Collection – Mixed Mode Data Collection”, where both parallel and sequential mixed modes are described together with the steps to follow during the survey process (from designing to conducting a business survey) to prevent data from mode effect. Besides examples of recent mixed mode designs in business surveys from NSIs are described. They provide evidence on how to get a high response rate by using specific mixed mode strategies with unaffected overall response rates and data quality.

After the appropriate data collection mode(s) has been chosen, researchers have to design and set up the appropriate contact strategy, that is when and how respondents are contacted and what material (questionnaire, cover letter, instructions etc.) is used in each contact. In the theme module “Data Collection – Design of Data Collection Part 2: Contact Strategies” readers can find recommendations and suggestions on how to design this delicate phase of the survey process. In particular it describes which factors have to be considered, how contact strategy varies according to the type of businesses, how reminder strategy can be tuned according to the chosen contact strategy³.

Besides, a hint is given to responsive design approach to be used for both the design and contact strategy phases and how they can be modified during the data collection process to improve response rate.

After the design phase, data collection instruments have to be built following the specifications generated during the previous design phase (Sub-phase “*3.1 Build data collection instrument*” of GSBPM). This means that, depending on the type of mode(s) used, one or more data collection instruments have to be built (paper or electronic questionnaires, SDMX hubs, systems to extract and receive data from administrative archives) and their contents and functioning have to be tested. During the building phase it is also extremely important to establish a connection between the collection instruments and the metadata system, in order to facilitate data comparability inside the entire collection system and to reduce the work in subsequent phases. Preparing also for collecting paradata will be of great help in improving the collection step (Kreuter, 2013).

After the building phase, the collection of data can start (Phase “*4.Collect*” of GSBPM). How collection is performed depends on the chosen technique. Anyway, a common set of steps to be followed in order to gather data and to get them ready to enter the subsequent phase (Phase “*5.Process*” of GSBPM) of the survey process, can be described for any data collection mode. These steps are:

- preparation activities
- collection phase
- monitoring phase
- finalisation phase

Preparation activities are those activities to be carried out in order to be ready to collect data (sub-process “*4.2 Set-up collection*” of GSBPM). They include:

- training collection staff;

³ According to the GSBPM, “Contact strategy” and “Reminder strategy” are steps of the survey process that are carried out during the set up and running of the data collection phase (respectively sub-processes 4.2 and 4.3).

- ensuring collection resources are available, e.g., laptops;
- configuring collection systems to request and receive the data;
- ensuring the security of data to be collected;
- preparing collection instruments (e.g., printing questionnaires, pre-filling them with existing data, loading questionnaires and data onto interviewers' computers etc.).

The set of preparation activities can vary according to the chosen techniques: training of collection staff, for example, plays a fundamental role for interviewer-administered modes since it has to make interviewers able to collect data in the most objective way in order to reduce as much as possible the interviewer effect, that represents the effects on respondents' answers deriving from the different ways that interviewers administer the same survey (SDMX, 2009). On the other side, activities like ensuring data transmission security or availability of resources, like laptops, are peculiar of computer assisted data collection techniques.

The collection of data is run with the different collection instruments used to collect the data. It includes the initial contact with respondents and any subsequent follow-up or reminder actions. It records when and how respondents are contacted and whether they have responded (sub-phase “4.3 *Run Collection*” of GSBPM). For CATI surveys, the management of contacts with respondents is described in the theme module “Data Collection – CATI Allocation” that focuses on this peculiar feature of CATI, represented by the scheduling of telephone calls among the interviewers.

The monitoring phase is run while data collection is in progress in order to allow researchers to keep it under a constant control. Monitoring is based on a set of indicators about different aspects of the data collection like interviewers' productivity, response rate, non-response rate, refusal rate, interview length etc. In general, a unified system of codes for each indicator, to be used for any business surveys run inside an NSI, would be of a great help in computing comparable indicators (Györki, 2012). Besides, it would be advisable to use these codes to build quality indicators (see also the module “Quality Aspects – Quality of Statistics”) that will help monitoring the different problems that might arise during data collection that can cause non response errors, coverage errors and measurement errors (Eurostat, 2009). Example of these problems are:

- frame problems (status of the statistical unit: dead, under liquidation, etc.), classification problems, accessibility problems;
- problems with the activity of the statistical unit (no business activity: now, never, temporarily)
- problems referred directly to respondent (refuse to provide cooperation, no successful contact with him, etc.).

Finalisation of data collection starts when the collection of data is over. This step includes loading the collected data into a suitable electronic environment for further processing (4.4. *Finalize collection* – GSBPM). How finalisation of data collection is performed strictly depends of the technique used, being the computer assisted ones able to facilitate and speed it up. In fact, for these techniques, data are already stored in an electronic format and (partially) checked during the data collection itself. The consistency of final data can be further improved if the survey questionnaire has been designed following a metadata-driven approach or any techniques for relational database design.

How the above steps, from the preparation activities to the finalisation of data collection, have to be managed for the various data collection techniques is described in the theme module “Data Collection – Techniques and Tools”. This module considers only the main collection modes and divides them in two groups: interviewer-administered and self-administered. The first includes CATI, CAPI and Direct Observation while the second contains Mail and Web surveys. Besides, administrative data as well as data transfer through EDI and XBRL are also described.

How to collect statistical information from other data sources different from surveys is described in the module “Data Collection – Collection and Use of Secondary Data”. The entire process of collecting already existing data is generally referred to as the collection of secondary data. The topic discusses the advantages and disadvantages of this approach from an official statistics point of view together with research strategies with secondary data. A classification of secondary data types and an overview on the different types of use of secondary data by NSIs can also be found in this topic.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Eurostat (2009), *ESS Handbook for Quality Reports 2009 edition*. Eurostat Methodologies and Working papers. http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-EHQR_FINAL.pdf

GSBPM (2009), Generic Statistical Business Process Model. Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS) Version 4.0 – April 2009.

Györki, I. (2012), GÉSA: The Tool for Survey Control, Quality Assessment and Data Integration. *Hungarian Statistical Review, Special number 15*, 48–78.
http://www.ksh.hu/statszemle_archive/2012/2012_K15/2012_K15_048.pdf

Kreuter, F. (2013), *Improving Surveys with Paradata: Analytic Uses of Process Information*. Wiley.

SDMX (2009), Content-Oriented Guidelines Annex 4: Metadata Common Vocabulary 2009.

Interconnections with other modules

8. Related themes described in other modules

1. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
2. Data Collection – Design of Data Collection Part 2: Contact Strategies
3. Data Collection – Mixed Mode Data Collection
4. Data Collection – Techniques and Tools
5. Data Collection – CATI Allocation
6. Data Collection – Collection and Use of Secondary Data
7. Quality Aspects – Quality of Statistics

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

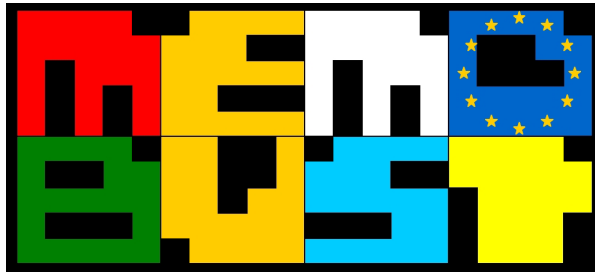
Data Collection-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-02-2012	first draft	M. Murgia	ISTAT (Italy)
0.2	03-08-2012	second draft	M. Murgia	ISTAT (Italy)
0.3	05-09-2012	third draft	M. Murgia	ISTAT (Italy)
0.4	19-11-2013	fourth version after EB revision	M. Murgia	ISTAT (Italy)
0.4.1	21-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:48



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Methods and Quality

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Methods and related objects	3
2.2 Conclusion.....	6
3. Design issues	6
4. Available software tools.....	6
5. Decision tree of methods	6
6. Glossary.....	6
7. References	6
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

This section describes what the dependencies are between the quality of the *statistical output* and the quality of the *method*. Various quality dimensions of the method are distinguished. Each dimension should be managed by taking the right measures. To attain a structural approach to the theme, the Object-oriented Quality and Risk Management (OQRM) model is used.

2. General description

In this section, statistical methods are placed in the context of quality (and risk) management. It will describe which measures or actions are taken in the Memobust project to assure the quality of the method description. In addition, this document assesses which measures should be taken by the users of the handbook in order to manage the quality of the statistical methods.

To attain structure in this document, the OQRM model (Van Nederpelt, 2012) is used. The next section concisely describes the OQRM model. This model is elaborated in the module “General Observations – Quality and Risk Management Models”. An important concept of the model is that the quality of the product is dependent on the quality other objects such as methods. An object is everything that can be perceived or conceived. Another concept of the model is, that quality of an object can be decomposed in attributes (also: characteristics or quality dimensions) of that object. For example, an attribute of the object *statistical output* is *accuracy*.

2.1 Methods and related objects

Looking at the object *method*, the following related objects can be distinguished. These objects can be regarded as part of the *method* family and have each their own set of attributes:

- Description of the method
- Values of the parameters of the implemented method
- Software implementation of the method
- Input and output data of the method
- Metadata output of the method

Methods are aimed at processing data in order to produce statistical output. A specific category of methods is the measurement processes for quality indicators and logging. A separate theme module in the handbook is dedicated to logging (“General Observations – Logging”).

In the Generic Statistical Business Process Model (GSBPM, 2009) the design of methodology is described at least in sub-process: 2.3 design data collection, 2.4 design frame and sample, and 2.5 design statistical processing. Figure 2 in the module “General Observations – GSBPM: Generic Statistical Business Process Model” shows which sub-processes have high (green), intermediate (light green) or little or no (light yellow) methodological content.

2.1.1 Method

The method can be distinguished in the method as such and the method as implemented in a statistical process. Both will be covered by this section. Attributes of a method are:

- Soundness
- Appropriateness
- Applicability, usability and stability (GSIM 0.3, 2012, section 35)
- Feasibility
- Complexity
- Efficiency
- Robustness

Soundness of a method can be defined as the extent to which the method(ology) used to compile statistics complies with the relevant international standards (SDMX, 2009).

The CoP (principle 7) states that soundness of methodology “underpins quality statistics”. Furthermore, principle seven says that “it requires adequate tools, procedures and expertise”. An important criterion is, that methods are accepted by the international community in the statistical domain. This handbook describes, however, also methods that have some drawbacks but that still will be used because better methods are not yet available. The *accuracy of the estimated value* is definitely dependent on the *soundness of the method*.

Appropriateness of a method is the degree to which a method can be applied to a specific statistical process. A related attributes is *applicability*. In the template of this handbook this focus area is promoted by the sections: purpose of the method (8), recommended use of the method (9), possible disadvantages of the method (10), logical preconditions (13) and the recommended use of the of the individual variants of the method (11). An inappropriate *method* will affect the *accuracy of the output*.

Usability of a method can be defined as the degree to which staff understand the method they use in production. It is required that they understand how they should tune the parameters of the methods not only initially but also in the course of time. This focus area should be managed by the methodologists who redesign a statistical process. A method that is not usable can affect *various quality dimensions of statistical output*.

Complexity of a method can be defined as the degree to which capacity is needed to use and implement the method correctly. A method that is too complex can take too much time to customise. A difficult method can be wrongly communicated, more easily be misunderstood, or applied wrongly. If a method should be implemented in custom made software it takes a more capacity to develop that software. So, there is a relationship between the *complexity of the method* and *costs of development* and *correctness of the software* too. Feasibility is related to complexity of the method.

This focus area can be addressed in section possible disadvantages of the method (10). Moreover, it should be managed by the methodologists in the redesign process. Methods that are too complex can affect the *cost* as well as the *timeliness of statistical output*.

Robustness of a method is the degree to which a method withstands different input data and produces nevertheless accurate output data. Stability is related to robustness of the method. This focus area can be addressed in section possible disadvantages of the method (10) and should be managed by the methodologists in the redesign process. A method that is not robust can cause inaccurate statistical output.

Efficiency of a method is the degree to which a method needs IT-resources or computing power. In case of large datasets, it could take too long to process this dataset using a specific method. The *complexity of the method* will certainly affect the *efficiency of the method*. This focus area could be addressed in the section possible disadvantages of the method (10) and should be managed by the methodologists in the redesign process. A method that is not efficient can affect the *cost* as well as the *timeliness of the statistical output*.

2.1.2 Description of the method

Completeness, correctness, clarity, unambiguity and consistency of the descriptions of the method. Relevant attributes of the description of the method are *completeness, correctness, clarity, un-ambiguity and consistency*. *Completeness of the description* is promoted by using a certain structure (template) for the descriptions. Furthermore, all focus areas are assured because the content is written and reviewed by experts. An inappropriate description of the method can cause various problems including problems with the *accuracy and timeliness of statistical output*. It can also lead to a misunderstanding of the actual method used.

2.1.3 Values of the parameters

Correctness of the values of the parameters. Part of the implementation of the method are the *values of the parameters*. Relevant focus area of the values of the parameters is the *correctness of the values of the parameters*. This focus area can be defined as the degree to which parameters of the method are sufficiently adjusted or tuned. In the template for the methods this focus area is addressed in section tuning parameters (14) and recommended use of the individual variants of the method (11).

The *correctness of the values of the parameters* can affect the *accuracy of the estimate* and the *costs of the statistical output*. For example, the sample size is a parameter of the sampling method. The larger the size of the sample, the larger the *costs of the statistical output* will be. The *accuracy of the estimate*, on the other hand, will increase.

2.1.4 Software implementation of the method

Correctness of the software. Relevant focus area of the software implementation is the *correctness of the software implementation of the method*. This can be defined as the degree to which the method is correctly implemented in the software. This focus area is not covered by the handbook. However, problems with this focus area will definitely affect the *accuracy of the estimate*. An action to assure the *correctness of the software* is, e.g., testing after development of the software and after each change of existing software. Another measure is to deploy competent and specialised software developers. Incorrect software will affect the *accuracy of the statistical output*.

Performance of the software. Another focus area is the performance of the software. This is the time the application needs to process input data. The *performance of the software* is among others

dependent on the *complexity of the method*. It is, moreover, dependant on other focus areas like *efficiency of the software, performance of the software tool and performance of the IT infrastructure*. Logging indicators (20) can be used to measure the *performance of the software*. Software that performs badly can affect the *costs and timeliness of statistical output*.

2.1.5 Input and output data of the method

Quality of input and output data is a specific field of expertise which will not be addressed here. A separate handbook could be compiled about this subject. However, bad quality of data (input, intermediate results) will certainly affect the *accuracy of estimates*.

2.1.6 Metadata output

Quality of metadata output. A statistical process can next to statistical data produce metadata such as logs and quality reports. Some metadata are related to the method used in the process. Other metadata are related to the input data, the process or the output data. In the template of the method, quality indicators for output quality are defined (21). Statisticians should take corrective actions if the agreed quality is not met or at least report this to the supplier of the input data or the user of the output data. The quality of metadata will not be elaborated here.

2.2 Conclusion

The focus area *accuracy, consistency, timeliness and the costs of the statistical output* are to a substantial degree dependent on the *quality of the methods* as well as to the quality of other objects. This handbook contributes to the assurance of the *quality of methods*.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Eurostat (2011), *European Statistics Code of Practice*. Adopted by the European Statistical System Committee, 28th September 2011.

Eurostat (2012), *Eurostat's Concepts and Definitions Database*. Website:

http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GL_OSSARY&StrNom=CODED2&StrLanguageCode=EN

- GSBPM (2009), Generic Statistical Business Process Model. Version 4.0 – April 2009. Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS).
- GSIM (2012), Generic Statistical Information Model. Version 0.3, March 2012.
- Van Nederpelt, P. W. M. (2012), *Object-oriented Quality and Risk Management (OQRM). A practical, scalable and generic method to manage quality and risks*. MicroData, Alphen aan den Rijn, The Netherlands. Website www.oqrm.org/English.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Quality and Risk Management Models
2. General Observations – Logging
3. General Observations – GSPBM: Generic Statistical Business Process Model

9. Methods explicitly referred to in this module

1. Micro-Fusion – Prorating
2. Micro-Fusion – Minimum Adjustment Methods
3. Micro-Fusion – Generalised Ratio Adjustments
4. Macro-Integration – RAS
5. Macro-Integration – Stone's Method
6. Macro-Integration – Denton's Method

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 2.3 Design data collection
2. 2.4 Design frame and sample
3. 2.5 Design statistical processing

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

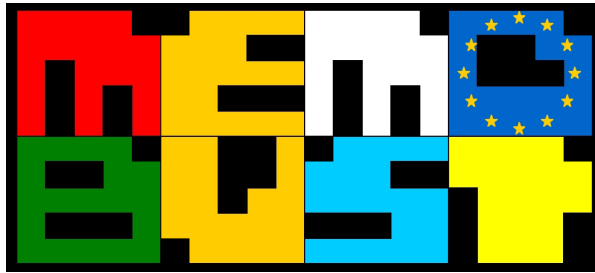
General Observations-T-Methods and Quality

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	02-04-2012	first draft	Peter van Nederpelt	Statistics Netherlands
0.2	10-06-2012	comment of Greece processed	Peter van Nederpelt	Statistics Netherlands
0.3	26-09-2013	comment of Editorial Board processed (first part)	Peter van Nederpelt	Statistics Netherlands
0.3.1	30-09-2013	preliminary release		
0.3.2	02-10-2013	comment of Editorial Board processed (second part: Leon Willenborg).	Peter van Nederpelt	Statistics Netherlands
0.4	04-10-2013	text about quality management moved to theme Quality and Risk Management models	Peter van Nederpelt	Statistics Netherlands
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:21



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Factors to consider when choosing data collection method	3
2.2 Different modes	4
2.3 How to mix modes.....	11
3. Design issues	11
4. Available software tools.....	12
5. Decision tree of methods	12
6. Glossary.....	12
7. References	12
Interconnections with other modules.....	14
Administrative section.....	15

General section

1. Summary

The chapter gives an overview of factors to consider when choosing data collection method. It also gives a short presentation of different modes available, modes suitable for business surveys, advantages and disadvantages with each mode and a brief description about how to mix modes.

2. General description

2.1 *Factors to consider when choosing data collection method*

There are several factors to consider when choosing data collection method and each method has its pros and cons. A general idea is to choose the method that minimises the total survey error (TSE) given the budget constraints. Some factors affecting the choice of mode and data collection instrument are response burden, desired data quality (e.g., in terms of nonresponse and measurement error), available resources (budget and staff, but also IT-resources and technical conditions), topic of the survey and the questionnaire content, sampling frame, properties of the target population (e.g., type of industry) and timetable for the survey (e.g., Biemer et al., 1991; Groves et al., 2004).

For instance, response burden can be reduced by good questionnaire design, extracting files automatically or by pre-printing information from previous reporting periods in the questionnaire. Lower response burden may also be achieved by sample coordination and sample rotation. For long surveys with complex calculations, an electronic self-administered questionnaire that guides the respondent through the form with built-in helps and logic checks might be an appropriate alternative. Some electronic questionnaires might also allow the reporting person to save data temporarily and continue later on if figures have to be looked up in other systems or files. Regardless what method is chosen, a contact strategy must also be defined when planning the data collection; how and when the respondents will be contacted.

One major difference between household surveys and business surveys is that in business surveys (most often) many employees cooperate in the reporting task, something that makes the response situation more complex; see the module “Response – Response Process”. We do not know much about how the tasks are divided or communicated internally within the businesses, we can only suppose this complexity makes questionnaire design even more important. Some employees might forward the whole questionnaire including instructions to a colleague; while others might interpret the question themselves and just ask the colleague for a figure (i.e., the colleague will never see or read neither the question nor the instructions). In some businesses only a few persons are authorised to report, but this does not necessarily mean that the authorised person has the knowledge to report. The questionnaire might be sent around to different employees within the business who partially fill out and report the figures they have knowledge on. In some businesses paper questionnaires are preferred, because “paper walks”. Other businesses find electronic self-completion questionnaires easier to handle in the reporting situation. The differences in preferences are often related to factors like for instance business size, organisation levels (hierarchy) and type of industry.

Business surveys are also a bit special in the sense that business populations have distinct frame problems. Often they vary quite much in size and they are highly dynamic. Small businesses are born and die rapidly. Medium-sized or large businesses merge with others or split up into several units. The

business population also demonstrates a distinction between a legally defined entity and physical location (Groves et al., 2004). These are also factors to consider when designing data collection and choosing mode.

Another important step in planning the data collection is to consider how the final result, the statistics should be presented. Which variables should be reported and how detailed should they be? How shall we get hold of this information; shall the variables be collected from a register, shall they be collected directly through a questionnaire or are the variables so complex that they have to be created by compound calculations? These kinds of choices will not only affect the level of response burden in the survey, but also the level of accuracy during the data collection which is also an important design feature which should be reflected in the choice of mode. In an interview, the interviewer can give the respondent more support than in a postal questionnaire, where there are limited opportunities to help the respondent to fulfil the task. In electronic self-administered questionnaires, controls can be built in which can be both an advantage and a disadvantage for the respondent. When designing the data collection instrument, research problems have to be translated into questions in the questionnaire without creating a mismatch opening up for specification- and measurement errors. One also has to ensure that all topics are covered in the questionnaire, i.e., no variables are missing. The planning and design process is a continuous process where improvements are made by iterations. Instrument design and testing questionnaires are dealt with more in detail in the topic “Questionnaire Design”.

Each survey has its own conditions, specific errors and how to treat them. In general, little is known about the relationship between quality, time, costs and response burden and it is hard to implement measures to reduce the burden without the expense of quality. Too few quantitative before-after studies are at present documented and actions intended to reduce response burden should be monitored, reviewed, documented and published better in order to gain more insight (Giesen, 2011).

2.2 *Different modes*

The *mode of data collection* refers to what medium is used for contacting the sample members to get their responses to the survey questions. The principal modes for data collection are: *face-to-face surveys*, *telephone survey*, *mail surveys* and *web surveys*. Face-to-face surveys and telephone surveys are often referred to as *interviewer-administered modes*, whereas mail surveys and web surveys are referred to *self-administered modes*.

The data collection can also be divided into direct and indirect data collection, referring to the level of contact with the respondent. For instance, administrative records are an indirect form for data collection with no contact with the respondent and a low data collector involvement; this in contrast to many of the other modes which are methods for direct data collection. The table below gives an overview over different modes, the level of data collection involvement from the data collector and level of contact with the respondent.

Table 1. Modes to choose from when planning the data collection.

	High Data Collector Involvement		Low Data Collector Involvement	
	Paper	Computer	Paper	Computer
Direct				
Contact with Respondent	Face-to-face (PAPI)	CAPI	Diary	CASI, ACASI
Indirect				
contact with Respondent	Telephone (PAPI)	CATI	Mail, fax, e-mail	TDE, e-mail, Web, DBM, EMS, VRE
No Contact with Respondent	Direct observation	CADE	Administrative records	EDI

ACASI, audio CASI; CADE, computer-assisted data entry; CAPI, computer-assisted data interviewing; CASI, computer-assisted self-interviewing; CATI, computer-assisted telephone interviewing; DBM, disc by mail; EDI, electronic data interchange; EMS, electronic mail survey; PAPI, paper-and-pencil interviewing; T-ACASI, telephone ACASI; TDE, touch-tone data entry; VRE, voice recognition entry. *Source: Biemer and Lyberg (2003).*

The modes have different advantages and disadvantages when it comes to costs, measurement errors, nonresponse and coverage, flexibility and timeliness. Questionnaire complexity and the respondents' possible reporting preferences are also important factors to consider, something that sometimes leads to a mixed mode solution when collecting data for the survey. Mixed-mode design might help in satisfying the respondent's preferences and hereby the response burden might be lowered. Even if lower response burden is highly desirable, it might sometimes be wise not to offer too many different modes at the same time. This is because too many computer systems to look after for the national statistical institute (hereafter called NSI) will be costly in the long run. Mixed mode also opens up for possible different error sources that might be difficult to combine and handle later on in the statistical process.

Below follows a short review of some of the modes presented in Table 1. The review primarily focuses on the modes relevant for business surveys, but as always there are exceptions and differences between countries depending on domestic conditions, which might have the greatest impact on the choice of mode at the end.

2.2.1 Mail surveys

The mail survey is carried out by a paper questionnaire sent to the sample respondents by mail. The data collector has no control over the response process or who is actually responding to the survey (e.g., Biemer et al., 1991). The response process is as previously mentioned even more complex in business surveys and sometimes it is a challenge just to find the right person within the business to mail the questionnaire to.

Mail surveys are quite inexpensive to implement, which make them the preferred mode for low-budget surveys. At the same time, mail surveys often require a long field period with at least one reminder to achieve acceptable response rates (Biemer and Lyberg, 2003). The respondent deals with the survey on its own and there is no interviewer present who can provide support or explain difficult questions. Some NSIs have chosen to have a support centre or help desk for business surveys, which the business representatives can call and ask for help when reporting. It is also common to include a telephone

number to the person who is responsible for the publication or statistical analysis in the questionnaire or in the advance letter.

The potential problem with complicated questions can be eased by a well-designed questionnaire that motivates and guides the respondent through the questionnaire by good navigation, help texts and visual support (e.g., Groves et al., 2004). Visual support and technical facilities can be made extra efficient in electronic self-completion questionnaires (see next section 2.2.2).

The quality of the answers in a mail questionnaire is to a greater extent depending on the questionnaire design than in interviews. However, it has been shown that response order and question order is less important in a mail survey, as the respondent can easily navigate back and forth in the questionnaire (Biemer et al., 1991). There is also less risk of social desirable responses for sensitive issues in mail surveys than in the interviewer-respondent situation (Biemer et al., 1991). For mail questionnaires there is a greater risk of *primacy effects*, i.e., the respondent chooses one of the first response categories when answering the question (e.g., de Leeuw, Hox and Dillman, 2008). *Open-ended questions*, where the respondent has to formulate the response on his/her own are less suitable for mail questionnaires. The respondents have proven to give less and less thoughtful answers to such questions in mail surveys than in an interview situation where the interviewer can help the respondent in formulating the answer by probing. In business surveys open-ended questions might lead to a situation where the data collector does not know what is included in the numbers reported. Without the interviewer directly motivating the respondent to participate, mail surveys typically have lower response rates than interviews and the risk of item nonresponse is also bigger in mail surveys (Biemer and Lyberg, 2003). However, the nonresponse rate is in general not the biggest problem in business surveys, since reporting most often is mandatory and failure to report will lead to fines.

2.2.2 Web surveys

Web surveys are based on self-administered electronic questionnaires which are often viewed upon as a technical version of the mail questionnaire. Logic checks and visual guidelines can be built in, but advanced solutions cost hours of programming and there is a risk of ending up with higher response burden due to all the technical features if they are not well specified and tested.

Web surveys are perhaps the most common mode for business surveys today. Many NSIs introduce electronic versions of the survey due to aims in cutting the costs for data collection and/or data editing, with the intention to improve data quality, in order to offer safe communication with businesses or in order to make it easier to respond and thereby aiming to lower the response burden (Giesen, 2011, Chapter 5).

Web surveys might also be offered for specific surveys or specific groups of surveys where reporting on the web has been found to suit the survey topic well, or where different versions of the questionnaire are sent to different subgroups in the population (e.g., small businesses).

Computerisation allows lots of built-in features like customised wording, mouse-over-help, skips and jumps, edit checks and randomised question order. These features or refinements can be said to replace the role of an interviewer that helps the respondent through the survey. Visual elements like brightness, color, shape and position can be used in order to guide the respondent through the questionnaire (Groves et al., 2004). These features have shown to lead to less measurement error and less item non-response (ibid). The visual potential might also lower the response burden.

A factor to be considered when choosing the most suitable mode is that web surveys can be run on-line or off-line. As described in the module “Data Collection – Techniques and Tools”, these two ways offer the respondents the opportunity to compile the questionnaire directly on the survey web site or to download it, fill it out and send it back later on when finished.

Some examples of web-surveys in Europe: Statistics Norway introduced electronic reporting for all business surveys as a new data collection strategy; the primary data collection mode is nowadays the web (e.g., Haraldsen et al., 2011). Statistics Lithuania introduced web-surveys to create a favourable environment for the businesses in order to prepare statistical data at lower costs (e.g., Lapeniene, 2008). At Statistics Netherlands, more than half of the business surveys are available in electronic forms (e.g., Beukenhorst and Giesen, 2010) and in the latest years, work has been targeted on an electronic version of the annual Structural Business Survey (e.g., Snijkers et al., 2007) on the web. Further examples can be found in Raymond-Blaess (2011).

No matter the reason behind an electronic version of a self-completion questionnaire, there is no clear evidence that web-surveys do imply higher data quality and decreased response burden, even if some measurements suggests something in that direction (Snijkers et al., 2007; Giesen et al., 2009). Electronic data collection adds complexity to the response process which is already complicated within a business, and the respondent has to interact not only with the questions, but also with their internal records and the electronic instrument itself. Initially, switching from paper to electronic questionnaire might actually increase the (perceived) response burden and how well an electronic instrument will work in a business survey depends on several factors, such as the organisational structure, the size of the business, what industry the business operates in and the kind of products or services it sells (e.g., Goddeeris and Bruynooghe, 2011; Gravem, Haraldsen and Löfgren, 2011). Not all survey topics are suitable for electronic reporting. Sometimes a paper questionnaire is more convenient for the respondent because it is easier to handle in the reporting situation. On the other hand, electronic questionnaires can be designed to offer the same flexibility the respondent perceives it has with a paper questionnaire, something that can be achieved by creating a web-portal. The portal is not only a place to gather the surveys; it is also a system for survey administration - both for the respondents and the NSI. An example of a web portal is the AltInn-portal in Norway, where different informants can log-on and report on the parts they can contribute to and have knowledge on. This kind of web-portal solution is getting more and more common in Europe.

2.2.3 Administrative records

If existing administrative records can be used, there is not only money to save but also response burden since the respondents will not have to cope with another survey request. The error structure for administrative data is similar to those of other modes, this because the administrative records are produced on data collected somehow originally (Biemer and Lyberg, 2003). Administrative records might consist of data collected by some other institution than the NSI, but might also be data already collected by the NSI in a different survey. A good property with administrative records is that they most often cover the whole population. On the other hand, the drawbacks with administrative records is mainly that they may relate to a somewhat different population than the target population of the survey, leading to calls for further measures to achieve coverage. The content of the records is not always adapted to the wishes of statistics users and statisticians sometimes have no control over the record or how the record is updated (Biemer and Lyberg, 2003). Definitions, boundaries and variable

content may differ from those desired, so the parameters cannot be estimated easily and the NSI sometimes has to rely on model-based estimates. It is not unusual that the statistical purpose of a record comes in second hand, after the administrative ones which often are of primary interest. Different records have different data quality and this goes back to the main data collection or how the record is updated. Conceptual problems are common, especially when it comes to business surveys where there often is a mismatch between what data the businesses have and what data the NSIs ask for (Giesen, 2011).

2.2.4 Touch-tone Data Entry (TDE)

TDE is an alternative to mail collection and is a method where the respondent calls a computer linked to an automatic answering machine and reports by pressing the touchtone phone buttons. Usually, the answers are also read back for the respondent for verification (Biemer and Lyberg, 2003). TDE is only a good option in very short surveys with few questions where the answers are related to numerical information. There are, unfortunately, not many surveys that meet these requirements and there are also some up-front costs associated with using TDE in a survey, e.g., to program the hardware. The possibilities for editing during the process are also limited under this mode (Cox et al., 1995).

2.2.5 Electronic Data Interchange (EDI)

Electronic exchange of information is nowadays standard in the business world as many businesses are moving towards a paperless environment. EDI offers businesses an electronic way to exchange common standard information like order forms, shipping notes and other documents (Cox et al., 1995). The possibility to submit data by removing a file from the system and sending it to the NSI has many advantages. The respondents extract the needed data in a pre-specified format from their computer systems and transfer them to the NSI. Sophisticated EDI systems also offer direct on-line editing by the respondent (Cox et al., 1995). There is a minimal effort for the respondent, except for the first time when the base file has to be created, and response burden is therefore low. The quality of the data is dependent on the file but if it is created and updated correctly the quality might be good. The EDI technique may be used to collect large volumes of data and information from businesses.

2.2.6 Data provided by automatically extracted files (e.g., XBRL)

eXtensible Business Reporting Language (XBRL) is a technical standard for electronic communication of business and financial data and is based on the XML and Link technical standards. The idea of the XBRL language is to identify each concept (e.g., turnover) and add it into a “taxonomy”, which works like a dictionary. Once defined, they can be re-used by other users. The technique has potentials in reducing response burden (Allen and Junker, 2008) and offers flexibility to the businesses. XBRL might be a good solution for businesses of large size and/or businesses that do not report themselves, but use an external accountant that have to report on the same survey on a regular basis (Goddeeris and Bruynooghe, 2011).

The relationship between computerisation and quality is not straight forward. The main strength of computers is not that they do things right, but that they do things consistently. This means that in case of incorrect programming or linkage between the statistical need and the source of information, the computer program will consistently produce errors as a result.

The XBRL-technology also struggles with two kinds of updating problems. The first is linked to when questions in the survey are changed and the second is more related to changes in staff. When questions are changed, the software company has to develop a new version and implement it at the customers, which might be a diminishing problem as more and more software updates are available on Internet. Still, this fact implies that automatic data capture will work best in stable environments with fixed survey contents. The second problem is the transfer of competence when people leave a workplace; ensuring the knowledge and experience to link the administrative systems with the statistical ones will be transferred to someone else within the company (e.g., Haraldsen et al., 2011).

Many NSIs are active in this field with different development projects; for instance Statistics Finland developed an automated data capture procedure for hotel accommodations in 2005 (Savolainen and Vertanen, 2007; Orjala, 2010). Destatis in Germany developed the eSTATISTIK.core (2008) which uses the XML file format, and the statistical bureau in Spain – Instituto Nacional de Estadística – developed a XML based system for the hotel occupancy survey 2008 (INE, 2008). Another successful project that shows the potentials within this area is the Simplified Business Information system (Portuguese acronym IES) developed in partnership with different public entities, including Statistics Portugal. The system makes it possible to acquire administrative and statistical information in a coordinated manner, conducted electronically on one single occasion for the whole population of enterprises. The IES system also represents an improvement on the quality dimensions; coverage, coherence, punctuality, timeliness, comparability and reliability for business statistics (Pereira, 2011).

2.2.7 Face-to-face interviewing – PAPI and CAPI

Face-to-face (PAPI) interview is the oldest mode of interview since it does not rely on modern technology. The mode involves direct contact with the respondent and the data collector is highly involved. When a computer is used instead of paper-and-pencil in the interview situation, the mode is often referred to as CAPI.

PAPI and CAPI are not very common modes in business surveys; however they are used in some countries that for instance lack a business register and/or have problems in locating or contacting the businesses. There might also be some survey specific circumstances when the modes might be a good choice; e.g., when the respondent clearly would benefit the support from an interviewer (e.g., help in recalling events, amounts or frequencies of some phenomenon) or has no access to Internet.

PAPI and CAPI are by far the most expensive data collection methods especially when the respondents are spread over large geographic areas; mainly because of travel and lodging expenses for interviewers as well as interviewer training. In the case of CAPI the interviewer also has to be equipped with a computer. The mode has traditionally been associated with high quality, mainly due to the interviewer's presence and the positive effects from that. Besides for CAPI, the pc-support has the same advantages mentioned for web surveys.

This view has changed in recent decades due to the discovery of measurement error and the problems face-to face interviewing potentially brings, especially for questions on sensitive topics (Biemer et al., 1991). Personal contact is efficient when persuading respondents to participate, something often mirrored in the high response rates for face-to-face interviewing compared to other modes. A face-to-face interview may be longer and cover more complex issues than a telephone interview or a questionnaire sent by mail. At the interview the interviewer can control the response situation; that the

respondent has understood the question and ensure that the response is not influenced by other persons, or that it is the intended respondent who responds to the survey and not someone else. The latter is for instance something out of the NSIs control when sending out a questionnaire by mail.

Another advantage with the face-to-face interview is that the interviewer can use visual aids in the field work, e.g., cards with response categories; something that would not be possible in a telephone interview situation (Biemer et al., 1991). The presence of an interviewer can also have a negative effect on the responses and the quality of the data collected; interviewers affect the respondents' answers in a way similar to the clustering effect in cluster sampling. The responses are affected through the individual interviewers' behaviour and performance pattern during the interview. Different interviewers have different behaviour patterns and they ask the questions in their own style and pace and the question wording might not always be exactly as in the questionnaire. The interviewer effect is strongest particularly in face-to-face interviews and especially on sensitive issues where the interviewer's influence can lead to so called *social desirability bias* (e.g., Biemer and Lyberg, 2003). Social desirability bias is probably more common in household surveys, but can occur in business surveys too depending on industry covered and topic of the survey. For instance, businesses within an industry known for air pollution might report strategic or "brushed up" figures when it comes to environmental investments in cleaning technology or environmental protection with the intention to make them look better in public.

2.2.8 Telephone interviewing (CATI)

Telephone interviewing is the fastest data collection mode to implement from start to completion of data collection and is often used in combination with other modes in mixed-mode surveys (Biemer and Lyberg, 2003). The mode is not so common in business surveys in the data collection phase, but rather when it comes to call-backs, the editing phase when trying to fill out missing values or to reduce nonresponse. However, the mode is still used in business surveys in some countries, e.g., in agricultural surveys, and therefore it is included in this review.

By building common survey procedures directly into CATI systems, or into pre-packed setup modules, surveys with similar designs can be conducted more efficiently, even by staff with limited survey experience (Groves et al., 1988). The telephone interview shares some of the advantages and disadvantages with the face-to-face mode concerning the interviewer presence, as well as some of the advantages of the electronic questionnaires mentioned before. The interviewer effects and risks of social desirability bias are however lower than in face-to-face interviews. A disadvantage with telephone interviews is they are less flexible. Visual aids cannot be used, and neither the survey topic nor the survey questions (or the response categories) can be too many or too complicated in a telephone interview situation (e.g., Biemer et al., 1991). With too many response categories the respondent might forget and systematically pick the last response category read; something called *recency effect* (e.g., Biemer et al., 1991). The respondent might also interrupt the interviewer after the first response category has been read and say "yes" to that one, not letting the interviewer finish the job with reading the other response alternatives. This phenomenon is often referred to as *top-of-the-head-responses*. Top-of-the-head responses occur in all modes, but are perhaps more frequent in telephone interviews (see Biemer and Lyberg, 2003). The influence on data caused by recency effects and top-of-the-head-responses can be diminished by some programming if questions and response categories are allowed to be randomised within the questionnaire. If the survey questions require some

extra effort from the respondent like a check-up in computer systems or calculations, both face-to-face interviews and telephone interviews are less suitable modes. A growing problem in general with telephone interviews is that parts of the population may be difficult to reach since they are not listed in the telephone book (e.g., Biemer and Lyberg, 2003). This phenomenon is increasing as more and more people use only their mobile telephones and do not have fixed land line (e.g., Lepkowski et al., 2008). Naturally, finding the telephone number of the business is in general not a huge problem when conducting business surveys; the issue lies more within finding the right person within the business.

2.2.9 Direct observation

Direct observation in the field means that data are collected without direct involvement of a particular respondent; the observer assumes the role of the respondent (Biemer and Lyberg, 2003). The mode is often used in biology and qualitative research (de Leeuw, Hox and Dillman, 2008) but can also be used in data collections in business surveys for official statistics. An example of direct observation is when the goal is to estimate the proportion of trucks in traffic on a ring-road around a city, where observers register the number of trucks travelling at a random place during a randomly selected time period. Measurement errors for this mode may be introduced by the recording of observations by the observers in ways similar to the errors introduced by interviewers. The measurement errors may also relate to the instrument or device used to gather information. Large scale data collections using direct observation as mode are found in most agricultural surveys (Biemer et al., 1991).

2.3 How to mix modes

With all these mode possibilities there is a good opportunity to combine the strong points of each mode offering the respondent several modes for reporting. Such mode decision has to be planned carefully because it implies a more complicated, more expensive, longer and probably more challenging survey implementation. The usual goal is to find an optimal mix for data collection given the research question and the population under study given the restrictions (Biemer and Lyberg, 2003). The reason for mixing modes might be to collect follow-up panel data from the same respondent at a later time, but also to collect data from same respondents during a single data collection period. Mixed mode can be carried out to meet the respondents' preferences, but usually the main reason for mixed mode surveys is to battle the nonresponse. The general idea of mixed mode is to start with one main mode and when all possibilities are emptied for that mode a switch to another often more expensive mode is made, and so on. Allowing mixed modes or letting for instance businesses completely choose and define the agenda how they want to report might not be the best approach in the long run. Different modes have different ways of contacting the respondent which affects the answers, something that might cause problems in comparative surveys if the instruments are not well designed. It might also be costly to develop and maintain the data collection systems for each mode (e.g., de Leeuw, Hox and Dillman, 2008). A more detailed description on mixed modes can be found in the module "Data Collection – Mixed Mode Data Collection".

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Allen, J. and Junker, C. (2008), How far can IT standards and tools help to reduce response burden? Paper presented at the 94th DGINS conference, Vilnius, 25th-26th of September 2008.
- Beukenhorst, D. and Giesen, D. (2010), Internet Use for Data Collection at Statistics Netherlands. Paper presented at the 2nd International Workshop on Internet Survey Methods, Statistics Korea, Daejeon, South Korea, September 8 & 9, 2010.
- Biemer, P. P. et al. (1991), *Measurement Error in Surveys*. Wiley Series in Probability and Mathematical Statistics, New York.
- Biemer, P. P. and Lyberg, L. E. (2003), *Introduction to Survey Quality*. Wiley Series in Survey Methodology, New Jersey.
- Cox, B. G. et al. (1995), *Business Survey Methods*. Wiley Series in Probability and Mathematical Statistics, New York.
- e.Statistik.core (2008), Neue Wege zur Entlastung der Unternehmen. Statistische Bundesamt, Wiesbaden. (www.statistik-portal.de)
- de Leeuw, E. D., Hox, J. J., and Dillman, D. A. (2008), *International Handbook of Survey Methodology*. European Association of Methodology, Lawrence Erlbaum Associates, New York.
- Giesen, D. (ed.) (2011), Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes. BLUE-Enterprise and Trade Statistics, Small or medium-scale focused research project. (<http://www.blue-ets.istat.it/>)
- Giesen, D., Morren, M., and Snijkers, G. (2009), The effect of survey redesign on response burden: and evaluation of the redesign of the SBS questionnaires. Draft paper presented at the 3rd European Survey Research Association Conference 2009, Warsaw, June 29-July 3 2009.
- Goddeeris, O. and Bruynooghe, K. (2011), Administrative Simplification of the Structural Business Statistics. In: D. Giesen and M. Bavdaž (eds.), *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, Statistics Netherlands, Heerlen.
- Gravem, D., Haraldsen, G., and Löfgren, T. (2011), Response Burden Trends and Consequences. In: D. Giesen and M. Bavdaž (eds.), *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, Statistics Netherlands, Heerlen.

- Groves, R. M. et al. (2004), *Survey Methodology*. Wiley Series in Survey Methodology, New Jersey.
- Groves, R. M. et al. (1988), *Telephone Survey Methodology*. Wiley Series in Probability and Mathematical Statistics, New York.
- Haraldsen, G. et al. (2011), Utilizing Web Technology in Business Data Collection: Some Norwegian, Dutch and Danish Experiences. Paper presented at the New Techniques and Technologies for Statistics (NTTS) Conference, 22-24 February 2011, Brussels, Belgium.
- INE (2008), The response burden in business statistics - The Spanish experience. Paper presented at the 94th DGINS conference, Vilnius, 25th-26th September 2008.
- Lapeniene, V. (2008), Reduction of Data Collection Burden. Paper presented at the 17th Statistical Days, Radenci (Slovenia), 5-7 November 2007.
- Lepkowski, J. M. et al. (2008), *Advances in Telephone Survey Methodology*. Wiley Series in Survey Methodology, New Jersey.
- Orjala, H. (2010), Reducing the administrative burden in official statistics – Enterprise respondents in focus. Paper presented at the SIMPLY 2010 conference, Ghent 2nd-3rd, December 2010.
- Pereira, H. J. (2011), Simplified Business Information (IES) – Is coordination between public entities really possible? In: D. Giesen and M. Bavdaž (eds.), *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, Statistics Netherlands, Heerlen.
- Raymond-Blaess, V. (2011), Overview of measures used by NSIs to reduce response burden as reported in the literature between 2005 and 2010. In: D. Giesen and V. Raymond-Blaess (eds.), *Response burden measurement and reduction in official business statistics. A literature review of national statistical institutes' practices and experiences*, BLUE-ETS project, deliverable 2.1, 27–42.
- Savolainen, A. and Vertanen, V. (2007), Statistics Finland's measures to reduce enterprises' response burden. Paper presented at the Seminar, NordStat 2007, Reykjavik, June 2007.
- Snijkers, G., Onat, E., and Vis-Visschers, R. (2007), The Annual Structural Business Survey: Developing and Testing an Electronic Form. *Proceedings of the Third International Conference on Establishment Surveys*, American Statistical Association, Alexandria, VA, 456–463.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Main Module
2. Data Collection – Mixed Mode Data Collection
3. Data Collection – Techniques and Tools
4. Response – Response Process

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

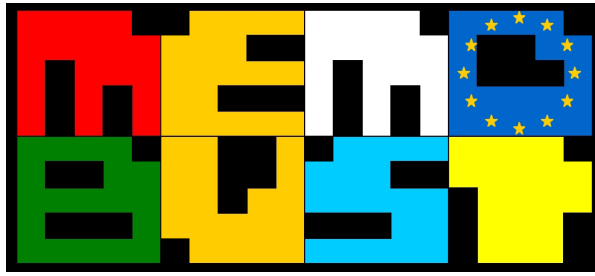
Data Collection-T-Design of Data Collection (Part 1)

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	25-01-2012	first version	Tora Löfgren	Statistics Norway
0.2	24-04-2012	second version – changes, added text and additional references according to review	Tora Löfgren	Statistics Norway
0.3	19-05-2012	third version – with some minor changes	Tora Löfgren	Statistics Norway
0.4	04-07-2012	fourth version with some minor changes according to review	Tora Löfgren	Statistics Norway
0.5	06-07-2012	last version	Tora Löfgren	Statistics Norway
0.6	21-11-2013	revised after EB comments	Tora Löfgren	Statistics Norway
0.6.1	21-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:48



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Quality and Risk Management Models

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 EFQM.....	3
2.2 ISO 9001.....	5
2.3 Code of Practice	5
2.4 ESS QAF	6
2.5 Object-oriented Quality and Risk Management model	6
3. Design issues	9
4. Available software tools.....	9
5. Decision tree of methods	9
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	11
Administrative section.....	12

General section

1. Summary

This module describes models for quality and risk management that are used in statistical institutes and their relevance for this handbook.

2. General description

Several models are used by statistical institutes for quality and risk management. These models include the following:

1. European Foundation for Quality Management Excellence Model (EFQM, 2013)
2. ISO 9001 (2008)
3. European Statistics Code of Practice (CoP) (Eurostat, 2011b)
4. Quality Assurance Framework of the European Statistical System (ESSQAF) (Eurostat, 2012).
5. Object-oriented Quality and Risk Management (OQRM) (Van Nederpelt, 2012)

In the next subsections we will describe these models briefly.

2.1 EFQM

The EFQM model distinguishes nine ‘criteria’:

1. Leadership
2. People
3. Strategy
4. Partnership & Resources
5. Process
6. People Results
7. Customer Results
8. Society Results
9. Key Performance Results

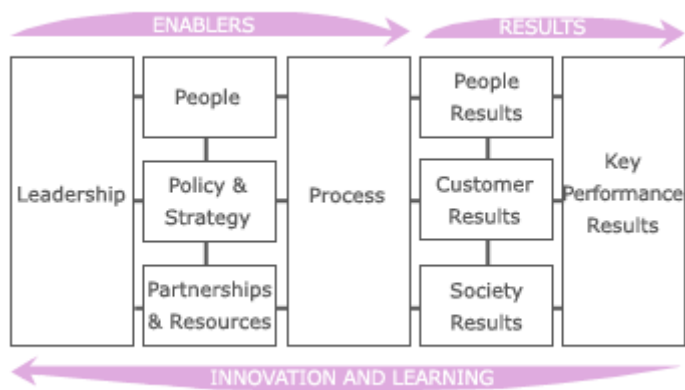


Figure 1. Nine criteria of the EFQM Excellence Model

These nine criteria are subdivided into sub criteria and recommendations respectively. To facilitate the integration of the prevailing ESS quality frameworks, namely the Code of Practice and the EFQM model, Eurostat mapped the Code of Practice on the EFQM Excellence model (Eurostat, 2005). Principal 7-15 regard statistical processes and statistical output and are relevant at the level of business statistics. The other principles of the CoP are at institutional level. As figure 2 shows, there is an overlap between indicator 7.1, 7.2, 7.7 (sound methodology), 8.5 (appropriate processes), 12.1-12.4 (accuracy and reliability), 15.1 and 15.2 (accessibility and clarity) and the EFQM Excellence model.

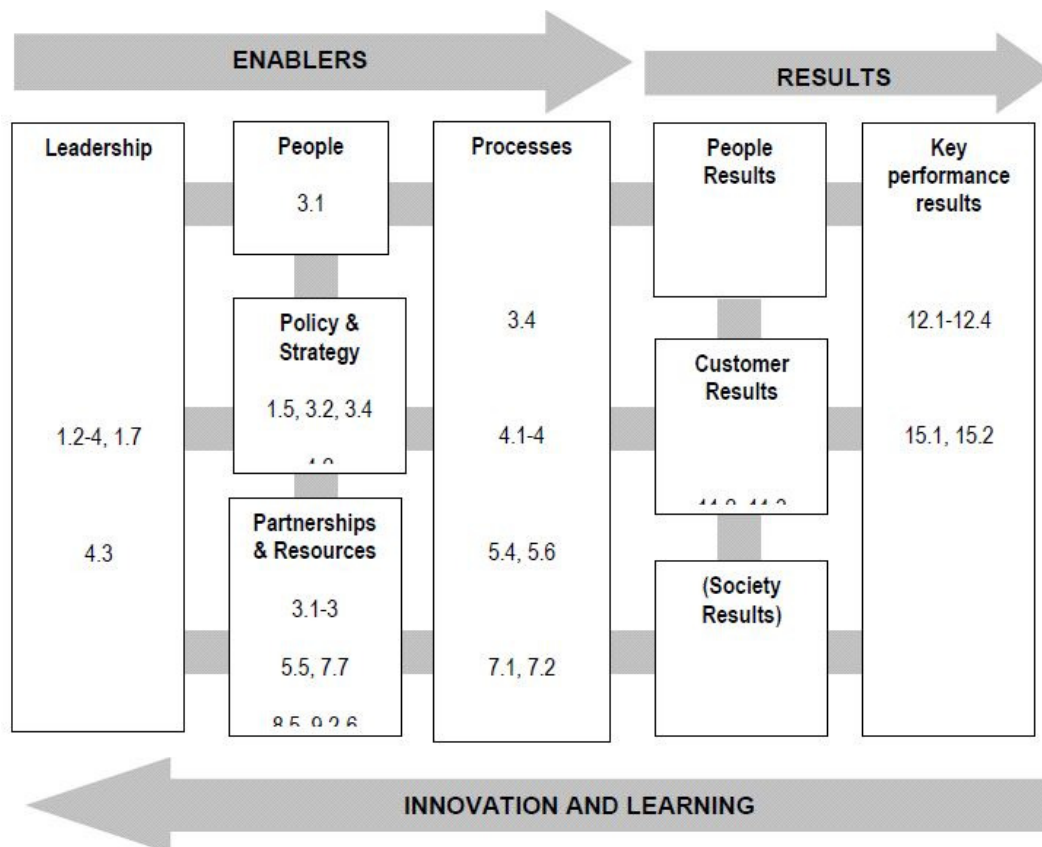


Figure 2. Mapping of the indicators of the Code of Practice on the EFQM Excellence Model

2.2 *ISO 9001*

ISO 9001 (2008) is a set of requirements for a quality management system (QMS). A QMS consists according to ISO 9001 of a number of elements. Some of them may be useful for business statistics such as:

- Requirements. Requirements related to the product are determined (ISO 9001, 2008, 7.2.1).
- Quality objectives. In planning product realisation quality objectives are determined (ISO 9001, 2008, 7.1.a).
- Characteristics of the product. The characteristics of the product are monitored and measured to verify that product requirements are met (ISO 9001, 2008, 8.2.4).
- Preventive action. Action is taken to eliminate the causes of potential nonconformities in order to prevent recurrence (ISO 9001, 2008, 8.5.3).
- Corrections and corrective actions. Necessary corrections and corrective actions are taken without undue delay to eliminate detected nonconformities and their causes (ISO 9001, 2008, 8.2.2).
- Records. Records of the result of the verification and any necessary actions are maintained (ISO 9001, 2008, 7.3.5).
- Training. Training is provided or other actions are taken to achieve the necessary competence (ISO 9001, 2008, 6.2.2).

ISO 9001 is widely accepted in the world of quality management and is applied by several NSIs too. However, the ESS QAF (Eurostat, 2012) is suitably adapted to quality management of statistics.

2.3 *Code of Practice*

The Code of Practice (Eurostat, 2011) is a set of 15 principles for statistical institutes that produce European statistics. These principles are divided in three categories: institutional, process and output. Each principle is subdivided into 'indicators'. The CoP is adopted by the European Statistical System Committee.

The principles of the CoP about output (11 until 15) distinguish eight quality dimensions of statistical output. These dimensions are generally recognised within the European Statistical System and are elaborated in the theme Quality of Statistics.

Principle 7 of the CoP about the soundness of methodology is especially relevant in the context of this handbook (figure 3) although the requirements are formulated at a high level of abstraction.

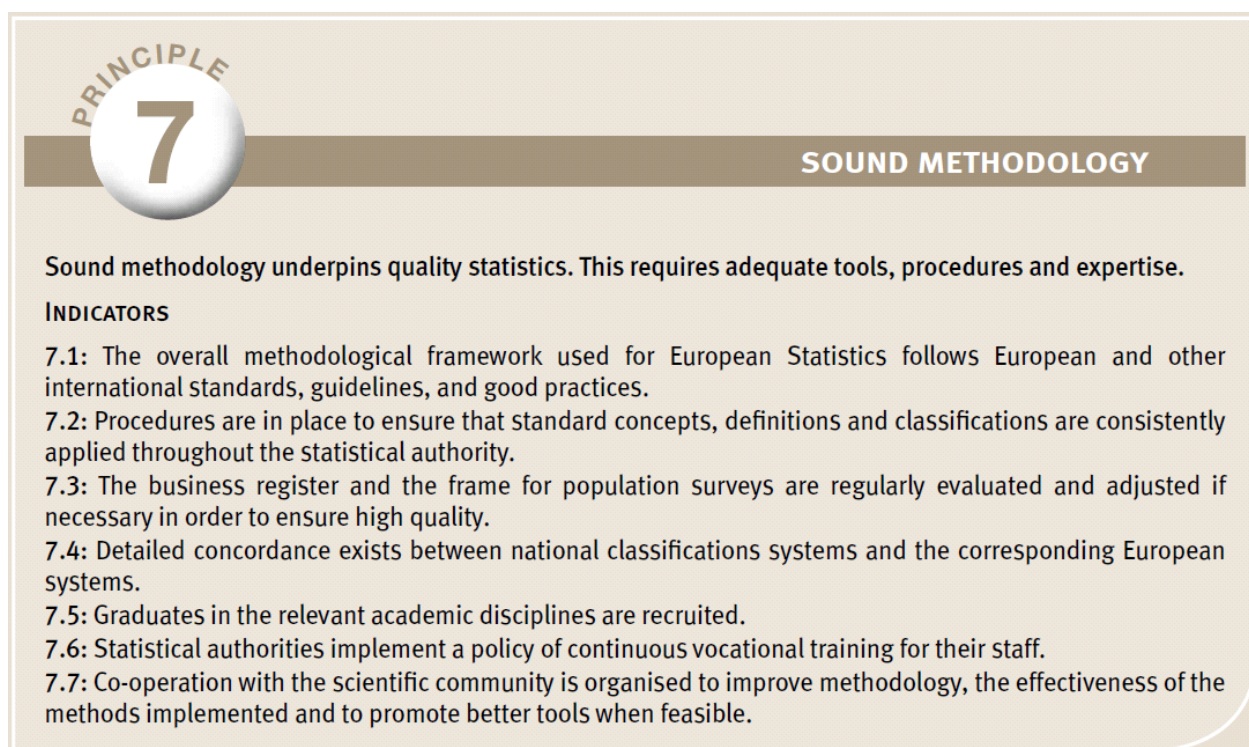


Figure 3. Principle 7 of the Code of Practice

2.4 ESS QAF

The ESS QAF (2012) is a set of good practices ('methods') for each indicator of the CoP (Eurostat, 2011). The QAF distinguished methods on institutional level and methods on process level. The methods on process level regarding principle 7 and 11-15 are relevant for business statistics.

2.5 Object-oriented Quality and Risk Management model

The OQRM model is meant to manage quality and risk. Important concept of the model that the quality of a product is dependent on the quality of other 'objects' such as processes, methods and data.

About the OQRM model and its applications, several papers have been published and a book (Van Nederpelt, 2012). The model has been presented at international conferences and can be used in any organisation, at any scale and in any field of expertise.

The OQRM model can be characterised as an 'empty' model. It is generic and does not contain any domain knowledge. The structure of the OQRM model is 'rich'. This model is, therefore, suitable for this handbook.

The model can be applied to objects like *output*, *processes* and *documentation* where an object is defined as everything that can be perceived or conceived. In this document, the model is only applied to the objects *statistical output* and *statistical methods*. The quality of *statistical output* is for a large part dependent on the quality of *statistical methods*.

Each object has a specific set of attributes (also called characteristics or quality dimensions). Each combination of an object and one associated attribute is called a focus area. An example of a focus area is *accuracy of estimates* where *estimate* is the object and *accuracy* the attribute. In this document

objects, attributes and focus areas are written in *italic*. The model also allows attributes that are not associated with quality like *costs*, *duration* and *capacity*.

OQRM defines the quality of an object as “the set of attributes of an object”. This is slightly different from the ISO 9000 (2005) definition: “the degree to which a set of inherent characteristics fulfils requirements”. The OQRM definition emphasises that attributes are associated to an object. It also uses the more general word attribute rather than characteristic. Finally, OQMR does relate quality to the required quality in separate steps in the model and not in the definition of quality.

A focus area is a unit that can be managed by taking the right measures (= actions, steps). The aim is to be or to get in control of a focus area. An organisation is in control of a focus area if the requirements for a focus area are met and/or an acceptable risk is taken regarding an objective like the quality of the *output*.

The OQRM model can be used to develop frameworks and to integrate existing frameworks. Requirement can be grouped by object respectively focus area. This improves the adaptability of the framework.

The model can also be used for quality assurance. It can help to find the right measures to control selected focus areas. By using the OQRM model, measures can be determined in a structural, analytical manner to get or to be in control of a focus area. This happens by taking the steps shown in figure 4. It can also be put that OQRM is a meta language that is valid in the area of quality and risk management.

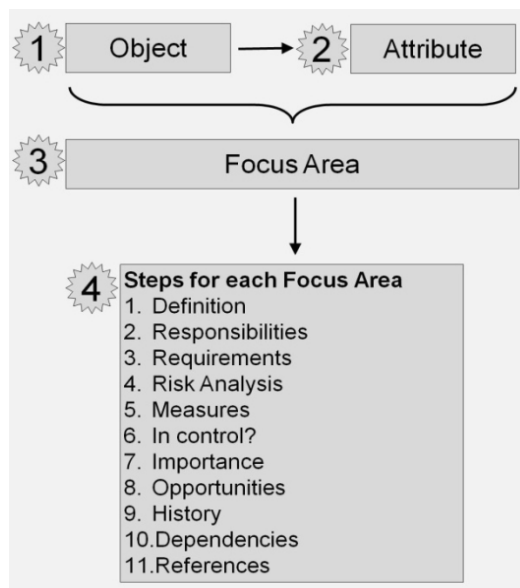


Figure 4. OQRM at a glance

Steps that will be applied in this document, are

- Step 1: Define what a focus area means.
- Step 3: Formulate requirements which a focus area must comply with.
- Step 4: Analyse the possible causes for problems with a focus area. Determine the possible effects of problems with a focus area.

- Step 5: Determine which measures have already been implemented, are still in progress or are already planned. We can distinguish four three of measures: signalling measures, preventive measures and curative measures.
 - A quality indicator is a signalling measure. It can detect a quality problem which can be followed-up by a curative measure to solve this problem.
 - A preventive measure is meant to avoid quality problems.
 - In case of a curative measure a quality problem has already occurred. The curative measure is meant to solve the quality problem.
- Step 7: Determine how important a focus area is for achieving certain objectives.
- Step 10: Determine what relationships there are with other focus areas.
- Step 11: Determine which references there are on a focus area.

The other steps (2, 6, 8, and 9) are less relevant for this handbook, because these steps are only applicable in a specific case.

2.5.1 Example of the application of the OQRM model

The OQRM model is in this subsection applied to the focus area *accuracy of estimates*. In this case *estimate* is the object and *accuracy* the attribute. Each step of the model is elaborated.

Step 1: Definition

Accuracy of an estimate is defined as closeness of computations or estimates to the exact or true values that the statistics were intended to measure (SDMX, 2009). Accuracy includes both bias and variance.

Step 3: Requirements

The CoP (principle 7) requires that “European Statistics accurately and reliably portray reality”. In practice this requirement could be defined more specifically like the mean square error is less than a specific value.

Step 4: Risk analyses

Problems with the *accuracy of estimates* are caused by errors. Examples of these errors are sampling errors, measurement errors and processing errors. In case of secondary data collection errors in the data sources can be added too. Another category of source of errors are inadequate methods. We will call this ‘method related errors’. This category is relevant in the context of this handbook.

Theoretically, in each step of the statistical process errors can occur. In the logistical and publication process errors can happen too. Originators of all kind of errors are the statistical agency, respondents and data suppliers.

Step 5: Measures

Possible measures to manage or control *accuracy of estimates* are multiple. An important measure is to implement the right *methods* and to implement the *methods* right.

Another possible action is to measure the quality of the *data*. The ESS Committee determined and elaborated a set of quality indicators (Eurostat, 2011a) concerning the *accuracy of estimates*. These quality indicators are meant for reporting purposes but can be used to improve the quality of the output too.

1. Sampling error – indicator
2. Over-coverage – rate
3. Unit non-response – rate
4. Item non-response – rate
5. Imputation – rate
6. Common units – proportion
7. Data revision – average size

Not all indicators are related to a method (2, 3, 4, 6 and 7). These indicators also do not cover all categories of errors, e.g., assumption errors and method related errors.

Other possible actions to assure the required *accuracy of the estimate* are to develop good software and test it, hire competent staff and train them, put monitoring processes in place, check the quality of the data sources. These measures are not elaborated in this handbook.

Step 7: Importance

There is no doubt that *accuracy of estimates* is a very important focus area for a statistical agency. It will certainly be an objective of a statistical agency to compile statistics that are sufficiently accurate.

Step 10: Dependencies

Accuracy of estimates is dependent on the quality of a list of objects, i.e., *data sources, processes, methods, software and staff*. In the module “General Observations – Methods and Quality”, we will focus on the object *method*.

Step 11: References

References about the focus area *accuracy of estimates* can be found in the reference list (SDMX, 2009; EU, 2009; Eurostat, 2011b).

- 3. Design issues**
- 4. Available software tools**
- 5. Decision tree of methods**

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

EFQM (2013), *EFQM Excellence Model 2013*. EFQM, Brussels, Belgium.

EU (2009), Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities. Also referred to as “StatLaw”.

Eurostat (2005), Mapping of intersections between the European Statistics Code of Practice, the LEG on Quality recommendations and the EFQM Excellence Model Criteria.

Eurostat (2011a), Final Report from the Sponsorship on Quality. 10th Meeting of the European Statistical System Committee. ESSC 2011/10/05/EN. Wiesbaden, 28 September 2011.

Eurostat (2011b), *European Statistics Code of Practice*. Adopted by the European Statistical System Committee, 28th September 2011.

Eurostat (2012a), Quality Assurance Framework for the European Statistical System (ESS QAF).

Eurostat (2012b), *Eurostat's Concepts and Definitions Database*. Website:

http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GL_OSSARY&StrNom=CODED2&StrLanguageCode=EN

ISO 1179-1 (2004), *ISO/IEC-FDIS 1179-1. Information technology – Metadata registers – Part 1: Frameworks*. International Organization for Standardization, Geneva.

ISO 9000 (2005), *ISO 9000:2005. Quality management systems – Fundamentals and vocabulary*. International Organization for Standardization, Geneva.

ISO 9001 (2008), *Quality management systems – Requirements*. International Organization for Standardization, Geneva.

Longman (2010), *Dictionary of Contemporary English*. Fifth Edition, third impression. Pearson Education Limited, Essex, England.

NQAF (2012), Glossary. Compiled by the Expert Group on National Quality Assurance Frameworks. 3 February 2012.

SDMX (2009), SDMX Content-oriented Guidelines. Annex 1: Cross Domain Concepts.

Van Nederpelt, P.W.M. (2012), *Object-oriented Quality and Risk Management (OQRM). A practical, scalable and generic method to manage quality and risks*. MicroData, Alphen aan den Rijn, The Netherlands. Website www.oqrm.org/English.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Methods and Quality
2. General Observations – Logging
3. Quality Aspects – Quality of Statistics

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

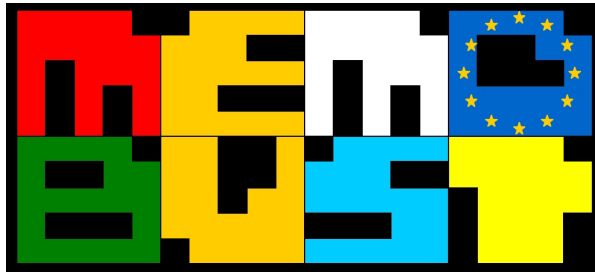
General Observations-T-Quality and Risk Management Models

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	02-10-2013	first draft	Peter van Nederpelt	Statistics Netherlands
0.1.1	31-01-2014	all EB's comment d.d. 22 January 2014 processed	Peter van Nederpelt	Statistics Netherlands
0.1.2	04-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:21



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Design of Data Collection Part 2: Contact Strategies

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Choosing the appropriate contact strategy.....	3
2.2 Defining the respondent	6
2.3 Specific treatment of new enterprises.....	7
2.4 Specific treatment of large enterprises	8
2.5 Feedback to enterprises	8
2.6 Responsive design issues.....	9
3. Design issues	10
4. Available software tools.....	11
5. Decision tree of methods	11
6. Glossary.....	11
7. References	11
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

This module deals with setting up a contact strategy to reach respondents. It deals with several sub-themes; the first part deals with choosing a contact strategy regarding when to send material to the respondents and what material to send, strategies for reminders, using penalties and fines et cetera. The second sub-part deals with finding the right respondent at an enterprise. These first two parts cover the whole of the sample, while following sub-parts deal with specific treatment of sub-populations. The third sub-part covers whether or not to treat new enterprises in any specific way, and the fourth sub-part covers specific treatment of large enterprises. Since giving feedback to respondents is a way of encouraging participation, a specific sub-part is devoted to that area. And finally, the last sub-part deals with making changes to the design during the data collection based on the outcomes so far, i.e., responsive design. Together with the previous module (“Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method”), the module gives an overview of the decisions to make before starting the actual collection process.

2. General description

2.1 *Choosing the appropriate contact strategy*

When the appropriate data collection mode(s) has been chosen, the next step is to choose an appropriate contact strategy. A contact strategy consists of when and how respondents are contacted, and what material (questionnaire, cover letter, instructions et cetera) is used in each contact. As was stated above (see “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method” regarding mixing modes), it is in many cases wise to exploit a data collection mode to the maximum before switching to another mode, i.e., using a sequential mixed mode approach. Another possibility when mixing modes is to utilise a parallel mixed mode design, giving the respondent the choice to choose his or her preferred mode. Some studies have shown that the choice of mode in itself can have a negative effect on response rates. This has been shown for voluntary surveys, where the most important question for the respondent at the initial stage is whether to participate or not in the survey. Making this choice more difficult by offering several possible ways to provide data may affect the willingness to participate in a negative way (not participating may seem as the easiest choice). For mandatory surveys, such negative effects are not as common. Since the respondent has to provide the data, being able to choose the best way to do it may have a positive effect on providing timely high quality data.

The choices to make regarding contact strategy are the following:

1. Timing of the first sending
2. Time given to respond
3. Material to include in the first sending
4. Number of reminders
5. Timing of reminders
6. Material to include in reminders

7. Penalties/Fines

Timing of first sending and time given to respond

Of course, the timing of the first sending is dependent on when the data collection needs to be finished for other processes to have appropriate time before the survey needs to be completed, and also on some of the other choices (number of reminders, using force), but it is also in many cases dependent on data availability. Many surveys (for example most monthly economic surveys) ask for very recent data, as soon as they are available. For example, economic data on revenues and costs, salary data, employment data et cetera are normally registered in the administrative systems at enterprises. But in many cases, there is a need to extract data in the form of reports to be able to provide them to statistical offices. For accounting data, there may be additional work done in order to finish an annual report or such, and there may be internal and external rules on when data may be released and to whom. External rules may even be legislated. The general advice is to consider what data the survey covers, examine when data may be available for statistical purposes and adapt the data collection period and the timing of the first sending after that. Normally, the best time for the first sending is just when data has become available. For an annual survey, it is normally possible to adjust the collection period to when data are available. However, for some short term statistics with a tight production schedule, this might interfere with the second point, time given to respond. It might then be better to send out the request a short time before data are available, so the respondents can prepare for providing data and are ready when data become available. It is also important not to send out the questionnaire too late. Data are in many cases archived at enterprises at specific times, so it is important not to send out questionnaires asking for old data and most importantly not to ask for data that have recently been archived, since that increases the risk of respondents misreporting current data instead. For example, accounting data are often archived one year after the accounting year, so that what are readily available at a specific time are data for the current and the previous accounting year, while older data would have to be retrieved from archives.

When asking for old data, it might be necessary to adapt the given time to respond to the survey, but this issue is important to consider in all surveys. The time given to respond should not be too short, since the respondent must be given a fair chance to fulfil the requests of the statistical office, especially if the survey is mandatory. But the time given should neither be too long, since that increases the risk of respondents forgetting about sending data, and also because it sends a signal to the respondents that the survey is not very important. Normally 10-14 days response time is appropriate for short term statistics, while a response time of 3-4 weeks, but not much longer than that, is appropriate for annual or one-time surveys.

First sending

The decision on what material to include in the first sending is based on the choice of modes to use:

Paper questionnaire: Given that the choice is a mail-out paper questionnaire, the material that is sent is the questionnaire itself and a cover letter describing the purpose of the survey, the due date et cetera. In most cases, a pre-paid return envelope with the sending address pre-printed is included so that no additional cost except the response burden is put on the respondent. If a parallel mixed mode approach

is chosen using paper as the main mode, the cover letter or the questionnaire contains a link to reply to the survey by web.

Web survey: For a pure web survey, what is sent is an e-mail with the description of the survey and everything needed, including a link to the survey and, if appropriate log-in information such as user-ID and password. However, a survey that is run by a National Statistical Institute and uses web as the primary mode can also choose another way – to use a mail-out sending for a survey with web as the response alternative. In that case, the sending consists of only a cover letter, including a web address to enter into the respondent's browser and, if appropriate, user-ID and password to log in. This method makes it easier to reach all respondents, since registers of postal addresses are normally more complete and more up-to-date than registers of e-mail addresses. This method has proven effective in many countries, and is recommended for all surveys where you don't have regular and close contact with all respondents (monthly surveys where enterprises selected remain in the sample over a longer period of time being the exception).

CATI and CAPI: For an interviewer-administered survey, the initial sending is a pre-notice letter that tells the respondent that the survey has started and that he or she will be contacted by an interviewer. In an interview survey, no additional material to this letter is needed.

Regarding pre-contacts with respondents in other surveys than interviewed-administered ones, it is sometimes done for new respondents to ongoing surveys. This is described further in Section 2.3 below.

Reminders

Enterprises that haven't sent their data by the due date need to be reminded. There are several methods to use for reminders:

- Ordinary mail
- E-mail
- Telephone

In CATI/CAPI surveys there are no reminders as such, rather a number of contact attempts to reach the respondents.

The number of reminders to use is a delicate matter, especially in voluntary surveys. On one hand, a voluntary survey that is not given a high priority by an enterprise may need many reminders to have a good enough response rate. On the other hand, sending too many reminders angers respondents, who may feel that not responding is the way to show that they will not participate in the survey. Very few respondents actually contact the statistical office to notify the survey that they are not willing to participate. Two or three reminders is the common practice in most countries. The last reminder may also be targeted at specific groups or strata where the response rate is low.

The timing of reminders is also to be decided. If the survey has a due date, which is common practice in enterprise surveys, the first reminder should be sent right after the due date. Waiting a long time after the due date to send a reminder may be interpreted by respondents that the survey was not that important, that the need for data is not that big. This may in turn have a negative effect on response rates. In short-term surveys, especially monthly surveys, there might actually be such a tight production schedule that the first reminder must actually be sent before the due date. This is normally

done by sending a “thank you and reminder” card or letter, thanking those respondents that have already sent data while at the same time reminding the rest that the due date is approaching.

A final decision regarding reminders is what material to include in reminders, this is of course decided by the chosen mode. The decision between ordinary mail and e-mail is the same as when considering the first sending. Of course, if the first sending was done by e-mail, reminders should normally also be sent by e-mail, unless a reminder includes a paper questionnaire, in which case ordinary mail is necessary. Including a paper questionnaire as an attached file in an e-mail is also an option, but it may be difficult to manage the practicalities if the questionnaire includes individualised information like pre-printed data or such. If reminders are carried out by phone, no material is included. The contact strategy may also involve a mixing of modes at the reminder stage as well, using for example mail or e-mail for some enterprises and phone for others.

Using penalties or fines

When the previously described ways of managing reminders do not work, country legislation in many countries gives Statistical Offices the possibility to use the force of the law to ensure that responses to mandatory surveys are received. In this case, the forms and methods available differ between countries, but may involve fines or other penalties for non-respondents. Therefore, it is difficult to make a general description of how this process should be run. However, it can be said that if it is up to the Statistical Office to use the force of the law or not, it is important to actually use it, to show non-respondents the importance of complying with mandatory requirements. If, for administrative reasons, cost reasons or other reasons it is not possible to use this force against all non-respondents, it is most important to use it against new respondents to ensure their future cooperation, and also to the enterprises that are most important to the quality of the survey (i.e., the largest enterprises).

2.2 *Defining the respondent*

One very important aspect in getting high quality data from enterprises is to be able to reach the best respondent within the enterprise. In small enterprises, this is normally not a big problem, since there are only a few employees, so the chance of reaching a relevant respondent is high. On the other hand, small enterprises often have their accounting done by a service bureau or similar, meaning the people within the enterprise might not always be able to respond to a survey themselves, without consulting the service bureau. But for small enterprises, the best way to get good responses is to address the survey to the enterprise itself. For larger enterprises, it might be more efficient to address a questionnaire to either a specific person or a designated role. The advantage of doing so is minimising the risk that the questionnaire is either held up by a “gatekeeper” within the enterprise – for example a person who opens all mail, but does not know whom to give a specific questionnaire to – or given to the wrong person within the enterprise. If the questionnaire is filled in by the “wrong” person – i.e., someone who is not suitable to respond based on the contents of the questionnaire – the risk of having measurement errors due to misunderstood questions et cetera is increased. Another risk in large enterprises is that if a questionnaire is first given to the wrong person and that person realises he or she is not the suitable person, it takes extra time for that person to find a suitable person in turn. This risks late responses or non-response to the questionnaire floating around within the enterprise. However, addressing a questionnaire to a specific person or a designated role may also have disadvantages. Starting with a specific person, this is normally the person who answered the questionnaire last time. This option is only available in ongoing surveys. If the last time was a year, or sometimes even three

months ago, the person may no longer be the most suitable person. He or she may have changed position (in which case he or she might be suitable in finding the new suitable person, but may not have the time to do so based on the new commitments) or may even have left the company. In the last case, this may lead to the enterprise sending back the envelope without even opening it, meaning extra effort to re-send the questionnaire to someone else. Sending to a specific person may therefore only be recommended in short-term surveys where the statistical office is in continuous contact with most respondents in the survey, or in surveys where the advantage of having the same contact over time is very large, e.g., when the content is extremely complicated. Sending to a designated role (e.g., “head of economic division”) may be an alternative to simplify the work of a gatekeeper within an enterprise to find a suitable person. However, this may also be risky. It is difficult to find role names that are known to all enterprises. If the role name chosen does not exist within an enterprise, the enterprise may simply send the envelope back. Otherwise, they will often have to translate the given role to something similar within their enterprise, the advantages of addressing questionnaires to a specific role can therefore be considered small. If a questionnaire is suitable for a specific type of person, it is probably better to state clearly in the cover letter what kind of data the questionnaire asks for and what type of person is a suitable respondent.

If the survey is large and covers several subject matters, it might not be possible for one person to fill in the whole questionnaire, he or she might need help or input from other people within the enterprise. This is more often the case in large enterprises than in small ones. If this is the case, it is good if the questionnaire is designed so that dividing the task is easier. This can be done by specifically putting parts that might need additional respondents on separate pages, and by allowing respondents to electronic questionnaire to save the data without sending and resume the questionnaire at a later time.

2.3 Specific treatment of new enterprises

Especially in an ongoing survey where the design normally is such that a sampled unit will be included in the sample over a period of time before rotating out of the sample, it might be worth to consider treating new enterprises in the sample a little differently. This can be considered not only in the data collection process, but also in the processing stage where perhaps they could be edited in another way than other respondents. But at the collection stage, if the questionnaires are sent to a specific person (see Section 2.2), no such person is available for new enterprises. One alternative in such a case is to have a pre-contact with new enterprises in the survey. The pre-contact can be done by phone or by sending new enterprises a specific letter before the collection starts to ask for a suitable person to send the actual questionnaire to. If the questionnaire contains specific data that have to be extracted from accounting systems (or similar) or requires specific actions from the respondents (calculations, estimations, perhaps even changing the accounting system to keep track of the data requested), a pre-notice letter or pre-collection telephone contact can also include such information and instructions. However, if the last example is valid, i.e., the enterprise has to make adjustments to the accounting system or in other ways arrange for data to be stored to make filling in the questionnaire possible, such a pre-contact must be done well in advance, so that the enterprise has a reasonable possibility to take action. Since sampling is often done close to the collection sending pre-notice letters or phoning so long in advance is not possible. So in practice, pre-contacts should mostly be used to find contact persons and inform about an upcoming survey. Since additional letters also increase the response burden on enterprises (it takes time to read a letter, react to it and take action) it

is recommended not to use them for easy surveys, they can only be motivated where data provision can be considered difficult or require special knowledge.

2.4 Specific treatment of large enterprises

Since large enterprises are more important than small enterprises in most business surveys with regards to quality, and since the largest enterprises are sampled in most surveys, many countries have decided to devote special attention to large enterprises. Some of the issues that this treatment includes are the following:

- Providing a single point of contact within the Statistical Office for the largest enterprises. This is both for when the enterprises need contact with the Statistical Office and vice versa.
- Building relations with the respondents within the largest enterprises.
- Profiling the largest enterprises with regards to statistical units – enterprise group, legal units, kind-of-activity units, local kind-of-activity units and so on.
- Helping surveys finding the right contact person within the largest enterprises for a specific survey. This can be aided by identifying functions and roles used in each of the largest enterprises. In some countries, these roles and functions are surveyed specifically and registered for future use.
- Helping the collection and editing staff with contacts with these enterprises.
- Helping the enterprises with support for their data provision.
- Enterprise-specific arrangements in data provision, such as specific questionnaires containing data from several surveys, aid in how to estimate difficult figures and such.
- Coherence analysis of data sent to different surveys.

In most countries that have tried a specific treatment of large enterprises, a special organisational group of people has been created. Within the group, one person deals with a specific number of enterprises, normally the same enterprises over time. This helps to build competence about the enterprise and its data, and it builds a relation with the enterprise. The enterprises themselves appreciate having a single contact person et cetera. Normally, one person deals with between 3 and 10 enterprises, and the group normally consists of 5-10 persons. The number of enterprises that should be covered by such a treatment may differ between countries, but in most countries it would be true to say that covering the 30-50 largest enterprises will cover a large part of the economy.

2.5 Feedback to enterprises

One final thing to consider when designing the data collection is whether to give respondents feedback on their responses or not. Here, we are not talking about a simple receipt that data have been received, which is normal procedure in electronic collection, but rather something more, using the actual data. Feedback can be of several types:

- General feedback, e.g., finalised figures and tables from the survey
- Specific feedback, using the actual data provided by each enterprise to provide an individualised feedback, maybe even comparing the enterprise to similar ones. Examples of such figures are key ratios or market shares.

Giving feedback can have several positive effects – it can show how the information is used by society and what is produced from the respondents' data, it can be, in some cases, of direct use to the respondents themselves and it can lower the perceived burden of providing the data. So it can be recommended to try to give feedback whenever possible. But feedback can be designed in several different ways.

In general, specific feedback is probably more efficient than general feedback, since it is based on the enterprises' own figures and could be of more use to the enterprise itself. The most attractive form of feedback is probably feedback directly after transmitting data. This requires that the transmitted data are immediately transformed into something nice-looking and sent back to the respondent. If this is possible, it can have very positive effect. But it is often not possible to use current data from other enterprises to make comparisons in such a feedback. Then, in this case, a good alternative could be an interesting feedback created by combining current data from the specific respondent and data from previous rounds for the whole sample, such a feedback can be used. But such feedback would have to be judged against better individualised feedback at a later stage. Therefore, if it is possible to create a better feedback by using current data but it is not possible to create that feedback immediately, then it is better to send the feedback at a later stage, after the collection is completed. When such feedback is sent later, it is good to allow respondents to choose (for example by ticking a box in the questionnaire) whether they want to receive the feedback or not, and also to give them the alternative to direct the feedback to a different person than the respondent.

If specific feedback based on data from the individual respondent is not relevant for a specific survey, or if it is very difficult to create, more general feedback based on the whole sample can be considered an alternative, it is still better than nothing. Another version of this is to give feedback of the data from last survey round when the request for data for the current round is sent out. This can result in a higher response rate or a lower perceived burden in itself.

2.6 *Responsive design issues*

Responsive design, which is a special case of adaptive design (Schouten et al., 2013), is a term that is used for modifying the design while the collection process is running. Of course, this is not done in a random manner, but rather following pre-specified rules. When statisticians talk about responsive design, they mostly refer to household surveys using CATI or CAPI modes. In those surveys, a lot of adaptations to the contact strategy can be made based on response rates in different strata, age groups et cetera, and those changes can have a profound effect on the actual collection work, since it deals with when different cases should be called and how many contact attempts should be made. For business statistics, responsive design is not discussed as much, partly because most business surveys use self-administered modes and also since, when CATI/CAPI is used, contacts with enterprises are limited to business hours. However there are still some adaptations to the design that can be made during the collection phase in business surveys:

- Utilising different modes – for example, in strata where the response rate is low, it might be considered to use a different, more expensive mode, than for other strata. This can be using CATI instead of paper/web questionnaires, CAPI instead of CATI et cetera.

- Utilising different reminder strategies – for example deciding which enterprises should get an e-mail reminder, which should get a reminder by paper mail and which should get a reminder by telephone.
- Utilising penalties/fines. In most cases, penalties/fines are not used for all non-responding enterprises, a responsive design approach can be used to decide for which enterprises penalties/fines will be used.
- Utilising different versions of the questionnaire. Some business surveys use different versions of the questionnaire, for example a longer questionnaire for larger enterprises and a shorter version for smaller enterprises. Responsive design can be used to decide if some enterprises should receive the shorter version instead of the longer one at the reminder stage.

In order to use a responsive design approach, the two following things must be done:

1. Set up decision rules for action. This involves several steps;
 - a. deciding on which different paths that can be taken (different modes, reminders, questionnaires, penalties et cetera, see above),
 - b. deciding which sub-groups to analyse (different strata or other groups)
 - c. deciding the thresholds when action is to be taken
 - d. deciding when decisions will be taken

The four points a-d will then resign in a set of specific rules, e.g., “At date X, all enterprises with a size over Y in all strata with a response rate of Z or lower will be given a telephone reminder instead of the planned e-mail reminder”. Of course, parameters like Y and Z can also be relative, compared to other enterprises or groups. For example, Z could be something like a response rate that is a percent lower than the average response rate in all strata. There has been some work done to come up with more advanced indicators that take response bias and the like into account, for example at Statistics Canada and Statistics Netherlands. Those indicators have mostly been designed for household surveys but there is some experience in business statistics as well. At the time of writing, it is too early to make a general recommendation on more advanced indicators.

2. Set up measurements. In order to utilise responsive design, there must be paradata available that can be used to make the decisions based on the rules that were set up.
3. Measure and take action. As the collection moves along and the decided measurement points are reached, measurement is made and appropriate action taken. This may also involve adjusting parameters like Y or Z in the example above, especially if it turns out that the action required is too costly compared to the survey budget.

It should be noted that responsive design is a relatively new concept that is still being developed within many statistical offices, and it is not yet possible to make general recommendations as to when and how to use it or not.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

De Leeuw, E. D., Hox, J. J., and Dillman, D. A. (eds.) (2008), *International Handbook of Survey Methodology*. Lawrence Erlbaum Associates, New York.

Schouten, B., Calinescu, M., and Luiten, A.(2013), Optimizing quality of response through adaptive survey designs. Component of Statistics Canada Catalogue no. 12-001-X Business Survey Methods Division – June 2013.

Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D. K. (2013), *Designing and Conducting Business Surveys*. Wiley, Hoboken, NJ.

Interconnections with other modules

8. Related themes described in other modules

1. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

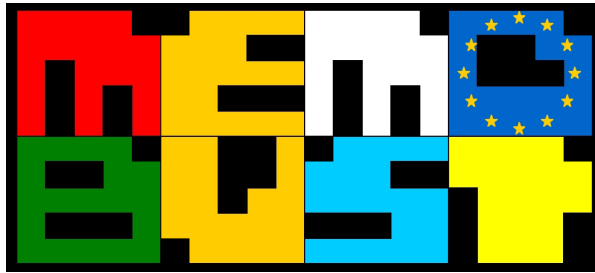
Data Collection-T-Design of Data Collection (Part 2)

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	17-02-2012	first draft	Johan Erikson	Statistics Sweden
0.2	19-06-2012	second draft after first review	Johan Erikson	Statistics Sweden
0.3	27-06-2012	revised after second review	Johan Erikson	Statistics Sweden
0.4	30-10-2013	revised after EB comments	Johan Erikson	Statistics Sweden
0.4.1	21-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:49



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Logging

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Purposes of logging	3
2.2 Logging indicators.....	5
2.3 Quality of logs and log information	5
2.4 Example of logging by τ - and μ -ARGUS	6
3. Design issues	6
3.1 Beforehand or afterwards	6
3.2 Structure of the log.....	6
3.3 Presenting log information	7
4. Available software tools.....	7
5. Decision tree of methods	7
6. Glossary.....	7
7. References	7
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

Logging is the activity of producing log information in a log. Log information is used to manage a statistical process and can serve various purposes. These purposes should be determined before implementation of logging. In this theme, we will define logging, logs and log information, and describe various possible areas of application of logging and possible technical solutions for logging.

2. General description

Log information is metadata produced during a specific run of a process. It includes all types of information such as the creation data of an output file, version of the software application that is used and the number of records that is processed. The definition of log information includes quality indicators. So, the definition of log information is quite broad.

Log information can be generated automatically or registered manually. Log information is not metadata that regards the statistical process in general such as process descriptions, methodological documents and descriptive metadata. Statistical output and intermediate results are not considered log information either because these are data and not metadata.

This section discusses the purpose of logging, logging indicators and the quality of logs.

2.1 *Purposes of logging*

The implementation of logging is dependent on the purpose of logging. Log information is used to manage the quality of the statistical process and output. Logs should be followed up by actions such as validating data (GSBPM sub-processes 5.3 en 6.2), reporting or even improvement of the process, method or system (GSBPM, 2009).

Purposes of logging can be related to the next factors:

1. Punctuality of the output
2. Accuracy of statistical output
3. Traceability and reproducibility of statistical output
4. Statistical confidentiality of statistical output
5. Multiple quality dimensions of statistical output
6. Efficiency of the process

Logs should be designed at the same stage of the design of the statistical methodology and the production system (GSBPM, sub-processes 2.5 and 2.6). Logs are an instrument to manage quality and can be used in validation processes (GSBPM, sub-processes 5.3 and 6.1).

In each subsection, we will describe for each factor how logging can be implemented.

2.1.1 *Logging related to punctuality of the output*

Some statistical processes have a tight time schedule. If the time to process a dataset takes a few hours or even days it is helpful to know why this process takes so much time. Logging of the performance of

the software that processes statistical data can be used to get insight in possible bottlenecks in the software or database.

2.1.2 Logging related to accuracy of the statistical output

The accuracy of the statistical output is dependent on various factors as mentioned below.

- a. Completeness of the units in a micro-dataset
- b. Validity of data in a micro-dataset
- c. Ability to statistically match units of two datasets

Ad a: The *completeness of units in a micro-dataset* can be measured by logging the number of records in a dataset and compare this number by an expected number. If units are imputed because of incompleteness data about these imputations can be logged too. Several methods for imputation are elaborated in this handbook; see “Imputation – Main Module”.

Example: The number of business units in a dataset is 38,000 while 45,000 units were expected.

Ad b: *Validity of data in a micro-dataset* can be checked by applying specific rules (constraints) to the data. This process includes checking on missing values and outliers. Wrong data will be edited. Apart from logging the old value of a variable the violated rule can be recorded. This last information can be used for analysing purposes and as input for improvement of the statistical process. Several methods for editing are elaborated in this handbook; see “Statistical Data Editing – Main Module”.

Example: The return of a business unit is Euro 2 million while the number of employees is 400.

Ad c: The *ability to statistically match units of two datasets* can be measured in the process of statistical matching two dataset. Matches and unmatched units can be logged. The ratio between matches and unmatched units is an indicator for the ability to match two datasets. Mismatches (false matches) are, however, hard to discover and cannot be logged. Several methods for matching (or record linkage) are elaborated in this handbook; see “Micro-Fusion – Object Matching (Record Linkage)”.

2.1.3 Logging related to traceability and reproducibility of the statistical output

If traceability and reproducibility is required, it is necessary to log the version of the datasets and the version of the software that are used. Moreover, if data are edited or imputed manually it is necessary to log the number of edits and imputations too in order to be able to reproduce the statistical output.

2.1.4 Logging related to statistical confidentiality of statistical output

While analysing the statistical confidentiality, the results of the analysis can be logged. The log could report which details should be or are left out or which data should be or are changed. The log report created for statistical confidentiality is confidential information and should be treated as such. It is only for the NSI concerned, and is accessible by a limited group of persons within the NSI only.

2.1.5 Logging related to multiple dimensions of statistical output

Quality indicators can be regarded as logging indicators. Indicators can cover multiple quality dimension of statistical output and processes. It depends on what indicator is selected.

Example: Quality indicator *item non-response* is an indicator for the accuracy of the output.

2.1.6 *Logging related to efficiency of the process*

Logs can be used to improve the efficiency of the process.

Example 1: Files can be found more easily as path and filename are logged.

Example 2: Staff capacity needed to run a process manually and automatically can be logged. This log information can be used to analyse if more or less staff should be or can be assigned to the process.

2.2 *Logging indicators*

There are a number of items that can be logged. It depends on the purpose of the log which items are relevant. Examples of these items are:

- Version of the software
- Version of a file
- Start and completion date and time
- Path and file names of a file
- Time used to create output
- Number of processed records
- Script files. These files make it possible to check if the right procedure is followed.
- Method and rules that has been applied.
- Tuning parameters for a method. Tuning parameters are specified in section 14 of each method module in the handbook.
- Flags: yes/no or more values. A flag indicates for example if a record is edited manually.
- Quality indicators. Quality indicators are specified in section 21 of each method module in the handbook.
- Violated rules: identification of the rule, frequency of violation.

2.3 *Quality of logs and log information*

Logs and log information should have the right quality too. Quality dimensions of logs and log information are (see the section on the OQRM model in the module “General Observations – Quality and Risk Management Models”):

- Relevance of log information. Log information should be useful and serve a purpose to be relevant.
- Completeness and correctness of log information. If log information is not complete or correct, it could even effect the quality of the statistical output in a negative way.
- Clarity of log information. Log information should be clear to be understood and efficiently used by the user of the log information.

- Accessibility of logs. It should be clear who is authorised to access specific logs.
- Confidentiality of log information. Some logs are confidential such as logs related to statistical confidentiality.

2.4 *Example of logging by τ - and μ -ARGUS*

τ -ARGUS is a software program designed to protect statistical tables. μ -ARGUS creates safe micro-data files. Both programs produce logs. The “ARGUS Report” contains the following log information (Argus, 2013):

- Creation date of the output table
- Path and filenames of the input files and output file
- Table structure
- Safety rules used
- Time used to protect the table
- Summary of the table, e.g., safe and unsafe cells.
- Version of the software

Separately, τ - and μ -ARGUS both produce a technical log (logbook.txt) that reports the following log information:

- Start date and time of the run
- Version of the software
- Structure of the input table
- Path and filename of the input file
- Start and completion statement

For τ -ARGUS, it is optional to produce a script file that can be used to rerun the program.

3. **Design issues**

3.1 *Beforehand or afterwards*

Logging is preferably developed in the design phase (GSPM sub-processes 2.5 and 2.6). However, it could be necessary to produce a log afterwards on an ad hoc basis. A log can be produced, for example, by comparing two files and log the differences. This method can always be used as a last resort if there are no relevant logs available.

3.2 *Structure of the log*

Logs can be designed as a structured file or as text. In case of a structured file, there are three ways to structure a log file:

- Add a separate file (log) for log information

- Add extra fields to the existing file with statistical data for log information

3.2.1 Separate files

A separate file is useful if the logs concern the dataset as a whole and not separate units in a dataset.

3.2.2 Extra fields

Extra fields are useful if the log information concerns each unit in the dataset. A ‘flag’ is an example of an extra field. A flag can have two values such as yes or no but it can also contain a code, e.g., a code for a violated rule.

A flag can also be used to generate summary reports. If, for example, a flag is used to indicate that a certain variable is imputed or not, then the total number of imputations can be derived from the flag.

3.3 Presenting log information

Log information can be presented by printing the logs. An alternative is to present the log information on screen and use the theme module in the process of editing for example.

4. Available software tools

Logging is often part of the software that processes the data and seldom a separate software application.

5. Decision tree of methods

A decision tree of methods is not applicable.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Argus (2013), Website Statistical disclosure control <http://neon.vb.cbs.nl/casc/glossary.htm>. Retrieved 25 October 2013.

GSBPM (2009), Generic Statistical Business Process Model. Version 4.0 – April 2009. Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS).

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Quality and Risk Management Models
2. Micro-Fusion – Object Matching (Record Linkage)
3. Statistical Data Editing – Main Module
4. Imputation – Main Module

9. Methods explicitly referred to in this module

1. All method modules in the handbook – Section 20: Logging indicators
2. All method modules in the handbook – Section 21: Quality indicators of the output data

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 2: Design
2. Phase 5: Process
3. Phase 6: Analyse
4. Quality management (as overarching process)

12. Tools explicitly referred to in this module

1. τ - and μ -ARGUS (Argus, 2013). These tools support statistical confidentiality control and produces standard log files. It is used as an example for logging indicators.

13. Process steps explicitly referred to in this module

1. Editing
2. Imputation
3. Statistical disclosure control

Administrative section

14. Module code

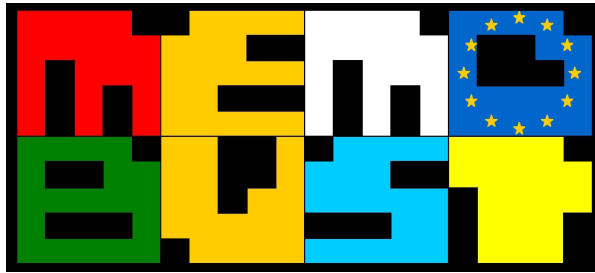
General Observations-T-Logging

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	19-03-2013	first draft	Peter van Nederpelt	Statistics Netherlands
0.1.1	11-07-2013	comment SN's reviewers processed	Peter van Nederpelt	Statistics Netherlands
0.1.2	25-10-2013	second round of SN's reviewers processed	Peter van Nederpelt	Statistics Netherlands
0.1.3	10-01-2014	comment HU processed	Peter van Nederpelt	Statistics Netherlands
0.1.4	16-01-2014	log item changed in logging indicator in order to be consistent with the template for methods	Peter van Nederpelt	Statistics Netherlands
0.1.5	03-02-2014	EB's comment processed	Peter van Nederpelt	Statistics Netherlands
0.1.6	04-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:23



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Mixed Mode Data Collection

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Uni-mode and mixed-mode designs.....	3
2.2 Modes of data collection	3
2.3 Mode quality, cost, and response burden considerations	5
2.4 Multi-source/Mixed-mode design drivers	6
3. Design issues	6
3.1 Parallel and sequential mixed-mode designs.....	6
3.2 The most appropriate modes during the business survey data collection process.....	7
3.3 Increasing web take-up rates	8
3.4 Improving the respondent experience.....	10
3.5 Planning for data quality and its assessment	12
3.6 Tailoring the mixed-mode design to business size	13
3.7 Implementation of mixed-mode design for business surveys.....	14
4. Available software tools.....	15
5. Decision tree of methods	15
6. Glossary.....	15
7. References	15
Interconnections with other modules.....	18
Administrative section.....	19

General section

1. Summary

In this theme module we will discuss mixed-mode data collection designs in business surveys, with the term ‘survey’ in the limited sense of ‘data collection’ from businesses. This topic is of great relevance, since at present many NSIs around the world are moving from uni-mode to mixed-mode designs. When talking about modes in surveys we can distinguish between modes (or channels) of communication as used in a contact or survey communication strategy (like advance and reminder letters, telephone follow-up, etc.), and data collection modes, i.e., the mode used for the questionnaire or the delivery of the data. In this module we focus on data collection modes.

First we will give a general introduction to uni-mode and mixed-mode designs, discussing the characteristics of the various data collection modes. From there we will move on to discussing mixing modes in business surveys, with a focus on the web as a primary mode. This includes a brief discussion on a design for large and multi-surveyed businesses. We conclude with an overview of implementation steps.

2. General description

2.1 *Uni-mode and mixed-mode designs*

In general, we can identify two kinds of survey designs with regard to data collection modes: uni-mode designs and mixed-mode designs. In a uni-mode design, only one mode is used for data collection for all sampled units in the fieldwork period. For business surveys this used to be paper: sampled businesses received a paper questionnaire, sent, e.g., by mail or fax. In a mixed-mode design we use multiple modes to collect data from the sampled units in the data collection period of one survey. Apart from paper, the design can include a self-administered electronic questionnaire to be accessed on the internet, an interviewer-administered telephone interview (based on a paper or electronic questionnaire), or other types of data entry modes like TDE (Telephone/Touchtone Data Entry).

2.2 *Modes of data collection*

An overview of modes used most frequently in business surveys is presented in Table 1. This table also indicates whether a mode is self or interviewer administered, as well as an average indication of effects on data quality and costs. (For a detailed discussion on modes, we refer to: the theme module “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method”; De Leeuw, 2008; Dillman et al., 2009: Chapter 12, on Surveying Businesses and Other Establishments; Groves et al., 2004: Chapter 5: Methods of Data Collection.)

The modes listed in this table mostly are traditional ones, like paper questionnaires (either sent out and returned by mail, or by fax). CATI (Computer-assisted Telephone Interviewing) is a very common mode in social surveys since the 1970s; in mixed-mode business surveys it can be used as an additional mode to increase response rates (as we will see in Section 3.2). CAPI (Computer-assisted Personal Interviewing) refers to an interviewer visiting a respondent; in business surveys this could be the case when a field agent visits a business to complete or help completing a questionnaire.

*Table 1. Data collection modes used in business surveys *)*

Mode of data collection	Self/interviewer administered	Effects on data quality: risks of non-response and measurement errors	Costs
Paper: - mail-out/mail-back, - fax	Self-administered	High	Low
Electronic questionnaires in CAWI: - on-line - off-line	Self-administered	High	Low
Smart phone web questionnaires	Self-administered	High	Low
Smart phone, using Apps	Automatic registration	Low	Low
TDE	Self-administered	High	Low
CATI	Interviewer-administered	Low (depends on the topic: higher for sensitive topics)	Moderate
CAPI	Interviewer-administered	Low (depends on the topic: higher for sensitive topics)	High

*) For definitions, see Glossary. See also Table 1 in the theme module "Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method".

As for electronic questionnaires, those can be run on an ordinary PC, laptop, notebook, or tablet (we do not consider a tablet to be a separate mode). An electronic questionnaire can be completed on-line, or downloaded from the internet (or installed from a CD) and completed off-line. In that case, the data can be sent back by e-mail (as a scrambled attachment), or via a secured internet connection (after logging-in on a secured web portal). Strictly speaking electronic questionnaires are used in all CAI modes, including interviewer-administered CAI modes like CATI and CAPI; in business surveys, however, the term electronic questionnaires often refers to the electronic equivalent of a self-administered paper questionnaire. When the internet is used for accessing the questionnaire, either for on-line or off-line completion, we speak of CAWI (Computer-Assisted Web Interviewing) or Web surveys. In general, when an electronic questionnaire is used for self-completion, the acronym CASI (Computer-Assisted Self-Interviewing) is used. This includes all kinds of electronic formats, like Excel-files that may be sent by e-mail. We could say that CAWI can be considered as a special form of CASI. Design issues that need to be considered with web questionnaires are trusted web portals, firewalls, and visual design and usability issues for various screen settings and web browsers.

The table also includes the use of Smart phones. This is a relatively new device that is just recently being explored for its use in survey data collection in general. Smart phones can be used in two ways. The first way is to use the device for the completion of web questionnaires (like on PCs and tablets), e.g., to be used as a small diary that can be completed every time a specific situation occurs. We consider this to be a separate mode since it involves the development of a separate questionnaire, tailored to the completion process and usability issues when using a small screen. The other one

involves the use of all kinds of Apps that register all kinds of variables, e.g., the route a lorry has taken during a day in a transportation survey. This research is still in its infancy, and will not be discussed further in this module; however, we foresee a lot of opportunities for this device in future survey designs.

2.3 *Mode quality, cost, and response burden considerations*

In general, the effects on data quality and costs are in the opposite direction: expensive modes result in the best data quality. With data quality we refer to non-sampling errors like non-response and measurement errors (see the theme module “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method” and Snijkers et al. (2013, pp. 83-125) for a detailed discussion on quality issues in business surveys). These errors are affected by a large number of survey design considerations like the choice of the mode, the design of the questionnaire, and the survey communication strategy.

The mode choice depends, e.g., on characteristics of the sampled population: if businesses cannot access the internet, CAWI is not a good choice; this will result in high non-response rates. Self-completion modes in general have a high risk of high non-response rates, since the persuasion power of these modes is low: a paper questionnaire (or an advance letter introducing a web questionnaire) can be left on a desk, forgotten about or left since it has no priority; they therefore require a very well designed contact and reminding strategy. On the other hand these modes are cheap in comparison to interviewer-administered modes, which require a well-trained and managed interviewer unit.

The mode impacts the questionnaire design: the questionnaire needs to be tailored to the mode, in such a way that usability of the questionnaire is high and response burden to complete it is low. The questionnaire should also be adapted to the response process within businesses. If this is not the case, this may result in measurement errors, item non-response, and maybe even unit non-response: respondents may skim the questionnaire and apply a satisficing response behaviour. When they get stuck they may quit completing. Self-completion modes require help desk or on-line web support. Interviewers, on the other hand, guide respondents in the response process and provide support, which generally results in less measurement errors (provided that interviewers are well trained in doing their job in order to reduce errors and interviewer effects as much as possible) The presence of an interviewer would assure that a questionnaire is fully completed. (Questionnaire design in a mixed-mode design context is discussed in Section 3.5.)

The communication strategy refers to the contact and reminding strategy, and should motivate and facilitate businesses to respond: business respondents should be motivated to complete a questionnaire. This also includes, e.g., the timing of dispatching a questionnaire (the communication strategy is discussed in Section 3.4.). The modes used in the communication strategy may differ from the data collection modes, e.g., a web questionnaire is used, while businesses are contacted using a paper advance letter, providing the web address and log-in codes.

The survey costs at the business side are called compliance costs or response burden. Response burden is affected by the three survey components discussed above: the mode of data collection, the questionnaire, and the survey communication strategy. High perceived and high actual response burden result in high risks of non-response and measurement errors. (For a detailed discussion on response burden we refer to Snijkers et al., 2013, pp. 303-358.)

When designing a survey, these considerations need to be taken into account, and trade-off decisions between quality and costs (both internal and compliance costs) need to be made.

2.4 Multi-source/Mixed-mode design drivers

Surveyors turn to mixed-mode designs when uni-mode designs are insufficient to achieve the required results within the available time and budget. In general, drivers for applying mixed-mode design are cost and non-response considerations. Additional considerations as mentioned by Dillman et al. (2009) are: improved timeliness of the response, improved coverage of the sampled population and reduced non-response errors.

The main drivers for introducing mixed-mode designs in official business surveys are budget cuts, response rate improvement, and response burden reduction: cheaper modes are introduced to get the same levels of response rates, while at the same time reducing response burden. As a consequence, trade-off decisions with regard to quality have to be made (and documented). The surveyor must be aware that new modes may impact the data quality (also when moving from one uni-mode design to another, as we have briefly discussed in Section 2.3); and in addition, the combination of modes may introduce mode effects, i.e., that respondents complete questions differently for the various modes (see also Section 3.5). This increases measurement bias. On the other hand, mixed-mode designs are introduced to reduce non-response rates and non-response bias caused by selective responses. As for costs, it should be noted that the data collection costs may be lower for mixed-mode designs, whereas the design costs increase.

To implement these goals, NSIs nowadays have data collection strategies, indicating how data for producing statistics are to be collected. In short, such a strategy may dictate that first data from various secondary administrative sources should be used, and only if this is not possible or insufficient, a survey can be conducted. The modes for data collection and their order of usage are also generally dictated by cost considerations: web is the dominant mode, followed by paper, and finally interviewer-administered modes. For example, Statistics Canada (Brodeur and Ravindra, 2010), Statistics Netherlands (Snijders et al., 2011), Statistics New Zealand (2011), and Statistics Norway (2007) have implemented such a policy. The result of such a strategy is a multi-source/mixed-mode design for business data collection (Snijders, 2008). In this module we focus on the mixed-mode aspects.

3. Design issues

In mixed-mode designs we try to find an optimal mix of modes, striving for the best possible data quality, while keeping survey and compliance costs (response burden) as low as possible. In this section we will discuss how to mix modes within constraints (as discussed above).

3.1 Parallel and sequential mixed-mode designs

Basically, there are two ways of applying a mixed-mode design in business surveys (Snijders et al., 2013, pp. 359-430):

- A parallel or simultaneous approach, in which all mode options are offered to the respondents at once and the choice is left to them. Even though these options are offered with the intention to make it easy on respondents and to increase response rates, the outcome may be the opposite: these options may confront respondents with making a choice

that is difficult to make, as they do not know the consequences of their decision. As a result they may stick to the traditional mode, as this is what they know, or they may not respond at all. (Note that this decision is made prior to actually opening the questionnaire, and is therefore directed by the communication strategy as we will see in Section 3.3.)

- A sequential approach, in which a primary mode is offered first, with alternative modes offered in later stages to facilitate respondents and to increase response rates. Typically the least expensive mode is offered first, with a switch to more expensive modes in the next stages of the data collection process; the most expensive mode is being used in the final stage. This approach effectively uses survey finances, diminishes the need for respondents to choose a mode, and if used in an appropriate way can increase response rates (or at least increase the take-up rate for the primary mode as we will discuss in Section 3.3).

3.2 *The most appropriate modes during the business survey data collection process*

In business surveys, the most common modes are the self-completion modes. This has to do both with the response process within businesses, as well as cost considerations by the survey organisation. As we have seen in Table 1, the self-completion modes are the cheapest ones. As for the response process within businesses, the completion of a questionnaire may require data from various departments (see Section 3.4), which makes the completion of an interviewer-administered questionnaire cumbersome, time-consuming and expensive.

The fact that no interviewers are present puts a heavy load on the questionnaire design to ensure that all questions and instructions are correctly understood (i.e., the cognitive response process is performed correctly), and the respondent navigates correctly through the questionnaire. Also the full completion of a questionnaire is hard to control. In former days (up to about 2000), Statistics Netherlands had field managers visiting businesses to help them in completing questionnaires. This help facility has however been deleted because of budget cuts. Instead, help desk and on-line web support (including for instance a FAQ) is offered.

Interviewers may, however, be used in the data collection process in a number of ways, since personal communication is more effective than one-way communication. In the pre-field stage, businesses may be contacted on the phone to introduce the survey, to gain survey cooperation from the business management, to deal with gatekeepers, and to get the name of a contact person (a competent and knowledgeable respondent). During the field stage, interviewers can contact non-respondents for reminding, and motivate them to complete the questionnaire. An appointment could be made as to when a business will return the questionnaire. Also businesses may request a questionnaire in another mode, e.g., a paper questionnaire instead of a web questionnaire. In case of short questionnaires with an easy data retrieval process, the data may be collected at the spot. Paxson et al. (1995) advocate to use telephone follow-ups in addition to self-completion modes to encourage response and obtain the needed data. In these ways, we use the power of personal communication, and the persuasion and motivation skills of interviewers to increase response rates (Snijkers et al., 2013, pp. 303-430).

The more expensive modes, like using the telephone to actually collect the data, may not be applied for small and medium-sized businesses. This may be left for the more important businesses (e.g., the larger ones, or those with more than average turnover) within the various industry sectors. The

application of this design requires monitoring the response rates for subgroups during the fieldwork. The result is a tailored sequential web/paper/CATI mixed-mode design.

In addition, for the very large, multi-surveyed and indispensable businesses tailored data collection procedures may be implemented and applied by a special unit, aimed at reducing non-response and measurement errors, as well as reducing response burden (as we will discuss in Section 3.6).

3.3 Increasing web take-up rates

As we have seen above when discussing the data collection strategy, NSIs turn to web or electronic questionnaires as their primary mode. A major question now is: how to increase the web take-up rate, i.e., the proportion of businesses using the web questionnaire, without affecting overall response rates and data quality?

The evidence based on experiences in case studies indicates that a sequential approach is superior; the results indicate that with the parallel approach a newly introduced mode will hardly be used. A small-scale study by Hak, Anderson and Willimack (2003) explains this. They conclude that a major reason for sticking to the existing mode was that respondents saw no reasons for changing a routine that was convenient to them. To get to this decision, respondents compared the burdens of the existing mode with the perceived burden of the new mode. This applies to recurring surveys.

For one-time cross-sectional surveys (or a first wave in a recurring survey), a large-scale experiment embedded in a social survey conducted by Statistics Sweden provides evidence. This study shows that respondents have a tendency to use the mode that they have immediate access to. Holmberg et al. (2010) call this the “mode-in-the-hand principle”. In the experiment, a number of sequential approaches were compared to a simultaneous approach. The results indicated that the designs in which the paper option was presented in a stage later in the data collection process showed the highest web take-up rates, with comparable overall response rates. The results also suggested that the later in the fieldwork the alternative paper option was introduced, the higher the web take-up rate. We may assume that the mode-in-the-hand principle also is applicable to business surveys as a guiding principle for increasing web take-up rates, and points towards the sequential approach.

We will now give a number of specific examples of recent mixed-mode designs in business surveys as conducted by NSIs. The examples provide additional evidence on how to get a high response rate by use of a sequential mixed-mode design, i.e., increasing the take-up rate for a newly introduced mode, with unaffected overall response rates and data quality.

3.3.1 Example 1: introducing TDE at the UK Office for National Statistics (ONS)

In the UK, TDE is used for several ONS business surveys that collect nine or less data items. When TDE was introduced, respondents were sent a paper questionnaire with TDE offered as a simultaneous data collection mode: the TDE mode was indicated on the paper questionnaire. With this approach, response via TDE was generally around 20 to 30%. After a redesign, TDE was presented as the primary mode of data collection, and paper was offered as an alternative data collection mode: respondents received a letter indicating the use of TDE; they had to request the paper questionnaire. This approach resulted in an increase of the response rate via TDE to 80 or 90% (Jones, 2011; Thomas, 2007).

3.3.2 Example 2: Introducing a mixed-mode design for the Structural Business Survey at Statistics Netherlands

A second example is the sequential approach that was adopted for the introduction of the web mode for the Annual Structural Business Survey (SBS2006) conducted by Statistics Netherlands in 2007. In the first two contacts, the 63,644 respondents were only offered the web questionnaire. The paper questionnaire, that was available on request, was not mentioned in the letter. For the last reminder (the second or third reminder, depending on sector of the economy) a paper questionnaire was enclosed in the envelope. In all letters the web option was clearly promoted: in the middle of the letter login codes were provided. Before implementing this approach for the survey as a whole, it was tested in a field pilot in 2006 (SBS2005) with hopeful results (Snijkers et al., 2007). Based on the pilot results, it was decided to use the sequential approach in the next year for the whole sample. For the SBS2006, the web take-up rate was 80%, while 20% of the responding businesses used the paper questionnaire. The overall response rate of the SBS2006 was comparable to previous years: approximately 80%.

Response analysis (Morren, 2008) revealed that the percentage of web reporting increased with business size: from 78% for very small businesses to 89% for large businesses. For 24% of the businesses the mode of data collection was changed: they received the paper questionnaire with the 2nd or 3rd reminder. From these businesses 57% responded. Even though these businesses received a paper questionnaire, about one third used the web questionnaire. The mode could also be changed on request by the business: this was done by 4084 businesses. An interesting result here is the high response rate for businesses that requested a paper questionnaire: 91% (out of 4084). A voluntary request by respondents to receive a questionnaire in another mode seems a strong predictor for response. This indicates that making an alternative mode obtainable only on initiative of the respondent, would be a way to stimulate response.

3.3.3 Example 3: Introducing web questionnaires at the Australian Bureau of Statistics

A third example comes from the Australian Bureau of Statistics (Black and Ang, 2013). Like more NSIs, ABS is currently in the process of introducing web questionnaires in business and social surveys. In the May 2012 Employee Earnings and Hours Survey (EEH; a biannual survey) a sequential approach was applied: sampled businesses were given a link to an on-line questionnaire; a paper questionnaire was only sent out upon request by a business. ABS reports that this approach proved very successful, with a web take-up rate of 90%.

In addition, an analysis was carried out to provide reassurance that the new design did not impact on EEH estimates. The small amount of data obtained by other modes, however, limited the ability to detect systematic mode effects. Nevertheless, the analysis showed no systematic mode effects in the estimates. The analysis consisted of four parts:

1. An exploratory analysis, comparing the distributions of variables of interest for web and non-web respondents.
2. Comparing of EEH responses with data from other data sources for the same units. This involved comparing EEH web and non-web responses at the employer level with corresponding data provided by the same employer in other data sources. The values of common variables of interest for these units were compared using scatter plots, to examine

if the distributions for the web and non-web responses differed significantly. A more formal analysis was then conducted using linear regression analysis.

3. Modelling earnings and number of employees. Data from the May 2010 and May 2012 Average Weekly Earnings surveys (both paper surveys) was used to estimate how units common to both the 2010 EEH and 2012 EEH surveys would have responded if these units were provided with a paper form in 2012. The relationship between the modelled and actual 2012 EEH values for the web and non-web businesses was then compared.
4. Propensity score sub-classification. A logistic regression model was created to estimate the probability that each business in the EEH sample would respond via web. The sample was then grouped into five categories based on these estimated probabilities. A web mode impact was estimated within each category separately, and these were then combined to form an overall estimated impact.

Based on these results ABS decided to adopt the sequential approach for all its business surveys, and have web as the primary mode.

3.3.4 Example 4: Increasing web take-up rates in Sweden and Norway

Between 2005 and 2007, Statistics Sweden conducted a number of studies for web data collection designs in business surveys (Erikson, J., 2007; Erikson, 2010). Haraldsen (2009) studied the web take-up rates for business surveys conducted by Statistics Norway. These studies showed results in the same direction:

- Spontaneous web take-up rates for the simultaneous approach are low, between 5 and 25 %.
- Web take-up rates are increased significantly by eliminating paper in the first contacts.
- To get a high enough final response rate, alternative modes like paper questionnaires can be used in later stage (e.g., with reminders).
- Take-up rates are higher for respondents who already have experienced non-paper modes.
- Small businesses show the lowest take-up rates for web.
- Over the years, web take-up rates have increased.

3.4 Improving the respondent experience

A necessary factor that we need to take into account is the respondent factor. Even if we do not leave a choice for respondents with regard to the mode, we have to facilitate and motivate them. Nowadays, however, many businesses are used to completing web forms (for government regulations). As a consequence, they ask for web questionnaires, as they have the perception that this will reduce the burden of survey compliance. Still, this requires well-designed survey components.

Three components in the survey design affect web take-up rates: the communication strategy, the questionnaire, and a web portal (see Dowling and Stettler, 2007, for a more detailed discussion). With these three components we can influence the respondent's preferences with regard to mode, and facilitate the response process that is going on within businesses (see Snijkers et al., 2013, pp. 39-82). We can push a primary mode, but still be service-oriented. The question respondents will ask themselves when using an electronic questionnaire is: Will this make my life easier, and can I do this job quicker by using an electronic questionnaire? This is a cost-benefit analysis to estimate response burden, as we have seen Hak et al. (2003) also concluded.

For businesses, the response process starts when they are contacted in some way (e.g., with an advance letter) and the questionnaire is provided to them. The communication strategy (see also the theme module “Data Collection – Design of Data Collection Part 2: Contact Strategies”) indicates how businesses are contacted and followed-up in case of non-response, and is aimed at receiving timely, accurate and complete responses. This strategy can be a one-sided approach in which a survey organisation sends out questionnaires and assumes businesses to respond. Such an approach receives much criticism by businesses, and does not help to get cooperative businesses. An extended approach is a tailored and persuasive strategy that communicates the survey request and related instructions and procedures, and motivates and facilitates businesses to comply with the survey request. The survey communication strategy is an objective-driven and coherent plan of communication measures and actions. Such a strategy will improve the chances of getting respondents in the right mood for survey participation, and reducing (perceived) response burden by facilitating their survey-related work. In Sections 3.2 we have seen how such a strategy interacts with mode considerations, and Section 3.3 discussed examples of a communication strategy to increase web take-up rates. (Business survey communication is discussed extensively by Snijkers et al., 2013, pp. 359-430.)

When contacting businesses, we need to make a difference between the decision to participate, which is taken at a management level; and the actual completion of a questionnaire at an employee level. After businesses have decided to participate, they will start completing the questionnaire. This can be a complex process, in which various participants (e.g., data providers from various departments) are involved (see Snijkers et al., 2013, pp. 39-82, for a discussion on the response process). If respondents don't like a newly designed questionnaire in a new mode (like when introducing a web questionnaire), they will not use it, and in future contacts (in recurring surveys) it will be hard to convince respondents to use this mode (because of the bad experience). The implication of this is that we need to invest in questionnaires by developing them well, and tailoring them to the response process (see Snijkers et al., 2013, pp. 303-358, for a discussion on questionnaire design). Also questionnaires need to be tested thoroughly before using them in the field. Testing involves technical performance testing, cognitive testing and usability testing (see Snijkers et al., 2013, pp. 253-302, for a discussion on development, testing and evaluation methods).

With web questionnaires, the completion process starts when respondents go to the internet, enter the web address of the portal, and log-in; this is followed by the actual completion of the questionnaire and ends with submitting the data, and getting a feedback note saying that the data have been received and thanking the respondent. Also benchmark information can be provided; if this is to be an incentive, it should be mentioned as a promised incentive in the advance letter. To ensure that this process works well, also web portals need to be designed carefully, and tested for usability. Keywords in web site design are: easy to find, accessibility, and usability. A pitfall is that is assumed that respondents know how the internet works, and that they know what to do. Therefore the process needs to be easy, straightforward and logical for respondents (Snijkers et al., 2007).

By these investments in the communication strategy, the questionnaire and the web portal, respondents are likely to experience a reduced response burden. Statistics New Zealand (2013) calls this “Improving the respondent experience”.

3.5 *Planning for data quality and its assessment*

When introducing a mixed-mode design we need to assess whether the collected data are affected by the various modes, like ABS did in example 3: are there mode effects? In other words: show the collected data real-world levels and developments, or are they impacted by the survey design. The occurrence of mode effects (as well as survey errors as a whole) can be prevented by planning for data quality from the beginning. This involves a number of steps in the various stages of designing and conducting a business survey:

1. In the questionnaire design stage. In a mixed-mode design, for each mode a questionnaire needs to be developed. Here, two approaches can be adopted: 1. the questionnaires in all modes are exactly the same (e.g., translating a paper questionnaire one-on-one into an electronic questionnaire); 2. the questionnaires are tailored to the mode (i.e., all questionnaires make optimal use of the functionalities in the mode, while keeping the question wording and definitions, and routing the same). We believe that the second approach is the best way to go. In fact, we see no other option, since for electronic questionnaires businesses expect intelligent functionalities to be included (like automated routing, automated calculations, built-in edits, etc.; Snijkers et al., 2007; see Snijkers, 1992, for an overview of properties of electronic questionnaires). This approach is also recommended by Snijkers et al. (2013, pp. 303-358): in the design first focus on the issues that need to be addressed for a paper questionnaire (like wording of questions, response options, instructions, and order of the questions), and then improve the questionnaire by incorporating electronic functionalities. He states (ibid., p. 304): “As long as we have done our best within the paper format, we are less worried about possible mode effects that arise from improvements made in the web version.” This approach is followed by Statistics Netherlands for redesigning the Structural Business Survey questionnaire (see Example 2; Snijkers et al., 2007).
2. In the prefield stage, the questionnaires need to be tested whether they are valid measuring instruments (Snijkers et al., 2013, pp. 253-302). Also in this stage, sampled businesses in recurring surveys may be pre-notified (by sending a pre-notification letter) about changes in the mode, so they can be prepared for the new situation (Snijkers et al., 2013, pp. 359-430).
3. Next a pilot (or experiment) may be conducted, as was carried out by Statistics Netherlands in Example 2. A pilot is conducted to test the design in real-life conditions, i.e., to test if the survey design works as planned, to test if the survey production process works properly, and to study mode effects. Even though we have reduced errors by pre-testing the questionnaires, we cannot be sure about the results: for a business survey there are many uncertainties in the field that affect the survey outcomes. Therefore it is better to first test the design with a small part of the sample, instead of risking a whole fieldwork to go wrong. If possible, the pilot group needs to be a representative subsample of the population, thus applying an experimental design. It is our experience, however, that conducting an experiment with a properly defined control and experimental group, is hard to do in practice. Partly because we cannot fully control how businesses will use the various modes.

The pilot needs to be evaluated and analysed with regard to data quality issues as is done, e.g., by the ABS in example 3. In general, the following analyses could be carried out for key variables:

- Comparing estimates for the various modes, if the groups can be made comparable.
- Modelling over time for the same survey: comparing estimates for the same groups over time.
- Modelling over data sources: comparing data for the same units but from various sources in the same time span.

If the results from the pilot study are acceptable, and within pre-defined quality limits, the design can be implemented for the survey as a whole.

4. In the post-field stage, it is recommendable to evaluate the survey and conduct a quality assessment as discussed for the pilot (Snijkers et al., 2013, pp. 253-302 and pp. 431-458). In addition to the analyses listed in step 3, the editing of key variables can be monitored, by monitoring the number of edits for these variables, as well as the differences in value for these variables prior to editing as compared to edited values (Snijkers et al., 2013, pp. 431-504). If edits are included in the electronic questionnaire, this should show in the monitoring results: less post-field editing should be necessary for the electronic questionnaire.

3.6 Tailoring the mixed-mode design to business size

The mixed-mode design as discussed in the previous sections can be applied to the sample as a whole, which would be the case in social surveys. In business surveys, however, it is recommendable to tailor the design to the size of the businesses. As we have already discussed in Section 3.2, telephone non-response follow-ups can be conducted for small and medium-size. The very large businesses however require a more personal approach.

Since these large businesses are key in producing economic statistics they cannot be missed, and as a consequence these businesses are multi-surveyed. Special procedures for dealing with large businesses become more important in the context of globalisation (UNECE, 2011), with businesses organised on a global rather than national level. Also their data across surveys need to be coherent and consistent. To deal with these key businesses, NSIs have established special units, e.g., Statistics Canada (Sear et al., 2007), Statistics Netherlands (Vennix, 2012), Statistics Sweden (Erikson, A.-G., 2007), and US Census Bureau (Marske et al., 2007).

These units consist of specially trained managers, so called customer relationship managers, large enterprise managers, or account managers. Among other things, these relationship managers assist these businesses in the completion process, and provide information about the surveys (background, output, and changes). Their knowledge of the businesses in their portfolio is essential in this business-oriented approach: they know the situation of the businesses, know the surveys they receive, and understand the burden of receiving many questionnaires. These procedures are aimed at building and maintaining a good relationship with these key businesses, and over all surveys obtaining coherent, as well as timely, complete, and accurate data.

3.7 *Implementation of mixed-mode design for business surveys*

To conclude, the steps below can be considered for the implementation of a mixed-mode design for business surveys (Snijkers et al., 2013, pp. 359-430). With the multi-source/mixed-mode data collection strategies in mind, a general guideline is stated by Dillman et al. (2009, p. 422): plan for a mixed-mode design from the beginning.

Pre-field stage steps:

1. Use existing data sources (like existing survey data and administrative data) as much as possible. Only if needed, plan for a survey, with a mixed-mode design (possibly with the web as primary mode).
2. Optimise the questionnaire designs for the various modes, taking the business context and response burden issues into consideration. Start with paper, and then optimise the design for the web. Make sure that the questionnaire works properly for all major hardware platforms (Windows PCs, Apple computers, tablets), and web browsers. Pre-test all questionnaires to improve the respondent experience.
3. Design web portals and conduct usability testing. Make sure that the web portal is trusted, and not blocked by firewalls.
4. If possible, within time and budget constraints, conduct a pilot, to test the mixed-mode design and the survey production process.
5. Inform businesses (in recurring surveys) about changes in mode. Don't ask for the mode they would prefer, but inform them about the new primary mode that is coming up, and what they can expect.

Field stage steps:

6. Restrict access to paper, and apply a sequential approach: don't provide a paper questionnaire in the first contacts, but as an alternative in a later stage. In the letters mention that a paper questionnaire is available on request.
7. A mode switch can be done sooner or later. When to offer the paper questionnaire depends on the number of reminders, and the development of the response rate during the fieldwork (if response is falling behind, it is recommended to change to another mode).
8. Facilitate the use of the primary mode. In all letters, also the reminder letters that come with a paper questionnaire, clearly promote the web option. Provide support by a centralised help desk and on-line help (FAQs, contacts for technical and non-technical support, etc.).
9. In addition to self-completion modes, businesses can be followed-up by phone to encourage responding, and even collect the data. This can be tailored to, e.g., size, turnover, and industry.
10. Make access to web questionnaires easy from start to end. The entire process of finding the questionnaire, logging-in, opening and completing the questionnaire, and sending the data should be simple and straightforward. At the end of the process, provide feedback by confirming the receipt of the data, and say 'thank you'.
11. A special unit for large and multi-surveyed key businesses is responsible for contacting and building a relationship with these businesses.

Post-field stage step:

12. Evaluate the survey, and assess quality of the data to check for mode effects and improve the design (according to the Deming cycle: plan-do-check-act).

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Black, M. and Ang, L. (2013), Measuring the Impact of webforms in the Survey of Employee Earnings and Hours. ABS Methodological News, A Quarterly Information Bulletin, Australian Bureau of Statistics, March 2013, 4–5.
- Brodeur, M. and Ravindra, D. (2010), Statistics Canada’s new use of administrative data for survey replacement. Paper presented at the European SIMPLY Conference on Administrative Simplification in Official Statistics, Ghent, Belgium, Dec. 2–3.
- De Leeuw, E. D., Hox, J. J., and Dillman, D. A. (eds.) (2008), *International Handbook of Survey Methodology*. European Association of Methodology, Lawrence Erlbaum Associates, New York, London.
- De Leeuw, E. D. (2008), Choosing the Method of Data Collection. In: De Leeuw, E. D., Hox, J. J., and Dillman, D. A. (eds.), *International Handbook of Survey Methodology*, Lawrence Erlbaum Associates, New York, 113–135.
- Dillman, D. A., Smyth, J. D., and Christian, L. M. (2009), *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 3rd edition. Wiley, Hoboken, NJ.
- Dowling, Z. and Stettler, K. (2007), Factors influencing business respondents’ decision to adopt web returns. *Proceedings of the 3rd International Conference on Establishment Surveys (ICES-III), Montreal, June 18–21*, American Statistical Association, Alexandria, VA, 1028–1039.
- Erikson, A.-G. (2007), Large enterprise management – higher quality and lower burden through better relations. *Proceedings of the 3rd International Conference on Establishment Surveys (ICES-III), Montreal, June 18–21*, American Statistical Association, Alexandria, VA, 825–829.
- Erikson, J. (2007), Effects of offering web questionnaires as an option in enterprise surveys: The Swedish experience. *Proceedings of the 3rd International Conference on Establishment Surveys (ICES-III), Montreal, June 18–21*, American Statistical Association, Alexandria, VA, 1431–1435.

- Erikson, J. (2010), Communication strategies in business surveys – implications when web data collection becomes the main mode. Paper presented at the 3rd International Workshop on Business Data Collection Methodology (BDCM), Wiesbaden, Germany, April 28–30.
- Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004), *Survey Methodology*. Wiley, Hoboken, NJ.
- Hak, T., Willimack, D. K., and Anderson, A. E. (2003), Response process and burden in establishment surveys. *Proceedings of the Section on Government Statistics, Joint Statistical Meetings, San Francisco, Aug. 3–7*, American Statistical Association, Alexandria, VA, 1724–1730.
- Haraldsen, G. (2009), Why don't all businesses report on web? Paper presented at the 4th International Workshop on Internet Survey Methodology, Bergamo, Italy, Sept. 17–19.
- Holmberg, A., Lorenc, B., and Werner, P. (2010), Contact strategies to improve participation via the web in a mixed-mode mail and web survey. *Journal of Official Statistics* **26**, 465–480 (available at www.jos.nu).
- Jones, J. (2011), Effects of different modes, especially mixed modes, on response rates. Paper presented at the Workshop on Different Modes of Data Collection, Eurofond, Dublin, April 6–7.
- Marske, R., Torene, L., and Hartz, M. (2007), Company-centric communication approaches for business survey response. *Proceedings of the 3rd International Conference on Establishment Surveys (ICES-III), Montreal, June 18–21*, American Statistical Association, Alexandria, VA, 941–952.
- Morren, M. (2008), The 2006 Structural Business Survey: Response Analysis (in Dutch: “De PS-2006: een evaluatie van de respons”). Research report, Statistics Netherlands, Heerlen.
- Paxson, M. C., Dillman, D. A., and Tarnai, J. (1995), Improving response to business mail surveys. In: Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S. (eds.), *Business Survey Methods*, Wiley, New York, 303–317.
- Sear, J., Hughes, J., Vinette, I., and Bozzato, W. (2007), Holistic response management of business surveys at Statistics Canada. *Proceedings of the 3rd International Conference on Establishment Surveys (ICES-III), Montreal, June 18–21*, American Statistical Association, Alexandria, VA, 953–958.
- Snijkers, G. (1992), Computer assisted interviewing: Telephone or personal? In: Westlake, A., Banks, R., Payne, C., and Orchard, T. (eds.), *Survey and Statistical Computing*, North-Holland, Amsterdam, 137–146.
- Snijkers, G., Onat, E., and Vis-Visschers, R. (2007), The Annual Structural Business Survey: Developing and Testing an Electronic Form. *Proceedings of the Third International Conference on Establishment Surveys (ICES-III), Montreal, June 18–21*, American Statistical Association, Alexandria, VA, 456–463.
- Snijkers, G. (2008), Getting data for business statistics: A response model. Paper presented at the 4th European Conference on Quality in Official Statistics, Rome, July 8–11.
- Snijkers, G., Göttgens, R., and Hermans, H. (2011), Data collection and data sharing at Statistics Netherlands: Yesterday, today, tomorrow. Paper presented at the 59th Plenary Session of the

- Conference of European Statisticians (CES), United Nations Economic Commission for Europe (UNECE), Geneva, June 14–16 (available at www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/20.e.pdf).
- Snijders, G., Haraldsen, G., Jones, J., and Willimack, D. (2013), *Designing and Conducting Business Surveys*. Wiley, Hoboken, NJ.
- Statistics Norway (2007), *Strategy for Data Collection*. Statistics Norway, Oslo (available at www.ssb.no/english/about_ssb/strategy/strategy2007.pdf).
- Statistics New Zealand (2011), *Statistics New Zealand's Collections Strategy for 2010-20: Transform Collections*. Statistics New Zealand, Wellington.
- Statistics New Zealand (2013), *Statistics New Zealand's Respondent Experience Strategy for 2013-20: Willing respondents, finding it easy*. Statistics New Zealand, Wellington.
- Thomas, P. (2007), *Using telephone data entry as a data collection mode for business surveys*. Paper presented at the 56th Session of the International Statistical Institute (ISI), Lisbon, Aug. 22–29.
- UNECE (2011), *The Impact of Globalization on National Accounts*. United Nations Economic Commission for Europe, New York/Geneva.
- Vennix, K. (2012), *The Treatment of Large Enterprise Groups within Statistics Netherlands*. Paper presented at the 4th International Conference on Establishment Surveys (ICES-IV), Montreal, June 11–14, American Statistical Association, Alexandria, VA.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Main Module
2. Questionnaire Design – Electronic Questionnaire Design
3. Data Collection – Main Module
4. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
5. Data Collection – Design of Data Collection Part 2: Contact Strategies
6. Data Collection – Techniques and Tools
7. Data Collection – CATI Allocation
8. Response – Response Burden

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 2.3: Design data collection methodology
2. GSBPM Sub-process 4.2: Set up collection
3. GSBPM Sub-process 4.3: Run collection

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

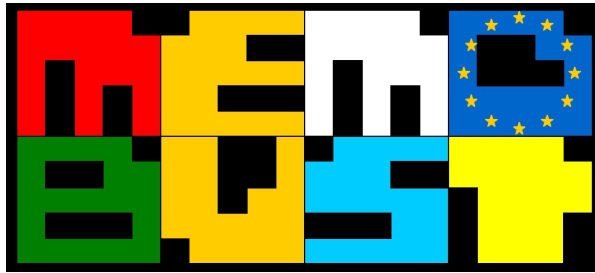
Data Collection-T-Mixed Mode

15. Version history

Version	Date	Description of changes	Author	Institute
0.0.5	14-02-2013	first version	Ger Snijkers, Rob van de Laar	CBS (Netherlands)
0.1	24-06-2013	first complete version	Ger Snijkers	CBS (Netherlands)
0.2	06-12-2013	second version	Ger Snijkers	CBS (Netherlands)
0.3	27-01-2014	third version after review Editorial Board	Rob van de Laar	CBS (Netherlands)
0.3.1	28-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:49



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Different Types of Surveys

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Different types of surveys	3
2.2 STS v. SBS	5
2.3 Types of statistical processes.....	10
3. Design issues	11
4. Available software tools.....	11
5. Decision tree of methods	11
6. Glossary.....	11
7. References	11
Interconnections with other modules.....	13
Administrative section.....	14

General section

1. Summary

Business surveys provide information on different aspects of the economy and economic activity of enterprises, e.g., production, employment, wages and salaries, trade, financial results, etc. They enable us to observe progress and changes in given domains of the economy, as well as monitor developments in the whole economy and identify specific phases of economic conditions of a country. This information is also used to make decisions concerning individual business activities as well as formulate economic and monetary policies. Different types of surveys are designed for different purposes. A distinction can be made between two main groups of business surveys – Structural Business Statistics (SBS) for annual or multiannual statistics and Short-Term Statistics (STS) for monthly and quarterly information on economic activity of enterprises.

The aim of this module is to present general information about the above surveys – STS and SBS – taking into account, among other things, the subjective and objective scope of surveys and type of statistical output. The module also describes differences between the two kinds of surveys, especially concerning the output information and its role in the informational system.

The module is also devoted to the classification of surveys in terms of characteristics, sources and methods of data processing. It highlights the role of classification for statistical processes.

2. General description

National Statistical Institutes (NSIs) conduct a number of surveys aimed at providing statistical information about the economy and society of their countries. Despite the diversity of surveys, some methods of data production are common to many of them. As a result, we can classify surveys into several groups. This section gives an overview of a typology of surveys based on common features related to the process of statistical production. There are different criteria of survey classification – from sources of data, methods of data collection and processing, to the type of output information. For example, business surveys can be divided into two main types, STS and SBS, which differ from each other in terms of output information.

The structure of section 2 is organised as follows. Subsection 2.1 describes two main types of business surveys, i.e., STS, SBS, and differences between them with respect to their scope and aims. Section 2.2 presents a classification of surveys based on a typology of surveys used in the Central Statistical Office (CSO) of Poland. Subsection 2.3 deals with a division of statistical surveys in terms of the type of data sources and methods of data processing.

2.1 Different types of surveys

Each statistical survey is usually unequivocally described by a set of characteristics, such as the objective and subjective scope of a statistical survey, form and frequency of data collection, type of output information, etc. These characteristics are the basis of a classification of statistical surveys, mentioned below:

Survey topic

A statistical survey can focus on many **areas, domains** and **phenomena** of economic or social life and analyse them with statistical methods (SCO, 2012b). For example, in the case of a business survey, there are, among others, surveys on business tendency, financial results of enterprises, production, labour market, retail trade, production of manufactured goods, etc. One survey can cover one or more areas. For example, the monthly report on economic activity of enterprises, conducted in Poland, monitors many aspects of business operation, e.g., it delivers information on sold production, turnovers, new orders, wages and salaries and employment. On the other hand, some surveys are devoted to one topic or activity-related entities to obtain more detailed information about a given domain (e.g., industrial production; construction activity; transport and communication; materials, fuel and energy market; retail trade, etc.) (CSO, 2012a).

Survey type - scope of the surveyed population

In terms of the surveyed population, we can distinguish a **complete survey of the entire population**, for example an economic census, economic activity of large enterprises, or a **sample survey** based on a randomly (e.g., economic activity of micro-enterprises) or purposefully (e.g., producer prices) selected sample from a given population (CSO, 2012b). Full and detailed information on sampling methods and consequences of their applications in business surveys can be found in the topic entitled Sample Selection; see “Sample Selection – Main Module”.

Subjective and objective scope

Statistical surveys may vary in terms of the scope of collected data and units obliged to provide data, taking into account, among others, such characteristics of entities as legal and organisational form, type of activities performed, size of enterprises, etc.

Sources of statistical data

NSIs obtain data from different sources. Generally, we can distinguish surveys based on one source, such as statistical reports, administration data systems, pay cards, mobiles, web services, own estimates, etc., or those based on mixed sources, where some of the sources mentioned above are combined. In recent times, in the case of business surveys, there has been a tendency to combine data from surveys and administrative sources (being one of the category of secondary sources). The module Collection and Use of Secondary Data presents advantages and disadvantages of this approach and practical issues concerning using secondary data in statistical production (link to Collection and Use of Secondary Data). An overview of methods of data integration and problems concerning linking different data sources at micro level (data sources composed of units) is available in the topic Micro-Fusion; see “Micro-Fusion – Data Fusion at Micro Level”.

Obligatory reporting of statistical data or voluntary participation in a survey

Statistical offices can collect data on the basis of obligatory or voluntary participation of respondents in a survey. Voluntary participation generally applies to social surveys, so it does not concern entities which conduct business activity. In the case of obligatory surveys, subjects are obliged to provide to statistical offices in due time **complete** and **exhaustive information** in a predetermined scope and form.

Form of data transfer

Statistical data are generally collected by means of reporting forms and questionnaires, which can be delivered to NSIs in electronic (e.g., CAWI - Computer Assisted Web Interviews) or written form (e.g., by post). Another, less popular, way of obtaining information in business surveys is to interview respondents, using either CAPI (Computer Assisted Personal Interviews) or CATI (Computer Assisted Telephone Interviews) mode. The newest, recently developed method of gathering information for statistics is EDI (Electronic Data Interchange), which enables businesses a direct transfer of data from their internal systems.

All the above techniques, conditions of their application, advantages and disadvantages of using them in business surveys are described in the module “Data Collection – Techniques and Tools”.

Frequency

Considering survey frequency, the following types of statistical surveys can be distinguished: **one-time** – also known as ad hoc surveys, with no plans for repeat performances (e.g., domain- and activity-related surveys on demand for information) and surveys based on **repeated observations** of major areas and domains of economic or social life. Business surveys are mostly carried out on a regular basis, e.g., monthly, quarterly, semi-annually, annually or multi-annually. In the case of business statistics, regularity and repeatability of surveys is required to provide not only data about the level of economic development at a specific time but also information about changes in the surveyed population or variables over time.

Detailed information on conducting repeated surveys, especially in terms of frame construction, sample design and estimation is presented in the module “Repeated Surveys – Repeated Surveys”.

Type of output information

In terms of output information we can distinguish surveys which provide **qualitative data** (e.g., conducted in Poland monthly Business Tendency Survey, based on entrepreneurs opinions and related, among others, to the current and prospective production, demand, financial situation, prices, employment, and barriers faced in respect of the conducted activity (CSO, 2012a)) and **quantitative data**, such as: indicators (e.g., STS), monetary and count values (e.g., SBS), prices (e.g., producer prices).

The above classification of surveys is based on a typology applied by CSO of Poland (CSO, 2012b). Using this typology, we can classify surveys conducted in other NSIs, but it does not exhaust all aspects of surveys and other classifications (like those described in subsection 2.3) can be used.

2.2 STS v. SBS

Structural business statistics, abbreviated as **SBS**, present the structure and main characteristics of economic performance in the European Union (EU) and each of the EU Member States.

Data are produced under the legal basis provided by **Regulation (EC) No. 295/2008 of the European Parliament and of the Council of 11 March 2008** concerning structural business statistics and a number of Commissions Regulations implementing and amending the Council Regulation, among

others (Eurostat, 2013b): Commission Regulation (EC) No. 251/2009 of 11 March 2009 and Commission Regulation (EC) No. 250/2009 of 11 March 2009.

Regulations establish a common framework for the collection, compilation, transmission and evaluation of statistics on the structure, activity, competitiveness and performance of businesses in the EU. In particular, according to the Regulation (EC) No. 295/2008, statistics provide information to analyse:

- the structure and evolution of the activities of businesses;
- the factors of production used and other elements allowing business activity, competitiveness and performance to be measured;
- the regional, national, EU and international development of businesses and markets;
- business conduct;
- small and medium-sized enterprises;
- specific characteristics of enterprises related to particular breakdown of activities.

SBS covers business economy, according to NACE Rev. 2, Sections B to N and Division 95, which contains industry, construction, distributive trades and services. SBS does not survey agriculture, forestry, fishing, public administration and largely non-market services, such as education and health.

There are two kinds of units being under observation within the confines of SBS surveys: enterprises and kind of activity units (KAU). Most statistics are created as a result of observations of enterprises or parts of enterprises (local units) conducting economic activity. When an enterprise consists of several legal units (sometimes at a few locations) and/or performs more activities, all surveyed variables are compiled under the enterprise's principal activity, which normally generates the largest amount of value added (Eurostat, 2012). The use of kind-off activity units for the compilation of statistics is specified in the sector specific annexes to the Regulations. For example, Member States are obliged to prepare KAU characteristics for industry and construction, i.e., value of production, turnover, number of persons employed, wages and salaries, number of kind-off activity units.

According to the Regulations, Member States may obtain required data using a combination of different sources: compulsory surveys, other sources (e.g., administrative) and statistical estimations procedures. Choosing the collection method, NSIs should take into account the cost of obtaining data, the response burden on enterprises and quality of data. In most countries SBS data are usually collected through statistical surveys and/or administrative sources. The advancement of the EU countries in using administrative data for producing SBS statistics is presented in the module "Data Collection – Techniques and Tools".

Main variables, compiled within the confines of SBS surveys are (EC, 2008):

demographic statistics:

- structural data, e.g., number of enterprises, number of local units;

enterprise statistics:

- accounting data, e.g., turnover, production value, value-added at factor costs, total purchases of goods and services, personnel costs, wages and salaries;

- data related to the capital account, e.g., gross investment in tangible goods;
- data on employment, e.g., number of persons employed, number of employees.

Detailed information on statistics transmitted to Eurostat (first reference year, frequency, activity coverage and level of activity breakdown) is included in sector specific annexes for industry, trade, construction, insurance services, credit institutions, pension funds, business services and business demography, being a part of mentioned regulations. Requirements concerning characteristics differ depending on sectors. Each sector have wider, then mentioned above, list of statistics to be compiled for the study of special subjects.

The reference period for STS data is calendar year, which usually corresponds to the fiscal year. Most of statistics is transmitted to Eurostat annually, however some specific characteristics (burdensome in collection) are compiled only multi-annually. The annual national enterprise statistics are available to the four digits level (classes) of the NACE classification. A subset of SBS information (e.g., wages and salaries, number of persons employed) is also accessible for European regions, according to NUTS (Nomenclature of Territorial Units for Statistics) classification, as well as enterprise size-class – defined by the number of employed persons (or by size of turnover in retail trade), combined with three digits level (group) of NACE classification (Eurostat, 2013c).

Results of SBS surveys are generally presented as monetary values, mainly concerning operating income, expenditure or investment and as counts, covering business demography and employment, e.g., numbers of enterprises, employees and persons employed. This constitutes the main difference with respect to short-term statistics, where data are shown as monthly and quarterly indices generally calculated with reference to a base year.

Short-term business statistics, also called short-term statistics and abbreviated as STS, describe current developments of the economies of the whole European Union (EU) and each of the EU Member States. STS indicators are used by many national institutions and organisations, like governments and central banks, companies and financial markets to analyse current economic situation in their states. Short-term information is in great demand in the European Commission (EC) and European Central Bank (ECB) to monitor the situation of the EU and the euro area and to conduct the monetary policy.

STS surveys are conducted on the legal basis provided by Council Regulation (EC) No. 1165/98 of 19 May 1998 concerning short-term statistics, amended by the Regulation No. 1158/2005 of the European Parliament and of the Council of 6 July 2005 concerning short-term statistics, the so-called STS Regulations (STS-R), and a number of Commissions Regulations implementing and amending the Council Regulation, among others (Eurostat, 2013):

- Commission Regulation (EC) No. 1503/2006 of 28 September 2006,
- Commission Regulation (EC) No. 657/2007 of 14 June 2007,
- Commission Regulation (EC) No. 1178/2008 of 28 November 2008,
- Commission Regulation (EC) No. 329/2009 of 22 April 2009,
- Commission Regulation (EU) No. 461/2012 of 31 May 2012.

Regulations establish a common European framework for collecting, processing and compiling short-term data on supply and demand, factors of production and prices in the European Union. They also stipulate ways of transferring data to Eurostat and confidentiality of sensitive data. Regulations oblige national statistical authorities to apply all these rules in the production of STS to ensure good quality of European aggregates, consistency and comparability between national statistics and make sure they reflect the actual condition of the economy.

The aim of STS is to provide current information on the situation of enterprises conducting economic activity in four major domains, defined by NACE rev. 2, as industry, construction, retail trade and repair and other services, for which, according to aforementioned regulations, the following indicators are compiled:

Industry:

- Production
- Turnover: Total, Domestic, Non-domestic
- Number of persons employed
- Hours worked
- Gross wages and salaries
- Producers prices (Output prices): Total, Domestic market, Non-domestic market
- Import prices

Construction:

- Production: Total, Building construction, Civil engineering
- Number of persons employed
- Hours worked
- Gross wages and salaries
- Construction costs, Material costs, Labour costs
- Building permits: number of dwellings, square metres of useful floor area

Retail trade and repair

- Turnover
- Number of persons employed
- Deflator of sales
- Hours worked
- Gross wages and salaries

Other services:

- Turnover

- Number of persons employed
- Producer prices (output prices)
- Hours worked
- Gross wages and salaries.

In contrast to SBS, STS statistics do not present absolute amounts or monetary values. They are released as indices generally with monthly (e.g., industrial production, retail trade turnover, producer prices in industry) or quarterly frequency (e.g., turnover in other services, producer prices in services, labour input indicators) to indicate recent developments in the European Union and in each of the EU Member State. In order to monitor or predict structural changes over time and show trends observed in the economies, indices are released in form of time series with reference to a base value, which is representative for a base year – i.e., for a monthly series, the base value is the monthly average during the base year (Eurostat, 2006). The base year (currently 2010 = 100), according to the STS regulations, is adjusted every five years (using the years ending with a “0” or a “5”).

STS data are very sensitive to the calendar effect. The number of working hours in a month affects, among others, the level of production or turnover. Indicators are also influenced by seasonal factors, such as holidays, the weather, events, tradition or habits. “In order to increase comparability between different periods, time series are adjusted for calendar effects (working-day adjustment) and seasonal effects (seasonal adjustment) ([link to Seasonal Adjustment](#)). Without such adjustments a figure for May (a month with many public holidays) might wrongly indicate a decline in economic activity. Similarly, a comparison between countries, e.g., between Sweden (holidays in June) and France (holidays in August) could be misleading” (Eurostat, 2011).

STS regulations don’t require (but allows) to transmit to Eurostat seasonally adjusted data, but they oblige Member States to compile working-day adjusted figures for six indicators (Eurostat, 2013a):

- Industrial production
- Production in construction
- Hours worked in industry and construction (since 31 March 2015 also in retail trade and other services)
- Retail trade turnover
- Retail trade deflator of sales
- Turnover in other services.

According to STS-R, in order to produce short-term statistics, Member States may acquire data using different collection methods: conducting compulsory surveys, using administrative sources, applying statistical estimations procedures, as well as combining data from mentioned sources. All Member States obtain most of data using statistical questionnaires. Some information, e.g., value of turnover, buildings permits or data concerning employment are derived by NSIs from administrative source. The existing practices in Member States for using administrative data for compilation of STS indices are presented in the modules “Data Collection – Techniques and Tools” and “Data Collection – Collection and Use of Secondary Data”.

2.3 *Types of statistical processes*

Considering different types of sources used during the statistical production and methods of data processing, there is another typology which is used in the European Statistical System (ESS). This division of surveys was created for the purpose of the handbook “*ESS Handbook for Quality Reports*”, which is aimed at providing detailed guidelines, recommendations and practical examples for preparation of comprehensive quality reports covering all steps of the statistical production processes and their outputs. According to the handbook we can distinguish the six following types of statistical processes:

“Sample survey. This is a survey based on a, usually probabilistic, sampling procedure involving direct collection of data from respondents. For this kind of survey there is an established theory on accuracy that allows reporting on well-defined accuracy components (sampling and non-sampling errors).

Census. This can be seen as a special case of the sample survey, where all frame units are covered. There are population, economic and agricultural censuses.

Statistical processes using administrative source(s). This sort of process makes use of data collected for other purposes than direct production of statistics” (Eurostat, 2009).

The overview of existing practises in the use of administrative data for producing business statistics can be found in the module “Data Collection – Techniques and Tools”. This module, in the subsection *Use of administrative data*, presents four domain, i.e., Business register, STS, SBS and PRODCOM, in which administrative data are applied to statistical purposes.

“When discussing accuracy, three main types of processes using administrative sources are distinguished: tabulations based on one register, integration of several registers, and event reporting systems.

Statistical process involving multiple data sources. In many statistical areas, measurement problems are such that one unified approach to sampling and measurement is not possible or suitable. For example, in a structural business survey in which basic economic data – production, finance, etc. – about businesses are aggregated, different units, questionnaires, sampling schemes and/or other survey procedures may be used for different segments of the survey. Furthermore, one or more segments may depend upon administrative data.

Price or other economic index process. The reasons for distinguishing economic index processes as a special type of statistical process can be described as altogether fourfold (not everyone being strong enough on its own): (i) there is a specialised economic theory to define the target concepts for economic indexes; (ii) their error structure involves specialised concepts such as quality adjustment, replacement and re-sampling; (iii) sample surveys are used in several dimensions (weights, products, outlets), mixing probability and non-probability methods in a complex way; and (iv) there is a multitude of these indexes playing a key role in the national statistical systems and the ESS.

Statistical compilation. This statistical process assembles a variety of primary sources, including all of the above, in order to obtain an aggregate, with a special conceptual significance. Mainly, but not only, these are economic aggregates such as the National Accounts and the Balance of Payments” (Eurostat, 2009).

It is obvious that the diversity in methods of producing ESS statistics requires a typology of statistical processes, but according to the handbook's authors, "defining these six types should be regarded simply as a pragmatic device solely for the purpose of the Handbook. It is expected that in the future new categories and improved distinctions will emerge, so such a typology can be drawn up in a variety of different ways" (Eurostat, 2009).

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

CSO (2012a), The organization of the annual microenterprise survey compared with business surveys carried out by the Central Statistical Office of Poland. Paper presented on International Seminar "Statistical observation of small entrepreneurship", Svetlogorsk, Kaliningrad region, July 2012.

CSO (2012b), Statistical survey program of official statistics, 2012. Web page of CSO of Poland: http://www.stat.gov.pl/bip/76_ENG_HTML.htm

EC (1998), Council Regulation (EC) No. 1165/98 of May 1998, concerning short-term statistics. *Official Journal of the European Union* 5.6.98.

EC (2005), Regulation (EC) No. 1158/2005 of the European Parliament and of the Council of 6 July 2005 amending Council Regulation (EC) No. 1165/98 concerning short-term statistics. *Official Journal of the European Union* 22.7.2005.

EC (2006), Commission Regulation (EC) No. 1503/2006 of 28 September 2006 implementing and amending Council Regulation (EC) No 1165/98 concerning short-term statistics as regards definitions of variables, list of variables and frequency of data compilation. *Official Journal of the European Union* 12.10.2006.

EC (2008), Regulation (EC) No. 295/2008 of the European Parliament and of the Council of 11 March concerning structural business statistics. *Official Journal of the European Union* 9.4.2008.

EC (2009a), Commission Regulation (EC) No 250/2009 of 11 March 2009 implementing Regulation (EC) No 295/2008 of the European Parliament and of the Council as regards the definitions of characteristics, the technical format for the transmission of data, the double reporting requirements for NACE Rev.1.1 and NACE Rev.2 and derogations to be granted for structural business statistics.

- EC (2009b), Commission Regulation (EC) No 251/2009 of 11 March 2009 implementing and amending Regulation (EC) No 295/2008 of the European Parliament and of the Council as regards the series of data to be produced for structural business statistics and the adaptations necessary after the revision of the statistical classification of products by activity (CPA).
- EC (2009c), Commission Regulation (EC) No 329/2009 of 22 April 2009 amending Council Regulation (EC) No 1165/98 concerning short-term statistics as regards the updating of the list of variables, the frequency of compilation of the statistics and the levels of breakdown and aggregation to be applied to the variables. *Official Journal of the European Union* 23.4.2009.
- EU (2012), Commission Regulation (EU) No. 461/2012 of 31 May 2012 amending Council Regulation (EC) No. 1165/98 concerning short-term statistics and Commission Regulations (EC) No. 1503/2006, (EC) No. 657/2007 and (EC) No. 1178/2008 as regards adaptations related to the removal of the industrial new orders variables.
- Eurostat (2006), *Methodology of short-term business statistics*. Office for Official Publications of the European Communities, Luxembourg.
- Eurostat (2009), *ESS Handbook for Quality Reports*. Eurostat Methodologies and Working Papers, Office for Official Publications of the European Communities, Luxembourg.
- Eurostat (2011), Short-term business statistics, 2011. Web page of Eurostat:
http://epp.eurostat.ec.europa.eu/portal/page/portal/short_term_business_statistics/introduction/sts_in_brief
- Eurostat (2012). Structural business statistics overview, 2012. Web page of Eurostat:
http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Structural_business_statistics_overview
- Eurostat (2013). Short-term business statistics. Legislation, 2013. Web page of Eurostat:
http://epp.eurostat.ec.europa.eu/portal/page/portal/short_term_business_statistics/legislation
- Eurostat (2013a). Short-term business statistics, 2013. Reference Metadata in Euro-SDMX Metadata Structure (ESMS). Web page of Eurostat:
http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/en/sts_esms.htm
- Eurostat (2013b). Structural business statistics, 2013. Web page of Eurostat:
http://epp.eurostat.ec.europa.eu/portal/page/portal/european_business/introduction
- Eurostat (2013c). Structural business statistics, 2013. Reference Metadata in Euro-SDMX Metadata Structure (ESMS). Web page of Eurostat:
http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/en/sbs_esms.htm#stat_pres

Interconnections with other modules

8. Related themes described in other modules

1. User Needs – Specification of User Needs for Business Statistics
2. Repeated Surveys – Repeated Surveys
3. Sample Selection – Main Module
4. Data Collection – Techniques and Tools
5. Data Collection – Collection and Use of Secondary Data
6. Micro-Fusion – Data Fusion at Micro Level
7. Seasonal Adjustment – Introduction and General Description

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

General Observations-T-Different Types of Surveys

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	31-08-2012	first version	Monika Natkowska	GUS (Poland)
0.2	02-12-2013	updated version according to reviews	Monika Natkowska	GUS (Poland)
0.2.1	18-12-2013	preliminary release		
0.3	15-03-2014	updated version according to the remarks	Monika Natkowska	GUS (CSO)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:22



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Data Collection: Techniques and Tools

Contents

General section.....	3
1. Summary	3
2. General description.....	4
2.1 Interviewer-administered techniques.....	4
2.2 Self-administered techniques.....	11
2.3 Data collection using EDI and XBRL	20
2.4 Use of administrative data	25
3. Design issues	28
4. Available software tools.....	28
5. Decision tree of methods	28
6. Glossary.....	28
7. References	28
Interconnections with other modules.....	30
Administrative section.....	31

General section

1. Summary

This module provides a description of the main techniques and tools used by National Statistical Institute (NSIs) to collect data. Characteristics and peculiarities of each of them will be described together with organisational aspects to build data collection instruments and to set up, run and finalise data collection, in accordance with the sub-processes 3.1, 4.2, 4.3 and 4.4 indicated by the GSBPM model for “Phase 4. Collect”.

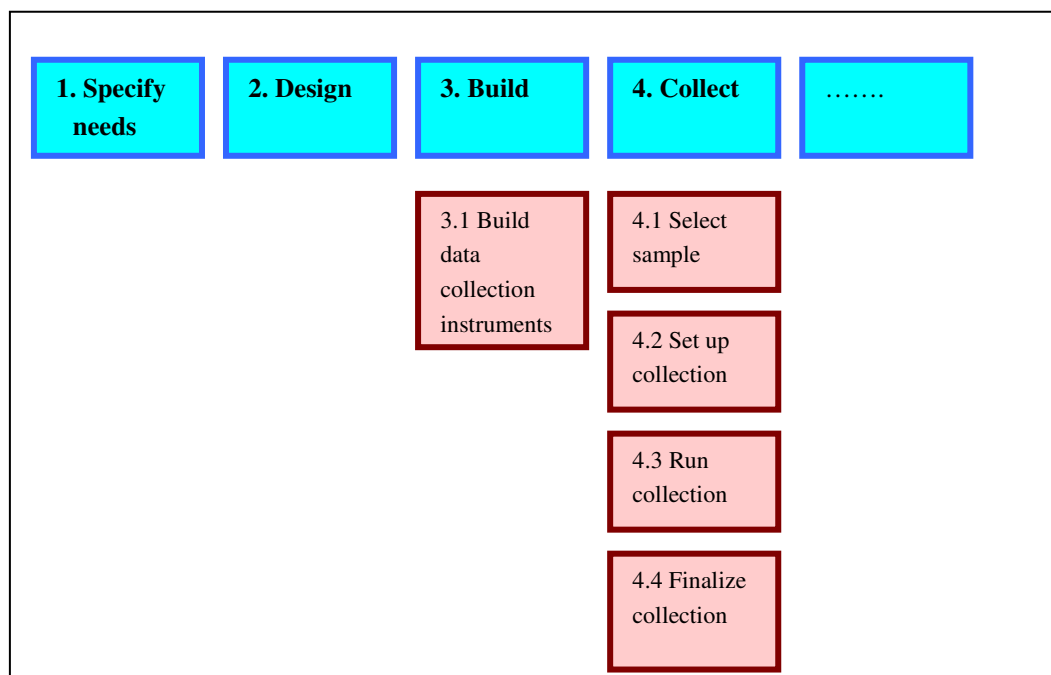


Figure 1. GSBPM – Phase 4: Collect

The choice of the most suitable technique or the way they can be combined is described in the module “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method” where the reader can also find in section 2.2 a possible classification of all available modes (Table 1).

In this module only the main and most used techniques and software tools to collect data for business surveys are described, dividing them into two main groups: interviewer-administered and self-administered modes. More specifically, the module is organised as follows:

Section 2.1 is about interviewer-administered techniques, CATI (Computers Assisted Telephone Interviewing) and CAPI (Computers Assisted Personal Interviewing). Here the advantages and disadvantages of the presence of interviewers and of the use of an electronic questionnaire will be described. The section will also make hint to the Direct Observation mode.

Section 2.2 is about self-administered modes: Mail and Web surveys. The potentialities of the electronic questionnaires typical of web surveys will be highlighted as well as the importance of good

questionnaire layout for mail surveys. The section will also talk about software tools for entering the information collected through the paper questionnaires used in mail surveys.

Section 2.3 talks about the structured electronic exchange of information based on EDI (Electronic Data Interchange) and XBRL (eXtensible Business Reporting Language) and the last section 2.4 is about administrative data, as their use is going to change the way NSIs organise their data collection process.

A note for readers: in the rest of this topic the term “respondent(s)” is used. With this term it is intended to represent all the “actors” involved in providing the information to be collected according to the surveys’ needs. Respondents can be defined as “*Respondents are businesses, authorities, individual persons, etc., from whom data and associated information are collected for use in compiling statistics*” (OECD Glossary of Statistical Terms). This definition, therefore, includes all the expressions like “reporting units”, “observation units”, “data provider”, etc. whose definition can be found the glossary, thus simplifying the reading.

2. General description

2.1 Interviewer-administered techniques

In this section interviewer-administered techniques are illustrated, focusing on CATI and CAPI modes. The last part is dedicated to Direct Observation that is treated apart since it is a particular interviewer-administered technique as it is not based on the interaction between interviewer and respondent.

Common features of CATI and CAPI are discussed right below, while characteristics of each technique (including also Direct Observation) are treated in the dedicated sub-sections.

The key features of CATI and CAPI are the presence of well-trained interviewers and the use of electronic questionnaires that, put together, allow the management of complex surveys, where complexity is in terms of survey content and structure of the questionnaire (questions, skipping and checking rules). In more details:

- the presence of well-trained interviewers allows to:
 - administer complex interviews in terms of survey content;
 - get in touch or interview difficult targets, like managers of large businesses;
 - find the right respondent especially in case one questionnaire has to be answered from different professional profiles inside the same company ;
 - collect data by directly observing a phenomenon (direct observation);
- the use of electronic questionnaires allows to:
 - have the paper questionnaire, the checking plan and the skipping rules, that are shown on the video screen of a pc, into the same software package;
 - increase data quality since editing can be performed during data collection (see “Questionnaire Design – Editing During Data Collection”) and, especially for CATI, very

complex checking plans can be managed. This aspect can also positively influence the editing and imputation phase making it simpler and faster;

- implement a set of indicators to monitor in real time data collection, making it possible to take the corrective actions in due time;
- reduce or avoid, with respect to self-administered techniques, follow-up calls;
- avoid the data entry, as data are sooner available in an electronic format, thus improving the timeliness of data;
- use data that are collected in previous waves of the survey or derived from other sources, like administrative data. In this way it is possible: *i)* to reduce the respondent burden since there is no need to ask again some information, *ii)* to improve data quality making comparison between these data and those collected during the survey;
- use the same electronic questionnaire for mixed mode surveys thus reducing the technique effect and programming costs.

The two features (presence of interviewers and of pc) combined together offer the opportunity to use, in CATI and CAPI questionnaires, any kind of question formats, including also open-ended questions. These can be coded on-line if the software, used for the questionnaire implementation, has got an assisted-coding function. The assisted coding is quite an important feature (see “Coding – Main Module”) because, although it might require a specific training session for interviewers, it guarantees high levels of standardisation and quality of the coded data since interviewers can use probing until they find a suitable code together with the respondent.

If the adoption of CATI and CAPI provides advantages in terms of response burden, data quality, amount of data entry and editing and imputation phase, on the other hand it requires considerable financial and organisational efforts. In fact:

- the presence of interviewers implies the setting up of the training and monitoring phases: the first is necessary to make interviewers able to collect data in the most objective way in order to reduce as much as possible the interviewer’s effect; the second is fundamental to keep under control the interviewers’ activity and to take the correct actions in due time;
- the IT feature implies some extra costs for software and hardware components: an initial cost for the questionnaire implementation and the creation of the entire IT infrastructure, plus a cost for testing and maintaining the applications. In particular for CAPI, there is the hardware cost represented by the laptops/tablets used by interviewers. Anyway these costs can be reduced if these techniques are used for periodic (not ad hoc) surveys.

The presence of interviewers makes it easier, with respect to self-administered techniques, to find the right respondent(s) for the questionnaire compilation. This is because interviewers can use appointments or can talk directly to people inside the enterprise that will address him/her to the correct target unit. Anyway, it is important, for both techniques, to plan a contact strategy to send to all the sampled units a pre-notice letter that will advise them about a future phone call or a visit from authorised personnel, thus creating a climate of trust. How to identify who the letter should be addressed to inside the company is fully described in the module “Data Collection – Design of Data Collection Part 2: Contact Strategies”. Here it is important to say that the letter has to specify the

content and aim of the survey, the deadline and, if possible, which is(are) the designated role(s) to answer the questionnaire. Sometimes a paper questionnaire is enclosed with the letter, like it happens for mail surveys and sometimes for web surveys, in order to let respondents know in advance which information is required. This is especially useful in case the questionnaire contains questions with very technical concepts only known by experts or when it is necessary to retrieve the information required, that can be contained in documents belonging to different business departments.

2.1.1 CATI – Computer Assisted Telephone Interviews

Data collection with CATI requires a central location where a call centre is settled. Interviewers work with desktop computers equipped with microphoned earphones. Each pc is a client connected to a central server that delivers the units to be interviewed according to the parameters that have been set in the scheduling system.

Being an interviewer-administered technique, a good management of the training phase and a constant monitoring of the interviewers' work during the entire data collection period are quite important for a successful data collection.

- The training phase is very important for data quality, because the communication with respondents is via telephone and it is therefore necessary that interviewers are able to create a climate of trust. In general, in CATI surveys, interviewers' training has to focus on:
 - how to assure respondents about the confidentiality of the information they provide;
 - the importance of finding the right moment to administer the questionnaire, by fixing appointments for days and times suitable for respondents;
 - how to probe for questions that are not immediately clear to respondents without influencing the answer in any way;
 - how to solve potential inconsistencies between answers through the use of error windows that contain messages whose text has to be customised according to the type of error.
- **Monitoring** the interviewers' job plays also a fundamental role on data quality and, in CATI surveys, this activity can be easily carried out because interviews are conducted in a centralised facility that allows for a daily storage of data on a central server always at disposal of the NSI. In this way a set of monitoring indicators can be implemented to indicate, day by day, if interviewers work respecting the instructions provided during the training session or, if they don't, which are the correct actions to be taken to improve their work. The set of indicators is designed by methodologists according to the survey's needs. Anyway it is advisable to use at least the following indicators:
 - number of completed interviews total and per interviewer;
 - number of other definitive contacts results (refusal and definitive interruption) total and per interviewer;
 - number of not definitive contact results (no-answer, busy, answering machine);
 - distribution of appointments per day and hours, total and per interviewer;
 - number of out of target units and reasons why;

- number of units with no contact.

This common set of indicators generally corresponds to the default one provided by the software itself and, therefore, methodologists have the opportunity to concentrate on the design and implementation of ad-hoc indicators about fundamental or very important survey variables that have to be strictly monitored. An example of an ad-hoc indicator can be the measurement of the time spent in assisted coding of open-ended variables, since in case it is too long methodologists can decide to train interviewers again or not to use this function as it can be too burdening for both respondents and interviewers, with negative effect on response rate. An ad-hoc indicator can be based, for example, on Control Charts that show how the values of the monitored variables varies along the time axis, indicating if variations depend on casualty or on systematic errors due to wrong interviewers' behaviour (Murgia et al. 2005).

Being a CAI (Computer Assisted Interviewing) technique, CATI can exploit the software potentialities to manage different aspects of the questionnaire, like different paths or branching or different types of controls (coherence controls, range controls and skipping rules) or texts' customisation for questions' wording and error messages (Blanke et al., 2006).

The questionnaire layout (see also "Questionnaire Design – Electronic Questionnaire Design") has to be designed in order to avoid the segmentation effect (one question per screen) and a too dense video screen to make the interviewer able to sooner find the information he needs, like explanatory texts, helps on line etc. Different colours and fonts should be used according to the different roles played by texts, for example, red for interviewer instructions, black for texts to be read to respondents, etc.

An important and peculiar feature of CATI is the call scheduler that allows methodologists to plan the contact strategy. Generally speaking, this means that methodologists have to establish in advance, on the basis of previous surveys or of pilot surveys, the time period of the day most suitable to run the interviews and how many times a unit (business) with no definitive contact results (no-answer, busy, appointments) has to be tried before assigning it a definitive contact result (interviewed, refusal, definitive interruption, not reachable) and then to substitute it with another one. Besides, thanks to the management of appointments, the scheduling system allows respondents to plan the interviewing time according to their needs thus improving the response rate. The scheduling system is described in detail in the module "Data Collection – CATI Allocation".

It has to be said that, among the CAI techniques, CATI allows for the implementation of the most complex electronic questionnaires because of the presence of well-trained interviewers and because interviewers work in call centres together with field supervisors that can provide immediate help in case of any doubt on how to proceed with the interview (while CAPI interviewers work alone) (Capparucci et al., 2009). Therefore, it is possible to make a "heavy" use of editing during data collection that can also manage many blocking edits (edits to be solved to proceed with the interview) more easily than other techniques. Anyway, the number and type of checking rules to be implemented in the electronic questionnaire have to be established by methodologist maintaining a good balance between data quality and response burden.

This unique feature of CATI must always be kept in mind when choosing the most suitable technique for a survey, but, anyway, it is not always usable in business surveys because, in general, complexity is synonymous of lengthy questionnaires that cannot be administered by telephone because long interviews surely decrease the response rate.

For this reason, the CATI technique is especially suitable only for specific types of business surveys or specific situations: *i)* for short interviews, *ii)* when data collection needs to be run in a very short period of time, like it happens for some agricultural surveys that have to produce timely estimates and preferably when *iii)* the compilation of questionnaires does not need the retrieval of the information from the respondent and *iv)* there is no need of different respondents to administer different sections of the questionnaire,.

To finalise data collection is quite simple in CATI surveys, because, at the end of the phase, data are sooner ready for the editing and imputation phase, that can be simplified because part of the data have already being checked during the data collection.

Nowadays CATI is mostly used for follow-up calls to non-respondent units (Parent et al., 1999) or to probe for answers that failed the edit procedure, or in mixed mode with other techniques like CAPI or CAWI (Computer Assisted Web Interviewing) Its use for business surveys is decreasing leaving room to Web surveys because, as explained later, web surveys are less expensive and are becoming more and more suitable to manage business surveys due to the increasing use of internet combined with the availability of administrative data and, in general, of other secondary source of information.

2.1.2 CAPI – Computer Assisted Personal Interviews

Data collection with CAPI requires a laptop or a tablet for each interviewer: data are stored in each pc and then periodically sent via LAN to the NSI's central server. As described in the following pages, CAPI requires a very good organisation to coordinate interviewers and to assist them both in terms of survey content and of software/hardware instruments. It is therefore a quite expensive data collection technique, that is generally used to administer interviews to the top management of large enterprises, that are difficult to reach by phone, or in mixed mode surveys, with CATI or CAWI, to cover those strata that have a response rate lower than the average one.

Like the other computer assisted techniques, CAPI presents the advantage of allowing the management of complex questionnaires (in terms of skipping and checking rules) and, in addition, the administration of long interviews. This is because interviews are run face-to-face and, in general, after having taken an appointment with respondents.

Being a CAI technique, the electronic questionnaire constitutes the core of the CAPI system. Anyway to run CAPI several other functions must be implemented (Budano, 2008). These are:

a) **interviewers database:** it is a centralised database necessary to organise at best the interviewers' job and to reduce their work burden. This database, therefore, has to be used to keep under control which are the active interviewers and which are not active (for holidays, illness, etc.) in order to organise their eventual substitutions with other interviewers;

b) **management of the laptops:** this aspect requires the organisation of local assistance to be provided on the territory in case of any software or hardware problems arise (i.e., possible pc substitution). Besides, a uniform configuration must be given to all the PCs that have to use the same strong authentication procedure - to inhibit the use by others – (Parent et al., 1999) and the same encryption program to guarantee a secure exchange of data. Finally, the organisation has to consider the management of the software package installed on the pc to keep it updated with respect to new software releases or to the software application in case of questionnaire changes. In general, a database containing all the events concerning the PCs is advisable;

c) **interviewers training:** it is a very important aspect for the success of the data collection because, differently from CATI, interviewers work alone and cannot rely on supervisors neither for software nor for content problems. The training phase must treat many different aspects (similar to those mentioned for CATI) like presentation of the interview and questionnaire content, electronic questionnaire management, how to face critical problems during the interview, technical aspects concerning the laptops management, how to manage the different functions of CAPI applications. A local contact reference with functions of supervisor is advised;

d) **allocation of sample units to interviewers:** the distribution of sample units among interviewers must be done according to the information contained in the interviewers database;

e) **contacts management:** respondents can be contacted by telephone to get an appointment or through visits at their location. In both cases, in order to guarantee all sample units the same chances to be contacted for an interview, it is necessary to plan a protocol of contacts that defines which are the possible sequences and amount of contact attempts to be done and which actions have to be taken before assigning a definitive non-contact result to the unit. Let's suppose contacts are made by telephone: in case, for example, of a sequence of "nobody answers" the telephone, then the action to be taken is "a visit at respondent's place of work"; or in case the unit has moved, then the action is "find the new address". At the end of the contacts sequence, the eligibility of the unit can be verified and it can be asked for an appointment to make the interview or it can be abandoned because *i*) it was not possible to contact it, *ii*) it was not possible to find the new address or *iii*) the unit refused the collaboration. All these contact results must be stored in the sample unit database and managed in a **contacts report** that can be electronically sent to the centre to monitor the interviewing phase;

f) **interviewers' agenda:** it is an important application in CAPI surveys to manage appointments and in general contacts with respondents. Therefore, it must be related to the contacts report chart because interviewers report in it all the events relative to the contacts with respondents like: changes of addresses, appointments to start or continue the interview, definitive contact results. The agenda is also related to the electronic questionnaire (in the same fashion as the scheduler for CATI surveys), because during the interview (when filling the electronic questionnaire) it is possible to register some contact results, like completed interview, appointment to complete the interview on another day, refusal to complete the interview. All these results will update the contacts report and the sample unit database;

g) **interview:** during the interview, interviewers put in practice what they have learned during the training phase. They have to read the question wording as it is, to give explanations when the respondent asks for them, to read carefully the error messages to try to solve consistency errors together with the respondent, to manage critic situations which could compromise the completion of the interview, to take notes of any problems/difficulties encountered;

For all these reasons, the electronic questionnaire design is fundamental (see also "Questionnaire Design – Electronic Questionnaire Design"). It must be made so as to reduce the so-called segmentation effect (Blanke et al., 2006), to make clear to the interviewer which parts are to be read to respondents and which not and where he/she can find help functions concerning both technical problems or variables definitions. A good electronic questionnaire design can be useful to reduce the interviewer effect. It must be easy to assign the contact results from the electronic questionnaire, so as to update automatically the contacts report, the interviewer's agenda and the sample unit database.

h) **exchange of data from/to the centre:** it necessary to manage the exchange of information from and to the centre (NSI): the NSI sends interviewers data about units they have to contact and receives from them data concerning completed interviews and contacts results (possibly after each working day). In this way it is possible to monitor the interviewing phase and to take in due time the right actions in case of problems. All electronic data exchange must be done guaranteeing security in terms of data integrity and privacy requirements. These features are obtained through the use of secure protocols and data encryption. The electronic data exchange function must be easily called by the interviewers, possibly by the agenda;

i) **monitoring of interviewing phase:** a monitoring system must be defined, to be daily analysed by the survey manager. It should be updated with data sent by interviewers and should process these data automatically to produce synthetic indicators to take under control different aspects of data collection. In particular, it should monitor: the state of the art of the interviewing phase (units to be contacted, not eligible, to be substituted, etc), the respect of contacts protocol by the interviewers, the interviewers productivity and any odd behaviours of interviewers, like, for instance, interviews too long or too short, too high units substitution rates, etc.;

j) **finalised data collection:** at the end of data collection, data are ready for processing. Like in CATI, the editing and imputation phase can be simplified as part of the data had already being checked during the data collection. Anyway, as interviewers work alone, it is advisable not to implement a “heavy” editing during data collection, leaving the correction of inconsistencies to the revision phase.

2.1.3 *Direct observation*

Direct observation is another way to perform data collection. With respect to the other modes, data are directly “observed” by the interviewers with no need of asking the information to respondents and therefore no response burden exists. It can be conducted with or without the support of computer and therefore all the elements relative to the support of the computer can be found in the previous sections.

The organisation of the data collection is similar to that of CAPI, since interviewers are spread over the territory, but the role played by the interviewer is even more delicate as he/she is the only observer of the phenomenon. Therefore, skilled interviewers on statistical methodologies and IT tools have to be used for this kind of survey and consequently a very deep and detailed training phase has to be managed by the NSI. For these reasons, Direct Observation is a very expensive technique. It is generally used for surveys on pricing and for some agricultural surveys aimed at estimating types and areas of crops (Statistics Canada, 2010).

An example of a survey based on Direct Observation is the “Survey on Prices of Consumption” carried out by Istat- Italian National Institute of Statistics. The collection of prices is performed in two different modes:

- a territorial collection for the most part of goods and services conducted by local offices;
- a centralised data collection performed by the central office about goods and services which have uniform prices at a national level.

The territorial collection is run through the direct observation of prices: interviewers use tablets where an ad-hoc software, developed in Istat, is installed. Data transmission to the central server is in real time through the use of the 3rd generation mobile technology.

The centralised data collection extracts information on databases that are available on the web or from specialised websites (for example, prices of train or air tickets).

2.2 *Self-administered techniques*

This section describes CAWI and mail surveys that, being both self-administered techniques, share several aspects to be taken into account when choosing the data collection mode. These common features are discussed right below, while peculiarities of each technique are treated in the two dedicated sub-sections.

When these techniques are used, respondents have, in general, more time to provide their answers and therefore these two modes can be adopted in case of long interviews or when respondents need to retrieve the information to answer the questionnaire or in case more respondents are needed to answer the same questionnaire.

At the same time respondents have to be guided and helped in the questionnaire compilation, as they cannot rely on the help of any trained interviewers (Couper, 2001) and, of course, they are not trained on how to answer. To this aim the questionnaire and its layout together with the instructions for questionnaire compilation are of an extreme importance:

- the questionnaire and its layout have to be designed with criteria different in the two techniques, but following two common rules: 1) they have to provide respondents with all the information they need without being chaotic or confusing, 2) they have to arouse and keep always high the respondents' interest (Istat, 1989).
- instructions have to cover two main information areas, one on content and one on technical aspects:
 - in terms of content, instructions have to make clear and understandable the meaning and the aim of each question. Besides, if necessary, they have to clarify which professional figure(s) has to answer the various questionnaire sections;
 - about the technical aspects, instructions have to inform on how to fill each question and on how to navigate among them. In other words, they have to make respondents immediately understand how and which questions have to be answered.

An important aid for respondents is represented by an “information point” they can easily contact to get any information they need. This is represented, in general, by a toll free line or an e-mail address managed by field staff that has been trained on how to answer all possible requests that can be about the content of the surveys, the technical aspects of the questionnaire or about organisational aspects of the survey.

Another issue to be managed for both techniques is the reminder strategy. As described in “Data Collection – Design of Data Collection Part 2: Contact Strategies”, it is important to plan in advance “when” and “how many” reminders should be sent in order to control the unit non-response rate without increasing the response burden with too many or not well addressed reminders. The reminder strategy has to be planned together with the reporting of unit non-response – coding of the reasons why a unit cannot be enumerated - in order to make reminders more effective and to take the correct actions in due time (substituting the unit, searching for the new address, telephone contact to the unit, etc).

Finally, a common element to be managed is the partial non-response typical of self-administered techniques: to keep it under control, it is necessary to properly design the questionnaire and the compilation instructions and to well organise the follow-up phase, that can be conducted by means of telephone interviews aimed at obtaining answers for those questions with missing values.

All these elements are common to CAWI and Mail surveys but they are implemented in different ways since the way these surveys are run is different.

2.2.1 CAWI – Computer Assisted Web Interviews

The use of web surveys is increasing in general and in particular for business surveys because among enterprises the use of software and hardware equipment is higher than for households/individuals. In fact, web surveys require, at respondent side, the presence of a computer equipped with internet services and obviously that respondents are acquainted with them. On the NSI side, they require a secure web server accessible from internet where to create web pages containing the questionnaire and the entire data collection IT infrastructure. Another important factor for its increasing use, is that the WWW offers the “lowest cost survey environment” especially for ongoing surveys: minor cost of data transmission, no postal charges and less cost for telephone fees (Clayton et al., 2000).

Web surveys are based on Computerised Self-Administered Questionnaires (CSAQ) that can be answered on-line or off-line:

- in the first case, respondents log on to a secure website and enter their data, or can upload dataset of survey data as explained in sections on EDI and XBRL;
- in the off-line case, respondents download, on their pc, an executable file or a “flat” file (excel, pdf, csv, etc.) containing the questionnaire or the structure (record layout) of microdata. Data are then sent back to the NSI via a secure e-mail system or by a fax-server and are automatically stored on the survey database.

On-line compilation assures timely data and a greater control on data collection from the survey manager. This implies a higher data quality but at the same time a higher response burden. Off-line compilation facilitate respondents’ co-operation especially in case more respondents are needed to answer the questionnaire, but there is a lower control on how questionnaires are filled in. To enhance response rate, both alternatives can be available. Besides, the paper questionnaire should always be downloadable: in this way respondents can print it and use the paper format as an aid to answer on the web.

To obtain a good response rate, it is necessary to consider many factors for the management and organisation of web surveys. The main ones are described below:

- **Cover letter:** a cover or pre-noticed letter about the starting date of the survey has to be sent to the sample units that were selected during the sample design phase. The letter should contain information about the survey, like its content, its aim and its deadline, together with the web address of the survey, the id-name and (temporary) password that respondents will use to register on the survey web site (see also “Data Collection – Design of Data Collection Part 2: Contact Strategies”). The letter can be sent also to the e-mail addresses of respondents, if an updated list of them exists. For short-term statistics the cover letter can be sent repeatedly at each survey round by e-mail or fax. The question about “to whom” address the letter (to a specific person or to a

designated role) is deeply described in the module “Data Collection – Design of Data Collection Part 2: Contact Strategies”. Here it would be sufficient to say that for business web surveys it is more crucial than for the other techniques to know in advance who the respondent unit is. This is because, apart from the need to retrieve information that can be located in different business departments or the presence of technical questions only known by experts, it can happen that respondents need authorisations before submitting the questionnaire. This last thing is especially true for web surveys because the questionnaire compilation requires a person able to use the pc that might not correspond to the target person. Therefore, it would be advisable to let respondents know in advance the content of the interview, by giving them the opportunity to download a paper questionnaire from the survey website, or by sending it by e-mail or as a last chance by post (environmental issues).

- **Login procedure:** for a well organised web survey the login procedure and the correlated issues of data security and privacy have to be managed. It is very important to implement an easy login procedure for respondents who, at the same time, have to feel sure about the respect of confidentiality of the information they send via web. Access procedures must be set-up in accordance with the national privacy laws. One way of managing them is to provide each respondent (using a paper letter or an e-mail) with a user-id and a temporary password for the first access and then allow the change of the password with a personal one that has to respect the common standards on passwords. For all subsequent logins for the same survey only the personal password can be used and, as it is known only by the respondent, it guarantees the respect of data confidentiality.
- **Questionnaire and its layout:** as web surveys are based on self-administered questionnaires, it needs to pay a great attention to the way the questionnaire appears on the video screen of the respondent’s pc. In fact, like the other CAI techniques, the layout has to be designed in order to avoid a screen too dense of information to make the user able to easily find what he needs to answer the questionnaire. In addition, for web surveys, it is important to implement an easy navigation of the questionnaire and to avoid vertical or horizontal scrolling that makes navigation more difficult and therefore burdensome (O’Neil, 2008). Like the other CAI techniques, it has to contain automatic skipping rules and customisation of questions’ wording. The questionnaire has to be designed in order to be easy to understand and to complete (Couper, 2001), it must keep respondents’ attention always at high level to make them able and willing to provide the optimal answers and has to make respondents sure about the confidentiality of their answers. Online help is extremely important: it should appear under an icon easily recognisable by respondents and has to contain information on words, concepts, questions’ aim and also instructions on how to fill in the questionnaire.

The pc support allows for the implementation of any type of questions including the open-ended ones for which an assisted coding function, easy to use, can be provided. Anyway, to avoid response burden, the use of open ended questions should be limited to variables easy to code (i.e., the place the establishment is located) or to variable respondents are used to answer (i.e., NACE¹ sector).

¹ NACE, is the nomenclature of economic activities in the European Union.

- **Editing during data collection:** this feature provides the same advantages described for CATI and CAPI. Anyway for web surveys, it is more important than for the other modes to establish good balance between edits and quality of data: since there are no trained interviewers for the questionnaire compilation (Couper, 2001), a too high number of error messages during the questionnaire administration can increase response burden, lowering the quality of the answers, and can induce respondents to skip questions or to stop their cooperation before the very end of the questionnaire. In general, for web surveys it is recommended to use edits during data collection only for crucial or important survey variables (that are defined during the questionnaire design phase). Furthermore, for these variables, edits can be blocking edits, meaning that the compilation cannot proceed until the error is solved. For the other not fundamental variables, it is instead advisable to implement warnings, also called soft edits, that have the advantage of making respondents aware of possible inconsistencies in the information they have provided and to leave them the choice of solving the edit failures or not: this “freedom” reduces respondent burden.

A peculiarity of web surveys is the possibility of managing edits on server-side and/or on client-side:

- usually only the edits on server-side are chosen: this means that each time the respondent presses the “submit button” (that could be placed at the end of each page/section or at the end of the questionnaire) data are stored on the database located on the central server. The database contains also all the checking rules and, in case of inconsistencies among the submitted data, errors messages appear to the respondent that is asked to solve the errors. This way of managing edits is the right one for short and not complex web questionnaires, like those generally used in business surveys. It is not suitable for long and complex questionnaires since it would imply too many edit messages at the end of the questionnaire or a too high LAN traffic in case data submission is done at the end of each page (Capparucci et al., 2009);
 - implementation of edits on client-side is strongly suggested for long and complex questionnaires. This solution has also the advantage of solving edits as soon as they happen and therefore to store only consistent data on the central server. The drawback is that edits on client-side need that software able to manage them (in general *Javascript*) is active on respondent’ pc. If this is not the case², the presence on checks on server-side will solve the problem, meaning that checks on server side must always be implemented in a web application³.
- **Hardware platforms, software systems and browsers:** a typical feature of web surveys is that respondents use PCs with different hardware components, different platforms and different software systems (Couper, 2001). The consequence is that the questionnaire might not work

² Respondents might be provided with a link to download and then install the needed software, but, in general, it is not a good practice to ask respondents to do this because it is burdening and because they are not so prone in installing new software on their pc.

³ The use of Ajax for asynchronous communication between client and server can be seen as a compromise between the two alternatives. Ajax is an acronym for Asynchronous JavaScript and XML. It is a group of interrelated web development techniques used on the client-side to create asynchronous web applications.

properly with some of them or it can be visualised in different ways. A typical example is the use of different browsers that might visualise a simple single-choice question in a completely different way. To avoid all of this a further effort is required to the NSI when designing and implementing the electronic questionnaire because it can be necessary to use ad hoc source code for the development of the electronic questionnaire. This extra work is quite important to reduce respondent burden and to control the non-response rate, because no extra efforts than questionnaire compilation have to be asked to respondents: the web application has to function properly without requiring the installation of any extra software components on the respondents' PCs or the use of specific web browsers.

As mentioned for the other CAI techniques, the presence of hardware and software technology can represent a drawback in a financial sense of the word (O'Neil, 2008), because it is costly to take all the above mentioned actions aimed at reducing the respondents' fatigue. Anyway, if the application is used for periodic surveys, the NSI can easily write off the initial cost.

- **Response time of the web application:** another aspect to be taken into account for the management of web surveys is the response time that represents here the period of time that a respondent (pc client) has to wait to get answers to his queries to the server. For example, the time to be waited after the submission of an answer and the administration of the following question. The entire web application must be implemented in such a way that response time is reduced at minimum levels, in order to avoid respondents abandoning the interview before completing it. Crash tests to establish how many contemporary accesses to the web site are possible with no delay in response time are therefore highly recommended before the beginning of the survey.
- **Partial submission:** it is important to give respondents the possibility of partial questionnaire submissions to let them answer the questionnaire in different moments of the day when they have time. This is particularly true for those surveys that need an information retrieval and/or different respondents to answer the various questionnaire sections.
- **Monitoring system:** in order to keep the data collection phase under control a set of real time indicators has to be implemented. Statisticians can build their own monitoring system that, anyway, should at least report the following indicators:
 - the number of completed interviews;
 - the number of partially completed interviews;
 - the number of refusals;
 - the number of those units that have made the registration but have not answered any questions;
 - the number of those units that did not register themselves.
- **Reminders:** as already said, the reminder strategy is fundamental for self-administered interviews to control the unit non response rate. In web surveys, reminders are generally done according to the results of the monitoring system and through different means of communication like the e-mail address, fax, or telephone.
- **Finalise data collection:** at the end of data collection data are ready for processing and for the editing and imputation phase that cannot rely on already checked data (if compared to CATI and CAPI), due to a limited use of editing during data collection. Anyway, the consistency of final

data can be more easily reached if the questionnaire has been designed following a metadata-driven approach or (Iverson, 2009) any techniques for relational database design, like the “Entity-Relationship scheme (E/R)” (Chen, 1976). In these ways data collected by filling the questionnaire are immediately stored into the relational database, underneath the survey, according to its designed structure that contains data tables, data links and data constraints.

The web site that hosts the survey plays a fundamental role for the success of the survey in terms of response rate, since it represents the “contact point” between businesses and the NSI. The web site should host not only the questionnaire but all the other data collection instruments aimed at supporting respondents in the self-administration. These are:

- instructions on how to access the web;
- instructions on how to answer the questionnaire,
- information on survey contents and aims,
- contacts for any questions or problems,
- list of FAQs,
- contact information of each enterprise (address, telephone number, e-mail address, etc.), that respondents can update after their registration.

All these elements have to be easily accessible from the web site that should be compliant with the following general requirements (Balestrino et al., 2006)⁴:

- to present a homogeneous and stable image of the Institute on the outside;
- to guarantee sender and receiver of each other identity;
- to guarantee the confidential nature of data and the comprehensive environmental security during the collection process;
- to minimise the impact on the operational environment of the external user;
- to replay to the user about the operation he carried out with a confirmation message;
- to favour the monitoring activity about data collection;
- to favour the internal management of the operations related to the data collection;
- to contain costs.

The future of web business surveys is represented by the “Business Statistical Portal”, a new way of organising and managing the data collection process for business surveys and already active in some European countries. This model allows to abandon the usual stovepipe model used for the production of business statistics which is “survey centred”, adopting a model which is “enterprises centred” and based on integrated production processes that will make NSIs able to organise more efficiently data collection, data processing and data estimation processes.

⁴ In Istat – Italian National Institute of Statistics – a web site dedicated to web surveys, named Indata (<https://indata.istat.it>), has been implemented since the ‘90s with the aim of presenting a unique front-end for respondents to any surveys.

The Business Statistical Portal will strongly reduce the response burden and therefore cost. This is thanks to the integration of administrative data and data provided by enterprises through the use of simple procedures that will allow asking only once information common to all surveys the enterprise is involved in.

A Business Statistical Portal should be compliant with the following requirements:

- it should allow for the sharing of data and metadata on the basis of a common data modelling;
- it should provide a centralised governance for the data collection processes respecting the businesses' needs;
- it should manage a back-office activity that allows to monitor the production of business statistics;
- it should allow the re-use of data that are already available in the statistical system or among the various public administrations, promoting also new protocols for data exchange;
- it should use IT instruments that make simpler and less expensive the exchange of information.

2.2.2 *Mail surveys*

Data collection for business surveys can be run by paper questionnaires which are sent to the target units by post and sent back to the NSI still by post or by fax. Due to low response rate and the environmental impact caused by the use of a great amount of paper, the adoption of this technique for business surveys is decreasing in favour of the use or the combined use (mixed mode) of those assisted by computer especially the web based ones. Anyway this mode has still a great importance to collect information for various reasons (explained in its advantages listed below) among which the low cost plays the major role.

It has advantages and disadvantages (see also “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method”) that are typical of self-administered techniques based on paper questionnaires and that influences the way data collection is set-up:

Advantages:

- as already said, it requires a low budget effort;
- it is quite useful when respondents need time to answer because: *i)* the interview is long or *ii)* it is necessary to retrieve information before answering the questionnaire or *iii)* the questionnaire contains questions with very technical concepts only known by experts or *iv)* because the questionnaire has to be filled in by different persons inside the same enterprise;
- due to its low cost, samples can be larger than those used with other techniques (keeping budget constant);
- the questionnaire can contain difficult questions (calculation, ordering, etc.);

Disadvantages:

- it has low response rate that requires to plan a reminder strategy (as described in the following) and therefore a longer data collection period;
- it has a high risk of partial non-response or of incomplete questionnaires that requires the design and setting of a more accurate editing and imputation phase;

- data are not soon available since there is a need to plan the data-entry phase to finalise the data collection (as described in the following).

In setting-up a data collection with mail surveys different elements should be taken into account as described in the following list.

- **Sending material to sampled units:** sampled units have to receive by post all the material necessary to participate to the survey (see also “Data Collection – Design of Data Collection Part 2: Contact Strategies”). Monthly or quarterly deliveries of all the material are generally planned for short terms statics. In these cases e-mail or fax are used instead of a mail-out system.

In general a unique envelope is sent, containing the following material (Istat, 1989):

- a cover letter that, apart from describing the content and aim of the survey (similarly to any other techniques) has to explain to respondents *i)* what they are asked to do, *ii)* the importance of their cooperation, *iii)* what they can do in case of any doubts, *iv)* which telephone number or e-mail address they can contact, *v)* how confidentiality is guaranteed and *vi)* acknowledgments for their collaboration;
 - instructions for questionnaire compilation, that explain the content of each question, the meaning of concepts, how to fill in questions and how to read navigation instructions. To avoid a too long questionnaire, instructions are in general written on separate paper sheets. It is advisable to re-write instructions (in a shorter format) on the questionnaire itself, next to the question they refer to. In fact, it has been tested (Istat, 1989) that respondents tend not to read instructions if they are written in documents different from questionnaire;
 - the questionnaire that, if possible, should be customised with pre-printed information like for instance enterprise master data (name, address, ect.);
 - a pre-paid returned envelope to send the questionnaire back to the NSI with no additional cost for respondents.
- **Questionnaire:** the questionnaire design and its layout (see also “Questionnaire Design – Main Module”), in other words, the way questions and instructions are organised and graphically represented, are extremely important for mail surveys even more than for web surveys, because in mail surveys the questionnaire is static and not dynamic (Couper, 2001) and therefore it is not possible to create different versions that are customised according to the interview flows. The first page of the questionnaire should contain a short presentation of the survey and must indicate a code (a bar-code or an alphanumeric sequence of characters) that represents the univocal key assigned to each respondent. This key is repeated in all pages to help finding and joining separated pages of the questionnaire and it is fundamental to link survey data to each respondent during the data registration phase. Besides in case an enterprise has more local units involved in the survey, it has to be created in such a way to link all the questionnaires.

Different fonts and colours should be used for texts according to their functions, in order to make respondents immediately understand if they are reading a question or an instruction or the section header. For questionnaire background a light colour should be chosen (Fanning, 2005) in order to create a contrast with texts that can be better read. Obviously it is important not to use too many colours that might increase respondents’ fatigue.

Questions should be organised according to a logic flow and, in general, this means to group into a section those questions referring to the same theme (see “Questionnaire Design – Main Module”). If possible a questionnaire page should correspond to a questionnaire section and, in any case, the beginning and ending of each section has to be made clear by using lines, boxes or other graphical elements.

Response items must be placed on the same page of the question they refer to, because if written on the next page there is the risk that respondent may miss reading them, creating a potentiality for measurement errors on data. Besides, to allow data-entry with specific software, they have to be numerated and a box for checking the answer has to be placed next to them.

Skipping instructions have to be graphically represented through symbols or short instructions (i.e., >>, →, goto, etc.) placed close to the filter questions in order to facilitate respondents in filling the correct branches of the questionnaire (see “Questionnaire Design – Main Module”).

Any types of question format can be used, although open ended questions should be used rarely. There are three main reasons for this recommendation: *i)* hand-written material is difficult to be registered both from the data-entry operator and the OCR (Optical Character Recognition) software; *ii)* the content of the answer could be meaningful or generic or ambiguous since there is no interviewer probing to get a meaningful answer and *iii)* a coding phase is necessary at the end of data collection that becomes more costly in terms of time and resources.

- **Reminder strategy:** reminders are necessary to increase the response rate and should start when questionnaires arrivals at NSI start decreasing regularly. This does not apply to short term statistics for which the first reminder is generally sent before the end of the data collection period (see “Data Collection – Design of Data Collection Part 2: Contact Strategies”). Reminders can be done by telephone or by post. As it may happen that some questionnaires are missed or do not reach respondents, it would be advisable to send, during the first or the second reminder, another questionnaire (Istat, 1989) paying attention, at the end of data collection, to the presence of duplicated questionnaires. The structure of a reminder (by letter or telephone) should be the following:
 - a kind but determined invitation to answer the questionnaire;
 - to re-state the importance of respondent’s cooperation;
 - apologies for those who already answered or did not receive the questionnaire.
- **Organising the data-entry phase:** the information collected through paper questionnaires has to be gathered and stored in an electronic format. This is done through data-entry with specific software or OCR systems that are described in the following two sub-sections.
- **Finalising data collection:** at the end of the data entry stage, the NSI has at its disposal a set of raw data as supplied by respondents, that is used as input for the editing and imputation phase where all types of inconsistencies are treated and solved. Besides, the raw data set enables statisticians to carry out systematic error analysis, which might be interesting for testing the clearness of questionnaires. Furthermore, by saving the original data the value added of editing operations can be determined. Thirdly, during subsequent stages of the processing, discussion

might arise as to the correctness of certain edits. This holds in particular when consistency checks with data from other surveys reveal differences between edited data (Willeboordse, 1998).

2.2.2.1 Data-entry with specific software

Data-entry can be done by means of an electronic questionnaire which is developed using specific software (like Blaise or CSPro)⁵. In case of mixed mode, the same program used for the other(s) CAI technique(s) can be used for data entry, thus saving implementation time and costs. The electronic questionnaire has to be developed in such a way to make the typists' job easy: this means that the electronic questionnaire layout should be quite similar to the paper one and no blocking edits have to be implemented. This is because typists are not trained on how to solve inconsistencies and the only thing they can do is to compare the entered data with the data on the form. In general, only soft consistency edits are implemented to reduce typing errors while blocking edits are about the questionnaire key number to avoid duplication of the same questionnaire. Anyway if the amount of questionnaires is limited the revision phase can coincide with the data-entry one. This requires the organisation of a training session for typist about how to solve inconsistencies.

2.2.2.2 Data-entry with Optical Character Recognition (OCR)

Another way to electronically store paper questionnaires is OCR that is particularly suited to large data collections. It allows simple edit checks, like valid values and value ranges. Readability is the crucial factor and shortcoming of this method: statisticians should take into account that the method does not enable systematic controls on the readability of the data reported, that numbers are more easily readable than plain text and that hand-written material is more difficult to be recognised than typed data. Although modern OCR packages use dictionaries and quite sophisticated software for texts recognition, the main caution to be considered is that OCR requires very accurate questionnaire layout and printing standards to ensure that the answers can be read by the sensors correctly.

At the end of the data collection by OCR, raw data are submitted to a program that checks which records have had problems in recognition of texts. Those records that fail this check are then submitted to the video screen correction and correct texts are inserted manually on the bases on what reported in the paper questionnaires. After this phase, the set of raw data is ready for the editing and imputation phase.

2.3 Data collection using EDI and XBRL

2.3.1 EDI: Electronic Data Interchange

EDI represents the “*Electronic exchange of data usually in forms that are compatible so that software or a combination of individuals and software can put the data in a compatible form at the receiving end if necessary*”. (SDMX, 2009).

⁵ Blaise is a computer-assisted interviewing (CAI) system and survey processing tool developed by Statistics Netherlands. CSPro - Census and Survey Processing System - is a public domain software package for entering, editing, tabulating, and disseminating census and survey data.

EDI offers businesses the opportunity to retrieve information electronically from their internal systems and to forward that information to trade partners/suppliers/customers/government through a communications network (from Context of SDMX, 2009).

The use of EDI requires standardisations from both technological and conceptual sides (Willeboordse, 1998):

- since businesses develop their own information system on the basis on their needs, it is necessary to map the concepts and then to standardise them accordingly to the statistical use. Conceptual dissimilarities may concern: *i*) the naming and coding of data items, *ii*) the level of aggregation of data items - a statistical item may be composed of different accounting items -, *iii*) the existence of data items - a statistical concept may have no accounting counterparts-. This standardisation has to end up with a standardised set of metadata to be used in any kind of business surveys;
- data should also be organised in a standard technical form in order to be readable by the NSI.

Therefore, implementation of EDI for data collection comprises the design of an electronic *translation* facility (Willeboordse, 1998) in order to bridge the technological and conceptual gap between the worlds of respondents and of NSI. As a consequence, enterprises have to use software for the translation and therefore they will have some start-up costs to adapt their information system. This cost also includes an overlapping testing period where both the old data collection method and the new EDI system are used. Besides these costs depends on the nature and complexity of edification projects that varies among surveys and depends on (Willeboordse, 1998):

- the distance between business accounting systems and information needs, with respect to the technological and conceptual dissimilarities as mentioned above;
- the degree of standardisation of business accounting practices.

The greater the distance and the lowest the degree of standardisation the higher the initial costs that anyway can have a counterpart in the reduction of respondent burden that can also be reduced if the NSI supplies standard software packages for free to the respondents.

This means that the use of EDI as a mean of data collection has an impact on the entire statistical process. It reduces respondent burden because it avoids the compilation of questionnaires and because it requires the harmonisation of similar or equal questions asked in different surveys. It improves the timeliness of data since it reduces the time that elapses between data collection and data processing, it can improve statistical integration since same data can be used for different statistical figures (Hans R. Stol).

Examples of the use of EDI are UN/EDIFACT and GESMES.

UN/EDIFACT - United Nations / Electronic Data Interchange For Administration, Commerce and Transport (<http://www.unece.org/trade/untdid/welcome.html>) is the international EDI standard developed under the United Nations. It comprises a set of internationally agreed standards, directories, and guidelines for the electronic interchange of structured data, between independent computerised information systems. In particular the EDIFACT standard provides:

- a set of syntax rules to structure data;
- an interactive exchange protocol (I-EDI);

- standard messages which allow multi-country and multi-industry exchange.

Recommended within the framework of the United Nations, the rules are approved and published by UNECE in the UNTDID (United Nations Trade Data Interchange Directory) and are maintained under agreed procedures. EDIFACT has been adopted by the [International Organization for Standardization](#) (ISO) as the ISO standard ISO 9735.

GESMES - Generic Statistical Message

(http://www.sdmx.org/docs/1_0/SDMX%201_0%20SECTION_04_SDMX-EDI.pdf)

It was developed by a group of European statistical organisations working within the international UN/EDIFACT standards body. GESMES was accepted as UN/EDIFACT Status 1 messages in 1995 and was first published in the UN/D95A directory. The statistical office of the European Union, EUROSTAT, who has lead the development of statistical UN/EDIFACT messages, is implementing GESMES into the data flows between it and the Member States of the EEA (European Economic Area) and promoting the use of the messages by other international organisations and by other sectors.

GESMES has all the features required to exchange multi-dimensional arrays and time series data, including metadata (such as attributes and footnotes). The advantage of using GESMES, in preference to a proprietary data format, is that it is an internationally agreed standard which is both open and fully functional. It is not tied to the format and constraints of one particular application. In particular GESMES supports the exchange of: metadata, multi-dimensional arrays, time series, administrative data.

An application of GESMES is GESMES/TS - GEneric Statistical MESsage for Time Series – (<http://stats.oecd.org/glossary/detail.asp?ID=5874>) which is a [data model](#) and message format (a GESMES profile) allowing the exchange of statistical time series, related attributes and structural definitions using a standardised format. The initial name of GESMES/TS was GESMES/CB (GEneric Statistical MESsage for Central Banks), but has been changed in order to reflect its wider application. The model and format are maintained under the auspices of the [SDMX](#) initiative. In this context, GESMES/TS is known as SDMX-EDI. In the same context it must be mentioned SDMX-ML which is the XML syntax used by the European Central Bank and the national central banks in the web dissemination of statistics.

At present, the use of the web and all instruments correlated to it and based on the XML standard have allowed the implementation of the XBRL described in the following section.

2.3.2 XBRL: eXtensible Business Reporting Language

While different EDI solutions may be very efficient in some cases, as shown in the previous section, there is also a strive for more generalised technical structures and formats that may aid statistical offices and other collectors of business information connect with businesses in an even more efficient way. This could involve sharing of data between authorities or the possibility for businesses to re-use the data in their administrative systems for many purposes. XBRL, short for eXtensible Business Reporting Language, may very well become this standard format. The XBRL format, developed and maintained by a consortium of regulators, accountants and software builders, can offer a link between the data kept in book keeping systems and the data terms of regulators, such as national statistical and tax offices. XBRL offers the same advantages as other EDI solutions; a possibility of reduced respondent burden and data collection costs, especially after over time since the first implementation

may invoke some costs. The main difference between XBRL and other more specific EDI solutions is that XBRL is an open format that is intended to be used for many different purposes from the business side; exchanging data within the enterprise (or enterprise group), exchanging data with accountants, sending data to government authorities and also sending data to any interested party such as banks, analysts et cetera.

What XBRL is has been described by Roos (Roos, 2008). XBRL is an XML-based computer language specifically developed for the exchange of business facts between computer systems. Business facts are defined as administrated events that are of economic interest to the company or other related organisations. The XBRL-standard provides a precise, predictable structure for describing and expressing those business facts in a way that can be used and processed by computer systems.

One advantage of XBRL compared to other file formats is that it is an open standard based on the globally well-known language XML. The idea of XBRL is rather simple. Instead of treating information as a block of text, like on a website or in a written document, XBRL tags the individual information in a document with the necessary information. This makes each piece of information readable and possible to interpret electronically.

In order for the systems to understand each other, an agreement on terms and definitions is needed. This is defined in a taxonomy. An XBRL taxonomy defines variables and the relations that may exist between those variables. A taxonomy may also refer to variables defined in other taxonomies. The taxonomy is developed by for example a data collector to describe which information is required. If this is done, a software provider or businesses themselves can link the data in the administrative systems to this taxonomy, and provide the requested information automatically as soon as the link is set up. Taxonomies can be created globally or locally, and relate to any kind of concept. On an international level, there are for example taxonomies created based on the accounting standards US GAAP and IFRS (for more information, see www.xbrl.org). It should be noted though that even though XBRL is an open format, the immaterial rights to the XBRL format are owned by XBRL International Inc. Therefore, when developing a taxonomy it is important to follow the specifications and guidelines given by the international consortium, and also to follow how others use the standard to ensure that the taxonomy created is in line with other ongoing initiatives. According to Bohlin et al. (2009), there are three fundamental design principles for the creation of taxonomies:

- Immaterial rights: The development and maintenance of taxonomies must follow the Intellectual Property Policy of XBRL International. They can be found at www.xbrl.org
- Technical guidelines: The taxonomy must follow XBRL 2.1 and should follow the Financial Reporting Taxonomy Architecture (FRTA) of XBRL International as much as possible. The FRTA can also be found at www.xbrl.org
- International “best practice”: The taxonomy must strive to follow similar design as other established taxonomies to ensure comparability and interoperability.

There are a number of factors influencing how effective introducing XBRL in the collection of statistical information can be:

- If XBRL is used for other means of sharing information or not. If other government agencies also use XBRL, it is easier to set up something. For example, if XBRL is used to send annual reports or tax reports, there is already an experience of using taxonomies and mapping business information

to them. Moreover, if some data can be used to fulfil data provision to several authorities, there might be a better business case to get software providers and others interested.

- How the taxonomy is set up. If the taxonomy is set up according to the requirements above that is a good start, but it must also be usable and understandable to businesses. The terms used must be clear and unambiguous.
- Mapping the taxonomy to the business systems. The possibility to map a taxonomy to business systems may vary a lot depending on the situation in different countries. Some countries have mandatory or very well-spread standardised accounting systems, meaning that it is possible to create more general mappings (for example, by software providers or even the statistical office itself) that can be used by many enterprises. For example, Statistics Finland has created a taxonomy that relates to hotels (Kontinen, 2012). In other countries, more or less every single enterprise would have to make their own mapping. For enterprises to be willing to make such a work, there must be potential gains over time, for example the possibility to re-use the mapping many times. For large enterprises, this might be certain since they are almost always included in the statistical samples, while for smaller enterprises they might very well rotate out of the sample after only a short while. Such enterprises might be less willing to make a mapping exercise.
- Mapping the taxonomy to the metadata systems at the Statistical office. Using definitions in a taxonomy in the data collection is an undertaking that must also be upheld. Changes to variables and definitions linked to XBRL taxonomies that are used by several parties cannot be done without informing the others, or indeed the businesses using these for reporting data. The links between the metadata systems and the taxonomy must be upheld and maintained, but it can also help in giving standardised definitions that are generally agreed upon, meaning the statistical office does not need to prepare its own definitions.
- Legislative issues. In a few countries, using XBRL has become mandatory for some reporting, in a few cases also statistical reporting (mostly for financial enterprises at the time of writing). Such a support of course makes using the XBRL solution for statistical collection much easier.
- Building technical solutions. After the mapping, there needs to be a technical solution to transfer the data from businesses to the statistical office. There are three possible cases to cover:
 - Complete coverage (everything requested in the statistical requirements is covered by the mapping and there is no need for adjustments).
 - Adjustments needed (for example, everything is covered but several sources at the business are involved and need to be combined, or some values have to be recalculated, e.g., adjusted from an accounting period to a calendar year).
 - Incomplete coverage (not everything in the statistical requirements is covered).

In the first case, it would be enough to build some mechanism for transferring the file from the business to the statistical office. It is probable that business would request an encrypted possibility. In the other two cases, there is also a need to build some tool to do the adjustments and add the missing data. This requires a more sophisticated technical solution. It can be built by an outside provider or by the statistical office itself, depending on what is deemed most suitable. Regarding this point, there are large similarities between systems for XBRL data and other EDI solutions.

Considering the limitations outlined above, it is clear that it is still too early to recommend a generalised common solution for implementing XBRL in the statistical collection all across the European Union. It is however clear that XBRL is a standard format which is used more and more extensively for many purposes, and statistical offices need to consider the possibilities to exploit in the statistical collection in the extent possible given the country specific situation. It can be foreseen that the use of XBRL will continue to grow, also in the statistical area. If a country is contemplating implementing an EDI solution (see above), making use of XBRL should also be considered. Using XBRL has a large potential in reducing response burden as well as data collection costs, but there is still a long way to go before it is a widespread possibility for businesses to provide data through XBRL on a broad scale.

2.4 *Use of administrative data*

Although administrative data do not represent themselves a data collection technique, they have to be mentioned in this topic since their use is going to change the way NSIs organise the data collection phase of the survey process.

This section will briefly describe the advantages and disadvantages in using administrative data and when and how they can be used in the statistical survey process. Full and detail information on this subject can be found in the module “Data Collection – Collection and Use of Secondary Data”.

Administrative data are “*the set of units and data derived from an administrative source*” and an administrative source is “*a data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations*” (SDMX, 2009).

The ESSNet on the “*Use of Administrative Data for Business Statistics*” ([Admin Data ESSNet 2011](https://essnet.admindata.eu)) <https://essnet.admindata.eu>, which is part of the European MEETS program, is aimed at developing recommended practices on the use of these type of data in business surveys. It also reports information on projects, carried out by NSIs, to improve or increase the use of administrative data.

The use of administrative data has the great advantage of reducing data collection costs as well as respondent burden and, sometimes, to improve the timeliness of data delivery because surveys can use already existing data (Statistics Canada, 2010). But, as these data are collected for administrative purposes and not for statistical ones, their use in surveys has to be done under certain bounds/limitations. In fact, administrative data are collected by public organisations to administer or to control or to tax or to regulate the activities of enterprises or individuals. This approach is different from that followed by NSIs that collect data to study and analyse individuals or enterprises. Besides, administrative data may differ from statistical data because public organisations and NSIs may adopt different definitions of units, different definition of variables and different classifications (Calzaroni, 2010).

Very important is the issue of the quality of administrative. The definition of quality is quite complex and therefore not yet commonly shared among NSIs (Casciano et al., 2011). Apart from definition, the problem about quality lays on the fact that NSIs do not control the data collection process which is set up by public organisations that use their own control procedures that can be based on different and less stringent criteria than those used by NSIs (Statistics Canada, 2010). Besides, the quality level could be lower for those variables which are not fundamental for the administrative study but are important for

statistical purposes. An overview of projects and approaches for assessing the quality of a secondary source can be found in the theme module “Data Collection – Collection and Use of Secondary Data”.

The existing practices in the use of administrative data, among the NSIs, for producing business statistics, are reported in the table below. The use of administrative data combined with survey data, is divided in the four domains studied by the Admin Essnet: Business Register, Short-term statistics (STS), Structural Business Statistics (SBS) and Prodcom statistics.

The use of administrative data – from Admin Data ESSNet						
<i>Countries of the EU & EFTA by combination of direct sources used for producing business statistics and business statistics domain (end of 2010) – Table 4</i>						
DOMAINS	COMBINATIONS				NON-RESPONSE	TOTAL
	Admin/ register data only	Admin/ register & survey data	Survey data only	Not specified		
BUSINESS REGISTER	12	16	-	-	2	30
Turnover	2	15	12	-	1	30
New orders	-	10	19	1	1	31
STS Production prices/costs	-	14	16	-	1	31
Building permits	14	2	13	1	1	31
Employment	3	16	10	-	1	30
Annexes I-IV	1	23	4	-	2	30
Annex V	13	11	3	-	3	30
Annex VI	16	8	2	1	3	30
SBS Annex VII	13	7	4	3	3	30
Annex VIII	-	13	14	-	3	30
Annex IX	20	7	1	-	2	30
PRODCOM	-	10	13	1	2	26

Elaboration from Admin Data ESSnet WP1, *Deliverable 1.2/2010. Database “Overview of Existing Practices in the Uses of Administrative Data for Producing Business Statistics in the EU and EFTA”* (2011).

Briefly, this table shows (Admin Data ESSNet 2011- Deliverable 1.1, pages 20-28), that: for the Business Register the majority of countries update it also by means of regular surveys; for Short-term and Structural statistics the exclusive use of administrative data is not common, but that administrative data do exist although they cannot be used as the unique source of information for reasons of quality,

comparability, timeliness, etc.; from Prodcum, due to the nature of its statistics, the use of administrative data instead of direct surveys is limited.

Anyway (Admin Data ESSNet 2011- Deliverable 1.1, page 13), the statistical use of administrative data is increasing because it is recognised and sustained by national statistical laws and because the cooperation among NSIs and public bodies is improving as well as the organisation for their collection and transmission. Obviously, the situation in Europe is not homogeneous with countries that represent the optimum and others quite far from it. Examples of the optimum are represented by France and Scandinavian countries: in the first case the NSI directly manages the business register, called SIRENE, which is used for both administrative and statistical purposes; in the second case there is a very good cooperation between the NSIs and the public organisations that hold the administrative data to set up and define strategies for collecting and using these data.

Examples of use of administrative data inside the statistical survey process are:

1. Direct processing or analysis: when administrative data can replace survey data.
2. Indirect processing: when ad hoc statistical surveys are run to cover lack of information of the register or to update it.
3. Indirect estimation: when administrative data are used as input in some estimation models.
4. Survey frames: when administrative data are used as survey frames or to update them.
5. Matching with statistical archives: this could be horizontal – different archives to obtain data for the same unit - and/or vertical – different archives to obtain information for different types of units. This also known as Hybrid Data Collection (HDC) that, hence, represents a collection process based on heterogeneous administrative archives, that can change also over time (Calzaroni, 2010).

The use of administrative data for statistical purposes should be done after a strict evaluation of many aspects like their quality, their coverage, the concepts and definitions they use. In their decision process, statisticians should consider and evaluate a set of factors whose composition depends on the type of source used. A not exhaustive list of the main factors is described in the following:

- **Response burden:** evaluate whether the use of administrative data really reduces the response burden (questionnaires are not administered or shorter versions of questionnaires can be used);
- **Cost:** evaluate whether the use of administrative data can eliminate some of the steps of the data collection process thus reducing cost;
- **Coverage:** evaluate whether the population the administrative source refers to is defined with the same criteria of the survey population;
- **Concepts and definitions:** evaluate whether the concepts, the definitions of units and variables as well as classifications are coherent and suitable for the survey needs;
- **Quality:** evaluate the control process used by the public administration and whether its criteria fit with those used by the NSI;
- **Timeliness:** evaluate whether the availability of administrative data fits with survey deadlines;

- **Consistency over time (stability):** evaluate whether data can change over time because of new administrative laws or rules or because of political changes;
- **Physical integration:** evaluate whether data are available in a convenient format in order to be easily matched with the statistical ones (if they are aggregated or not, which standardisation criteria have been used, etc.).
- **Legal issues:** be sure about the fact that their use is not limited by any privacy constraints.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Admin Data ESSNet (2011), Work Package 1 – Deliverable 1.1 “Main findings of the information Collection on the Use of Admin Data for Business Statistics in Eu and EFTA Countries” – June, 2011 (<https://essnet.admindata.eu>).

Balestrino, R., Macchia, S., and Murgia, M. (2006), Data capturing strategies used in Istat to improve quality. UNECE – Work session on statistical data editing, Bonn, 25-27 September 2006.

Blanke, K., Brancato, G., Hoffmeyer-Zlotnik, J. H. P., Koerner, T., Lima, P., Macchia, S., Murgia, M., Nimmergut, A., Paulino, R., Signore, M., and Simeoni, G. (2006), *Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System*. http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/Handbook_questionnaire_development_2006.pdf.

Bohlin, M., Holmgren, T., Persson, A., Rydell, L., and Thorling, P. (2009), *Guide till svenska taxonomier för årsredovisning och revisionsberättelse*.

Budano, G. (2008), *Design and implementation of the CAPI IT system for the Labour Force Survey* (only in Italian). Istat – Metodi e norme, n.36 2008.

Calzaroni, M. (2010), *The use of administrative sources for Statistical Registers*. Naples, October 2010.

Capparucci, L., Degortes, M., Landriscina, M., and Murgia, M. (2009), Comparative analysis among open source and commercial software for the development of electronic questionnaires for statistical surveys. NTTS 2009 – Bruxelles, February 2009.

- Casciano, C., De Giorgi, V., Luzi, O., Oropallo, F., Seri, G., and Siesto, G. (2011), Combining administrative and survey data: potential benefits and impact on editing and imputation for a structural business survey. UNECE - Work Session on Statistical Data Editing (Ljubljana, Slovenia, 9-11 May 2011).
- Chen, P. (1976), The Entity Relationship Model: Towards a Unified View of Data. *ACM Transaction on Database System* **1**, 9–36.
- Clayton, R. L., Searson, M. A., and Manning, C. D. (2000), *Electronic data collection in selected BLS establishment programs* (Clayton_R@BLS.gov).
- Couper, M. P. (2001), *Web Surveys: The Questionnaire Design Challenge*. Survey Research Center, University of Michigan (mcouper@umich.edu).
- Fanning, E. (2005), Formatting a paper-based survey questionnaire: best practices. In: *Practical Assessment, Research & Evaluation*, Volume 10 m.12, August 2005.
- Iverson, J. (2009), Metadata-driven Survey Design. *IASSIST Quarterly*, Spring – Summer 2009.
- Istat (1989), *Manual of surveys techniques*. Notes and reports, 1989, volume 3.
- Kontinen, J.-P. (2012), Rationalising data collection: automated data collection from enterprises. UNECE Seminar on New Frontiers for Statistical Data Collection, 31 Oct-2 Nov 2012.
- Murgia, M. and Simeoni, G. (2005), Improving the Quality of the Assisted Coding of Occupation in CATI Surveys through Control Charts. SIS Conference - Classification and Data Analysis - Cladag 2005 (Parma, June 6-8, 2005).
- OECD Glossary of Statistical Terms, <http://stats.oecd.org/glossary/>.
- O'Neil, G. E. (2008), *Developments in Electronic Survey Design for Establishment Surveys*. United States Census Bureau.
- Parent, G. and Jamieson, R. (1999), *The use of CAI for the collection of business surveys in Statistics Canada*.
- Roos, M. (2008), The Dutch Taxonomy Project and structural regulatory business reporting: impact for Statistics Netherlands. 94th DGINS Conference 25–26 September 2008, Vilnius, Lithuania.
- SDMX (2009), *Metadata Common Vocabulary*.
- Statistics Canada (1998), *Survey Methods and Practices*. Catalogue no. 12-587-X.
- Willeboordse, A. (ed.) (1998), *Handbook on the design and implementation of business surveys*.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Main Module
2. Questionnaire Design – Electronic Questionnaire Design
3. Questionnaire Design – Editing During Data Collection
4. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
5. Data Collection – Design of Data Collection Part 2: Contact Strategies
6. Data Collection – CATI Allocation
7. Data Collection – Collection and Use of Secondary Data
8. Coding – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

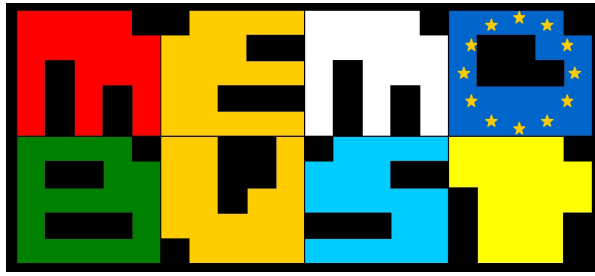
Data Collection-T-Techniques and Tools

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-02-2012	first draft	M. Murgia	ISTAT-Italy
0.2	08-05-2012	second draft	M. Murgia	ISTAT-Italy
0.3.1	05-09-2012	third version	M. Murgia	ISTAT-Italy
0.4	08-03-2013	glossary review	M. Murgia	ISTAT-Italy
0.5	19-11-2013	fifth version	M. Murgia	ISTAT-Italy
0.6	03-12-2013	EB revision	M. Murgia	ISTAT-Italy
0.6.1	05-12-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:50



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: The European Statistical System

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Tasks and responsibilities in the ESS.....	3
2.2 Governance and collaborations	4
2.3 Statistical production: status and future directions	6
3. Design issues	7
4. Available software tools.....	7
5. Decision tree of methods	8
6. Glossary.....	8
7. References	8
Interconnections with other modules.....	9
Administrative section.....	10

General section

1. Summary

An overview is given of the European Statistical System (ESS) as it operates currently. The way the ESS functions is likely to change in the coming decade, and some recent development directions, especially those related to the ESS.VIP programme are pointed out.

2. General description

2.1 Tasks and responsibilities in the ESS

According to the Eurostat website¹ *the European Statistical System is the partnership between the Community Statistical Authority, which is the Commission (Eurostat), and the national statistical institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European Statistics. This Partnership also includes the EEA and EFTA countries.* At the time of writing, the ESS consists of the 28 member states of the European Union plus Switzerland, Norway, Iceland and Liechtenstein.

The origins of the ESS can be traced back at least to 1952, when the High Authority of the European Coal and Steel Community established its Statistics Division. In 1959 the Statistical Bureau of the European Community (Eurostat) was established. The current legal basis for the ESS' governance system is laid down in a regulation of the European parliament and Council (EU, 2009). In short, this regulation establishes principles and a governance mechanisms for collaboration between partaking institutes. Below, some specific issues addressed by this regulation are highlighted. In the subsequent subsections we describe current collaboration modalities in the ESS and briefly discuss future directions.

Statistical Governance. Eurostat is responsible for development, production and dissemination of European statistics. It is also the sole authority on the area of statistical content, production and quality of European statistical publications. The European Statistical System Committee (ESSC) governs the development, production and dissemination of European Statistics; it consists of representatives from each NSI (usually director generals) and is chaired by Eurostat. In the following subsection the current governance and collaboration mechanisms will be described in a bit more detail. Eurostat is also responsible for preparing the multi-annual European Statistical Programme, which after consulting the ESSC, will be established by the European Parliament and Council. The multi-annual programme, which maximally covers a five-year period, prioritises certain developments. For example, the current programme states that the ESS shall produce updated indicators supporting the targets of Europe 2020 which include, amongst others, targets in the area of employment, energy and climate, and social integration. Besides the multi-annual programme, an annual work programme is presented by Eurostat to the ESSC. The ESSC receives advice on the multi-annual programme from the European Statistical Advisory Committee (ESAC), which is established in a separate resolution (EU, 2008b).

Statistical principles. European statistics should be produced in objective and impartial ways; they should be reliable in the sense that they are based on the scientific method; microdata must be treated

¹ http://epp.eurostat.europa.eu/portal/page/portal/pgp_ess/about_ess

confidentially and production must take place in a cost-effective manner. Minimisation of administrative burden is mentioned in this context as well. The principles are worked out further in the European Statistics Code of Practice (EU, 2011) which is maintained by the ESSC. The implementation of the Code of Practice is monitored by the European Statistical Governance Advisory Board (ESGAB) which is established through another resolution (EU, 2008a).

Statistical Quality. Statistics will be judged according to criteria relating to relevance, accuracy, timeliness, punctuality, availability and clarity, comparability, and coherence. Here, Eurostat has the authority to set the norms for these dimensions when regarding European Statistics. It is important to note that this authority pertains to the inner workings of Eurostat and not to those of the NSIs of member states or their products. Therefore, quality demands on statistical products delivered by member states to Eurostat are developed through various forms of collaboration within the ESS; the MEMOBUST project being one example that can be regarded in this context. We will return to this subject in Section 2.3.

European Collaboration. The regulation provides the option to set up (temporary) collaboration networks on specific statistical topics, provided that the results of such a collaboration are made available to the whole ESS. The regulation also provides the option for a “European approach to statistics” when this either improves quality on a European scale, improves cost-effectiveness or reduces administrative burden.

Collaboration with other bodies. Eurostat and the ESS shall seek collaboration with the other European Statistical Bodies, in particular the European System of Central Banks (ESCB), and especially where collaboration can reduce administrative burden. The regulation also explicitly provides for the option of exchanging (confidential) data between these bodies. An example of a collaboration between a statistical agency and a central bank is the development of the JDemetra+ software for seasonal adjustment. This software has been built in a collaboration between Eurostat and the Central bank of Belgium.

Confidentiality. The regulation provides extensive articles that guarantee confidential treatment of statistical (micro-)data. It establishes who may access confidential data and how they may be used (scientific, or for statistical purposes only). Also, members of the ESS are required to take measures to make violation of confidentiality punishable.

2.2 Governance and collaborations

Figure 1 gives an overview of the European Statistical System’s governance structure and collaboration mechanisms. The ESSC, consisting of Directors General of all participating member states functions as “daily management” and meets four times each year. ESSC meetings are prepared by the Partnership Group (PG). The PG consists of the DG and vice DG of Eurostat, an elected chairman from the member states, the chairman of the previous and the next period, and five other, chosen DGs of other member states. The DGINS meeting is an annual conference of all Directors General of the participating and candidate member states that is used to discuss organisational and thematic (statistical) current topics. For example, in 2012 the main themes were “green economy” and “geospatial statistics”, in 2013 the subjects are “new round of peer review” and “big data”.

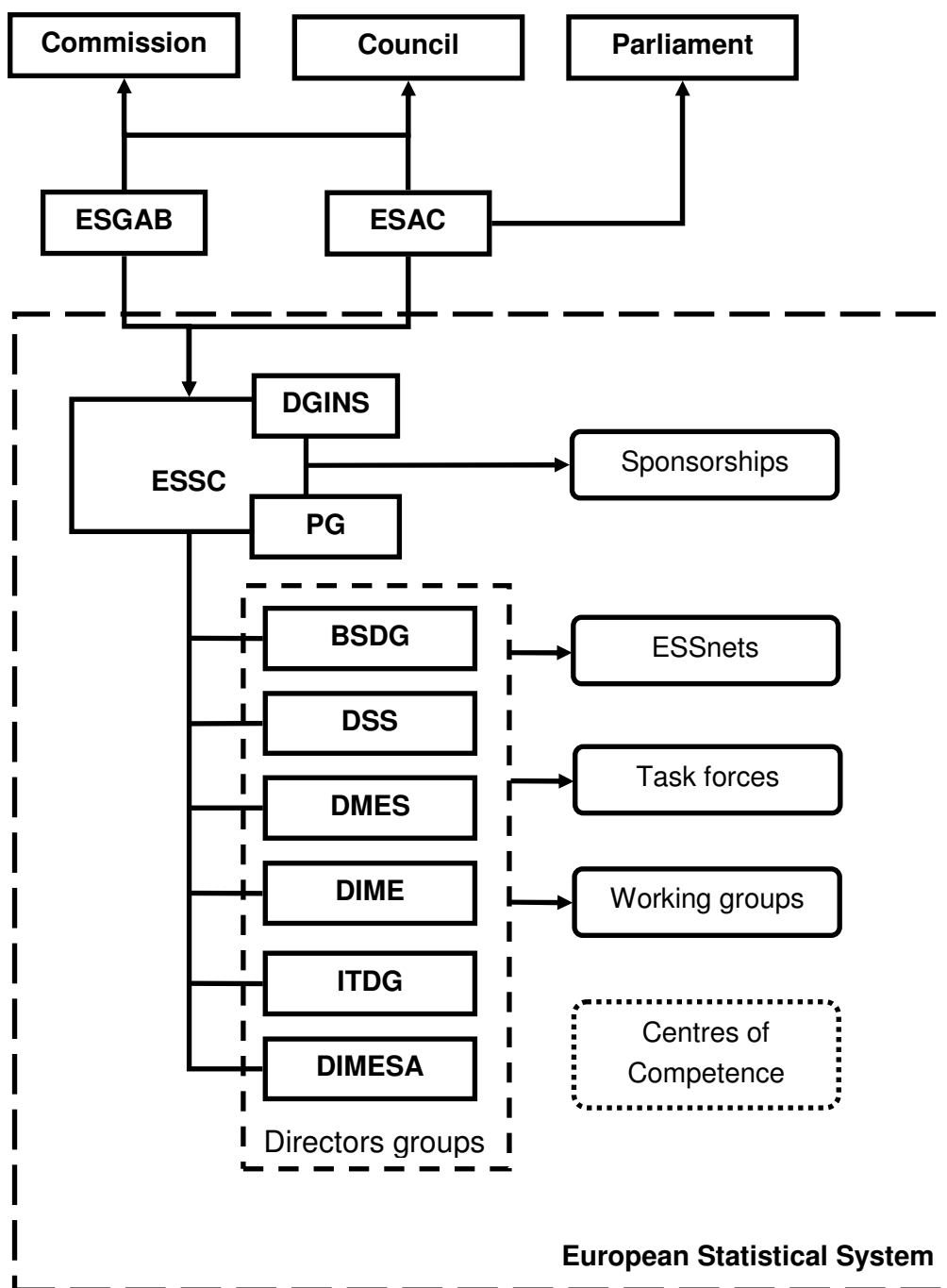


Figure 1. The European Statistical System and its relation to governing bodies of the European union.

As stated above, the ESSC is responsible for the multi-annual programme. The European Statistical Advisory Committee (ESAC) advises the ESSC on the content of this programme and reports directly to the European Commission, Council and Parliament. The European Statistical Governance Advisory Board (ESGAB) reports to the European Commission and the Council and guards of the implementation of the Statistical Code of Practice, for example, by organising peer reviews.

For specific strategic subjects or tasks, the ESSC can appoint a *Sponsorship*. A sponsorship consists of a delegation of several member states, usually at DG-level, which perform a specific task, for a fixed period of time. For example, in the period 2011-2013 the *Sponsorship on Standardisation* has investigated methods for improving standardisation processes within the ESS. This has eventually lead to an ESSnet on standardisation (2012-2014).

Under the ESSC there are six director's groups governing various subjects related to the production of statistics on a strategic level. These are the Business Statistics Directors Group (BSDG), Directors of Social Statistics (DSS), Directors of Macro-Economic Statistics (DMES), Directors of Methodology (DIME), IT Directors Group (ITDG) and Directors' Meetings of Environmental Statistics and Accounts (DIMESA). Some director's groups have a subgroup preparing the plenary meeting (not in Figure 1). For example, the DIME plenary meeting is prepared by DIME-Steering Group meetings, where the steering group has about ten members. The directors groups are mandated by, and report to the ESSC and govern ESS-activities where actual (statistical) development is done. There are several ways in which such activities can take place, some of the most common ones are stated below.

ESSnets are projects, subsidised by the Commission (Eurostat), performed by NSIs developing products to be used by the whole statistical system. This MEMOBUST handbook, for example, is the product of an ESSnet. ESSnets typically have duration of 2-4 years and can have a substantial amount of staff working on them from various member states.

Task Forces, consisting usually of a few experts from various NSIs, can be appointed to perform a specific task. One recent example is a Task Force which developed a Quality Assurance Framework for the ESS (Nov 2011-June 2012). Task forces need not be limited to members of the ESS and can involve experts from other organisations like the OECD or ESCB.

Working groups are collaborations between subject matter experts which are usually of a more permanent character. For example, the Working Group on harmonisation of consumer and price indices started in 1993 and still exists now. Working Groups need not be limited to the ESS but may involve experts from outside the ESS, such as with the SDMX technical working group which consists of members from NSIs, Central Banks, the OECD and the Word Bank. In many cases, working groups report to one or more of the Director's Meetings.

Centres of Competence. These do not exist right now, but there is currently a strong interest in developing them. Competence Centres will likely to serve a role providing (methodological, subject-matter) knowledge and expertise across NSIs and Eurostat within the ESS. Modalities for financing, governance, tasks and mode of operation are at the time of writing being discussed in various Directors Groups.

2.3 *Statistical production: status and future directions*

Currently, most European statistics are compiled by Eurostat based on aggregate figures delivered by members of the ESS and other partners (ESTAT, 2009). This means that the statistical production systems of Eurostat and other institutes are completely separated. Harmonisation of statistical concepts, methods and processes is established via both formal and informal, consensus-driven routes.

The formal side of harmonisation is established by fixing agreements in European regulations. For example, regulation N° 1893/2006 (EU, 2006a) establishes the NACE classification (Rev. 2) of economic activities. The NACE classification itself is referred to in the regulation that establishes the

short-term statistics (EC, 1998; EU, 2005) by specifying, in terms of NACE codes, on which type of activities member states should report. The same regulation also provides details on the type of statistical unit, level of detail, timeliness of data deliveries and so on. The interpretation of these regulations is aided by guidelines such as the *Methodology of short term business statistics* manual (EU, 2006b).

Regulations like the STS regulation and their corresponding guidelines do not establish explicit demands on methodology or the value of statistical quality measures such as confidence intervals. Such issues are mostly covered by in collaborative projects where the consensus on these matters is established. Recent examples include the production of a *handbook on precision requirements and variance estimation for ESS household surveys* (ESTAT, 2013) and the recommendations on the use of administrative data developed in the *AdminData* project (ESS, 2013). The documents produced by the MEMOBUST project should be regarded in this light as well. Such handbooks and documents provide recommendations by field experts, have been subjected to extensive peer review, and are ultimately presented to one or more of the ESS' directors groups for endorsement. In principle, such recommendations can ultimately be upgraded to actual standards, and the recent work of the *Sponsorship on Standardisation* and the ensuing *ESSnet on Standardisation* offer guidelines on when and how to standardise existing common practices and (quasi) standards.

To obtain the data from their suppliers, Eurostat currently uses a “single entry point policy”, meaning that all data deliveries should go through a single access point at Eurostat. The current implementation is provided by the eDAMIS (electronic Data files Administration and Management Information System) software, which data suppliers can use to upload their data to Eurostat.

In 2012, Eurostat unfolded ambitious plans to develop far-reaching integration of statistical production systems in the ESS under the title “ESS.VIP programme”, where VIP stands for Vision Implementation Projects. Here, *Vision* refers to a communication of Eurostat to the European Committee entitled “on the production method of EU statistics: a vision for the next decade” (ESTAT, 2009). In short, in the vision it is argued that the ESS should move away from stove-pipe oriented and separated production systems and replace it with integrated systems where production (software) tools and (micro-)data can be shared and reused securely. Needless to say, such a transition would have a tremendous impact on the way official statistics are produced in the ESS requiring not only changes in current business architecture but possibly policy changes at the political level to allow for the sharing of microdata, for example.

At the time of writing, the scope, governance, financing, and mode of operation of both the ESS.VIP programme and its final products are still being debated by the directors groups and at ESSC level. However, regardless of the outcome, it does seem likely that steps towards further integration will be taken in the future.

3. Design issues

Not applicable

4. Available software tools

Not applicable

5. Decision tree of methods

Not applicable

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

EC (1998), Council Regulation (EC) No 1165/98 concerning short-term statistics. *Official Journal of the European Communities* **162**, 1–15.

ESS (2013), All publicly available information and products of the AdminData project are distributed via <http://essnet.admindata.eu>.

ESSC (2011), *European Statistics Code of Practice*. ISBN 978-92-79-21679-4.

ESTAT (2009), Communication from the Commission to the European Parliament and the Council on the production method of EU statistics. A vision for the next decade. *COM(2009)* **404**.

ESTAT (2013), *Handbook on precision requirements and variance estimation for ESS households surveys*. Eurostat Methodologies and working papers, ISBN 978-92-79-31197-0.

EU (2005), Regulation (EC) No 1158/2005 of the European Parliament and of the Council. *Official Journal of the European Union* **191**, 1–15.

EU (2006a), Regulation (EC) No 1893/2006 of the European Parliament and of the Council. *Official Journal of the European Union* **393**, 1–39.

EU (2006b), *Methodology of short term business statistics, interpretation and guidelines*. Office for official publications of the European Communities, ISBN 92-79-01295-9.

EU (2008a), Decision No 234/2008/EC of the European Parliament and of the Council. *Official Journal of the European Union* **73**, 13–16.

EU (2008b), Decision No 235/2008/EC of the European Parliament and of the Council. *Official Journal of the European Union* **73**, 17–19.

EU (2009), Regulation (EU) No 223/2009 of the European Parliament and of the Council. *Official Journal of the European Union* **87**, 164–173.

Interconnections with other modules

8. Related themes described in other modules

1. Not applicable

9. Methods explicitly referred to in this module

1. Not applicable

10. Mathematical techniques explicitly referred to in this module

1. Not applicable

11. GSBPM phases explicitly referred to in this module

1. Not applicable

12. Tools explicitly referred to in this module

1. Not applicable

13. Process steps explicitly referred to in this module

1. Not applicable

Administrative section

14. Module code

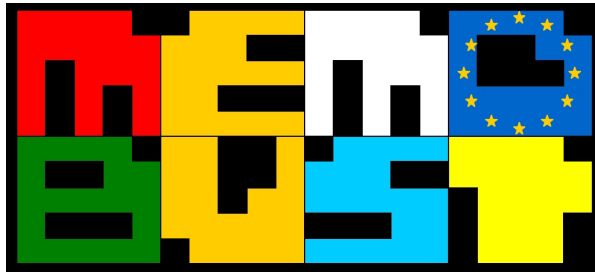
General Observations-T-The European Statistical System

15. Version history

Version	Date	Description of changes	Author	Institute
0.0.5	20-03-2013	first version	Mark van der Loo	Statistics Netherlands
0.0.6	21-08-2013	update after review	Mark van der Loo	Statistics Netherlands
0.0.7	26-08-2013	references updated to meet requirements	Mark van der Loo	Statistics Netherlands
0.0.8	02-09-2013	updates on references	Mark van der Loo	Statistics Netherlands
0.0.9	16-09-2013	removed double blanks; reference updates	Mark van der Loo	Statistics Netherlands
0.1	30-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:22



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: CATI Allocation

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction	3
2.2 On the scheduling and allocation problems in CATI allocation.....	4
3. Design issues	5
4. Available software tools.....	6
5. Decision tree of methods	6
6. Glossary.....	6
7. References	6
Interconnections with other modules.....	7
Administrative section.....	8

General section

1. Summary

To be able to perform CATI interviews sample elements have to be linked to CATI interviewers at some point, so that the interviewers can call and interview them by telephone. In order to be able to do this, several steps have to be taken. An important one is that interviewers have to be scheduled in a timetable, taking various conditions into account, such as days and/or parts of the day where interviewers are not able to work, the surveys for which the interviewers are trained and for which they can be deployed, who has to be called (in our case: businesses) and the time of day when this preferably should be done in order to increase the response rate. If the interviews are conducted from a call room, say at the premises of a national statistical office, its maximum capacity has to be taken into account (if the number of interviewers that can be employed simultaneously exceeds this capacity). The interviewers present in the call room (at a certain day, part of day (DPoD) combination) have to be fed with telephone numbers of sample elements to be contacted, as they were drawn into respective samples, so the interviewers can call them for interviews. The most difficult problem is the allocation of interviewers to DPoDs and surveys. This can be done by hand, but (preferably) by using optimisation models. It is possible to take deadlines for surveys into account in these models, or (salary) costs of the interviewer corps that should be adequate for its task, but not bigger than necessary. When these allocations have been made (and the interviewers know when they should come to work and for what surveys), they have to be ‘fed’ telephone numbers. We assume that the interviewers receive such telephone numbers ‘on their demand’, whenever they have indicated to be ready for a new interview.

A special issue is about the updating of an allocation that has been determined for a planning period. This is necessary as things change. Time flows and plan and reality tend to diverge more and more from each other as time passes. Interviewers get ill, go on vacation, have to take leave, quit their job, new ones get hired and trained. Also as time passes, new surveys come into sight that also have to be carried out. All these changes require a schedule to be updated, very regularly.

It should be noted that instead of a single method, in reality there is actually a complex of models, that are rather strongly related and that differ by using certain alternative constraints (or discarding such constraints), or certain objective functions that are different (but with the same set of constraints).

2. General description

2.1 Introduction

The subject of this theme is about allocation problems in CATI surveys. Ultimately it is about allocating sample elements to interviewers. This allocation happens in a few steps. As we assume that the interviewers work from a call room (say at the premises of an NSI) it has to be made clear which interviewer works when (on what DPoD combinations), and on what surveys (among those that are ‘active’ during the planning period). We assume that the allocation of sample elements to interviewers (and surveys) for CATI surveys is in the following order of steps:

1. **Scheduling:** The interviewers are allocated to admissible Day-Part of Day (DPoD) combinations, as well as to surveys they are supposed to work on.

2. **Allocation of workplace:** The interviewers are allocated to a specific workplace in the call room before they start their work on a particular DPoD combination.
3. **Allocation of telephone numbers:** Telephone numbers are allocated to interviewers when they indicate to be ready for a next interview. These telephone numbers are the direct links between interviewers and sample elements.

The first step is difficult and leads to all kinds of optimisation problems, depending on the goals that one pursues.

The second step is generally taken by some supervisor who knows the interviewers well and who knows which interviewers should sit together and which not. This step is not suitable for formalisation and automation, as it is likely to have little added value. For that reason it is not described here.

The third step is often carried out with specialised software, namely CATI call management software. Such a system assigns a suitable telephone number to an interviewer who asks for a number. The program then picks a suitable telephone number from a list. It ‘knows’ to which sample element it belongs, and in particular it ‘knows’ for which particular survey the sample element is intended. It also ‘knows’ for which surveys the interviewer asking for a new telephone number is qualified for. There are several variants possible for this step.

After the third step the interviewer can call the telephone number and try to interview the sample element to which this number belongs. This interviewing is not part of the allocation step considered in the present module. See, however, other modules in the topic “Data Collection”.

The steps above describe a static situation with a fixed planning period. In practice the planning is not static but dynamic. For all sorts of reasons reality may diverge from a planning: interviewers get sick, or move to another job, etc. As time proceeds, new days become available where new interviews can be planned. So in order to keep the planning up to date, new information has to be ‘injected’ into an existing (or current) planning. So this corresponds to a fourth step:

4. **Update the schedule:** Recent update information is used to update the current schedule.

We consider this step as a separate one, but it could be well considered an integral part of the first step. It is likely to be carried out with the same software.

2.2 *On the scheduling and allocation problems in CATI allocation*

The scheduling problem is a matching problem, where interviewers are linked to DPoD combinations and surveys. The matching uses various constraints, which reflect wishes and demands from the NSI that conducts the surveys, on one hand, and from the interviewers, on the other hand. Relevant input data for the scheduling problem are the following:

- It should be known which surveys are ‘active’ in the planning period considered.
- For each interviewer it should be known for which ‘active’ surveys he/she is qualified. More than one is possible. This means that such a person is allowed to conduct interviews for such surveys. This qualification is a result of successfully completing a relevant training.
- It should be known for each interviewer on which DPoD combination he/she is available.

- The interviewers are supposed to work from a call room, sitting at a desk (a workspace). This call room may physically consist of several rooms. Important in this case is that the capacity (the number of interviewers it can take simultaneously) is limited. This should be taken into account when planning the work. Another option would be that the interviewers work from home. In this case there is no limitation to the number of interviewers that can work simultaneously.
- It is possible to take into account specific wishes and demands. For instance the deadline of a survey and the maximum work load for interviewers. In case of a deadline, one should be able to estimate the number of hours needed to finish the work on time, that is, before the deadline.

There are also various objectives that can be considered, each resulting in specific objective functions. The scheduling problem is translated into an optimisation problem, which tries to maximise an objective function under a set of constraints. The problems that may arise in this way might be difficult to solve (currently), in which case one should look for simplified models that are tractable, and that yield approximate solutions that are good enough to be useful in practice.

The method allows for telephone numbers that have been called but without getting a response and can be called again. How often a number is called is part of the contact strategy. This is a separate topic, outside the present theme, but obviously related to it. It is obvious that in case call-backs are allowed, it is necessary to estimate how many there will be and how much time they take to handle. Otherwise it is impossible to say anything about deadlines.

Once the interviewers are at work in the call room they need to be provided with telephone numbers of sample subjects that they should call and interview. It is assumed that each interviewer gets a telephone number each time it is requested. A request indicates that an interviewer is ready for a next interview. There are various possibilities how telephone numbers can be allocated to interviewers.

As time proceeds, reality may deviate from the schedule calculate. As explained above this requires an update step, which is a separate step in the CATI allocation process.

Solving the CATI allocation problem can be done by formulating an appropriate optimisation model. See, e.g., Willenborg (2012) for details.

3. Design issues

We assume that it has been decided that a survey is needed to collect certain data from a target population. The details about this are not of interest to our allocation problem. What matters is that the elements in the sample have to be contacted and that this contact should result in information useful for the survey.

There are several options ('modes') available to contact sample elements: PAPI, CAPI, CATI, CAWI and mixed forms of these modes. What option to choose is a matter of weighing several factors: appropriateness, amount of work for the interviewee to provide the data using a particular interview mode, amount of work for the statistical office to prepare and execute a survey using a particular mode of data collection, effort needed to collect and prepare the data, quality of the data collected. (For more information on this see the theme module "Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method".) No mode is perfect. However, there is a trend to move away from PAPI interviews to modes that yield electronic data right away, that is CAPI, CATI

and CAWI. The first two of these employ interviewers, which has several advantages. CAWI is self-administered: it is (by far) the cheapest variant of the three mentioned, but it lacks the help and guidance of interviewers. CATI is much cheaper than CAPI, as no travel is needed from interviewers. But it poses limitations to the questionnaires being used concerning the kind, amount and complexity of information that can be gathered.

So depending on the kind of information that one wants to collect, and the amount of money available, a decision has to be made on what mode to use. It is even possible – and quite modern – to consider surveys that use a combination of these basic modes (mixed mode). More information can be found in the theme module “Data Collection – Mixed Mode Data Collection”.

4. Available software tools

Blaise® is a flexible and powerful computer assisted interviewing (CAI) system and survey processing tool for the Windows® operating system. Part of Blaise® is a CATI Call Management System. See <http://www.blaise.com/Description> for more information.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

<http://www.blaise.com/Description> (information on the Blaise system)

Willenborg, L. (2012), Allocation of sample units to interviewers in CATI surveys. Report, Statistics Netherlands, The Hague.

Interconnections with other modules

8. Related themes described in other modules

1. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
2. Data Collection – Mixed Mode Data Collection

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

1. Scheduling / planning – calculation of a plan / roster for interviewers so they know when to work and on which surveys
2. Matching

11. GSBPM phases explicitly referred to in this module

1. Data collection

12. Tools explicitly referred to in this module

1. Blaise (CATI call management system)

13. Process steps explicitly referred to in this module

1. Scheduling of CATI interviewers
2. Updating of CATI schedules
3. Allocation of telephone numbers to CATI interviewers

Administrative section

14. Module code

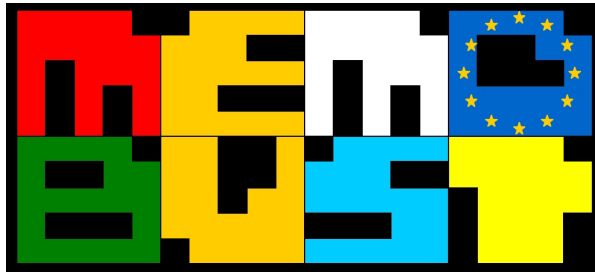
Data Collection-T-CATI Allocation

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	16-02-2012	first version	Leon Willenborg	CBS (Netherlands)
0.2	03-06-2012	remarks reviewers applied; references to other modules have been removed; technical terms have been removed; readability of text has been improved	Leon Willenborg	CBS (Netherlands)
0.3	12-12-2012	further remarks of reviewers applied; small corrections	Leon Willenborg Manuela Murgia	CBS (Netherlands) ISTAT (Italy)
0.4	19-11-2013	comments of EB review processed	Leon Willenborg	CBS (Netherlands)
0.4.1	19-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:50



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: GSBPM: Generic Statistical Business Process Model

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Versions of the GSBPM	3
2.2 The process structure of the GSBPM	3
2.3 Different uses of the GSBPM.....	5
2.4 The contents of the handbook.....	6
2.5 Remarks.....	8
3. Design issues	8
4. Available software tools.....	8
5. Decision tree of methods	8
6. Glossary.....	8
7. References	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

The Generic Statistical Business Process Model (GSBPM) is a means to describe statistics production in a general and process-oriented way. It is used both within and between statistical offices as a common basis for work with statistics production in different ways, such as quality, efficiency, standardisation, and process-orientation. It is used for all types of surveys, and “business” is not related to “business statistics” but refers to the statistical office, simply expressed. The GSBPM has been used as a basis for this handbook, in line with the instruction from Eurostat on using standards. The model is described below in brief.

The handbook is structured with some twenty topics. These topics are related to the GSBPM, but they do not correspond precisely with the phases and sub-processes. There is an inherent difference in that this handbook is restricted to methodology. A mapping between the GSBPM sub-processes and the handbook topics is provided, and some further comments are made.

2. General description

2.1 *Versions of the GSBPM*

Version 4 of the Generic Statistical Business Process Model (GSBPM), which was released in 2009, has been used by the Memobust project. In December 2013 version 5.0 was released, too late to use in the current handbook. The main differences are the following, as stated by UNECE (2013b):

- Phase 8 (Archive) has been removed, and incorporated into the over-arching process of data and metadata management, to reflect the view that archiving can happen at any stage in the statistical production process.
- A new sub-process: “Build or enhance dissemination components” has been added within the “Build” phase to reflect the growing importance of having a range of dissemination options.
- Several sub-processes have been re-named to improve clarity.
- The descriptions of the sub-processes have been updated and expanded where necessary. The terminology used has been changed to be less survey-centric, in recognition of the growing use of non-survey sources (administrative data, big data etc.).

Hence, there are no principal differences, which would cause negative effects. Please note that in this module the term survey is not restricted to direct data collection as above; it is used in a broad sense.

2.2 *The process structure of the GSBPM*

The Generic Statistical Business Process Model (GSBPM) is used by statistical offices. Statistics New Zealand was probably first and presented such a model around 2006. Much work has been done in joint UNECE/Eurostat/OECD Work Sessions on Statistical Metadata (METIS), see for instance UNECE (2009) and documentation from METIS. The GSBPM currently comprises four levels:

- Level 0, the statistical business process;
- Level 1, the nine phases of the statistical business process;
- Level 2, the sub-processes within each phase;
- Level 3, a description of those sub-processes.

The nine phases on the first level are the following:

1. Specify Needs;
2. Design;
3. Build;
4. Collect;
5. Process;
6. Analyse;
7. Disseminate;
8. Archive;
9. Evaluate.

The phases 1-3 can be regarded as preparatory, phases 4–7 correspond to the “obvious” production, phase 8 is a saving for the future (of essential data and metadata), and phase 9 summarises and formulates an action plan. Figure 1 below is the same as that presented by UNECE (2009), i.e., sub-processes on a two-digit level are included. Statistics Sweden, for instance, has found it useful to have further levels of sub-processes in some cases to elaborate more and give more detail and support. There is then a hierarchy with numbering on successive, more detailed, levels with 3 digits etc. This is the case for example for phase 4 *Collect* where the sub-processes are specified in more detail considering different collection modes, reminders etc. (There is unfortunately no reference in English.)

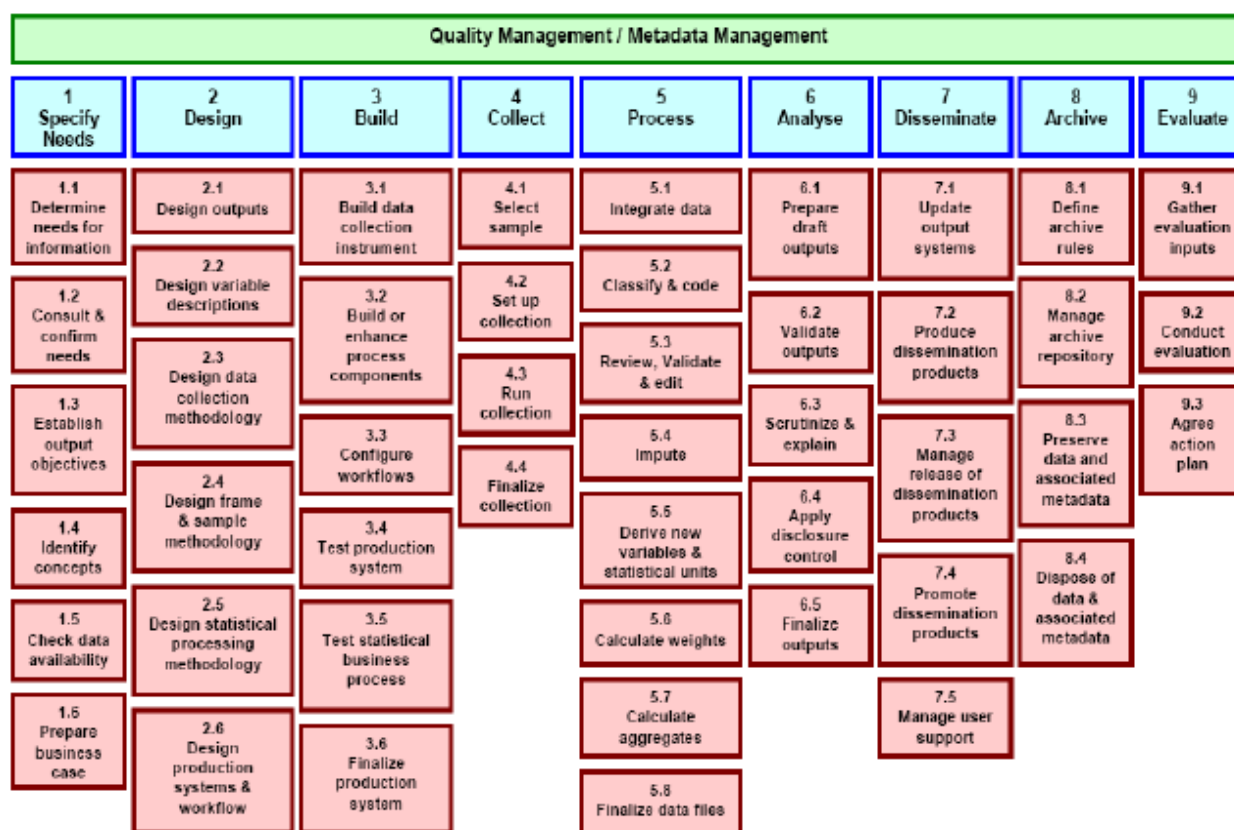


Figure 1. The GSBPM according to UNECE (2009)

All models are simplifications. Statistics production is not a simple process with successive sub-processes. There are, for instance, feedback loops, which are not explicit in the model. Many sub-processes need much more detail than shown here.

UNECE (2009) states that according to general process modelling theory, each sub-process should have a number of clearly identified attributes. It is clearly important to be aware of input(s), output(s), and purpose or value added. Furthermore, it is stated that other characteristics to take into account are the process owner with responsibility for the process, guides (for example manuals and documentation), and enablers: both people and systems. These attributes should be stated explicitly in order to make the statistics production flow smoothly. It is also stated that the attributes are likely to differ, at least to some extent, between statistical business processes, and between organisations.

2.3 *Different uses of the GSBPM*

The GSBPM model is useful as a common framework for all types of statistical surveys or processes, for instance both with direct data collection and when administrative data are used. Different types of statistical processes are described, for example by Eurostat (2009) in the European context and also in the theme module “General Observations – Different Types of Surveys”.

The six European types of statistical process described by Eurostat (2009) include “Statistical compilation”, which assembles a variety of primary sources to obtain an aggregate, with a special conceptual significance. Many of these are economic aggregates such as the National Accounts and the Balance of Payments. The GSBPM is then useful at least on a high level with data collection referring to the primary sources.

The GSBPM has been used also to describe production of registers, see UNECE (2011). There is some dissimilarity between maintenance of statistical business registers and production of statistics. The former is more or less constantly updated from a set of administrative and statistical sources. The latter is more from-start-to-end for each production round of a survey. The outputs from statistical registers are (i) registers and frames, and (ii) statistics based on the register. See the topic “Statistical Registers and Frames” (several modules) and the module “Dynamics of the Business Population – Business Demography”, for such outputs. The module mentioned gives examples of statistics based on the business register, for instance different types of birth depending on enterprise characteristics.

The GSBPM is useful for a new survey, when a survey is re-designed, and for continuous improvements of a repeated survey. The balance between the phases varies, for instance between situations such as those just mentioned. The three first phases may require considerable efforts for a new design and a re-design. They may be brief in repeated surveys, but they must not be skipped, when there is information to use, typically from the evaluation phase. The preparatory phases may then include, for instance, renewed contacts with users about priorities, modifications of some variables and possibly a data source, a renewed sample allocation, changes in the allocation of survey resources, and improvements of the IT-system.

Each sub-process involves methods, tools, and routines. The statistical office may choose to standardise or limit the ways in which a process can be run, at least for the majority of its statistical surveys. There could, for instance, be standard tools for sampling, for electronic data collection, for coding, and for imputation – tools that are broad enough to be useful for many surveys.

Similarly, the GSBPM is used also on the international level.

2.4 The contents of the handbook

It is necessary to note that this handbook is focused on methodology. This is especially obvious in method modules but also in theme modules. Hence, some sub-processes of the GSBPM get attention in several modules, whereas others are hardly mentioned. It has to be remembered also that this is an early version of the handbook. Some topics will expand later on, when there has been more development and agreement on recommended practices, and also more time for writing. Figure 2 below is a modified illustration of the GSBPM that is tied to the current Memobust project and handbook. The figure has two main goals.

Firstly, this version indicates the degree of methodological content in each sub-process. A scale with three categories is used: green for high methodological content, light green for an intermediate methodological content, and light yellow for little or no methodological content. The judgements are rough, and they have been made in the project based on UNECE (2009).

Secondly, there is a mapping between this GSBPM-version and the topics of the handbook. Again, the goal is to give an overview without details. Most topics are sorted into one or sometimes two sub-processes. This shows the main content of the topic. Hence, Figure 2 gives an overview in both directions between handbook topics and GSBPM sub-processes.

There are many handbook modules for phase 5 Process and phase 6 Analyse. Several of these include design aspects, in the module itself or in a separate module (an overall theme or a specific design module). Sub-process 2.5 Design statistical processing methodology is treated in several handbook modules. In Figure 2 they are shown in phases 5 and 6 but not in sub-process 2.5.

Some topics are quite broad and do not fit well into just one or a few sub-processes. Three such topics are:

- “General observations”
- “Overall Design”, which takes most of phase 2 *Design* into account with choices, allocations, and also coordination and optimisation aspects. This topic could, to some extent, be put in sub-process 2.6 *Design production systems and workflow* to emphasise its focus on overall issues.
- “Repeated Surveys”, which emphasises both user aspects and producer possibilities; these are typical when a survey is made regularly.

A few further topics are:

- “Response”, with modules about the response process and response burden, provides information to several sub-processes, for example about variables, data collection, and sampling with regard to response burden.
- “Quality Aspects”; this topic is shown in sub-process 6.3 *Scrutinize and explain*. It is related also to sub-process 6.5 *Finalize outputs* and it has more information, for instance about variance estimation and quality components in general.

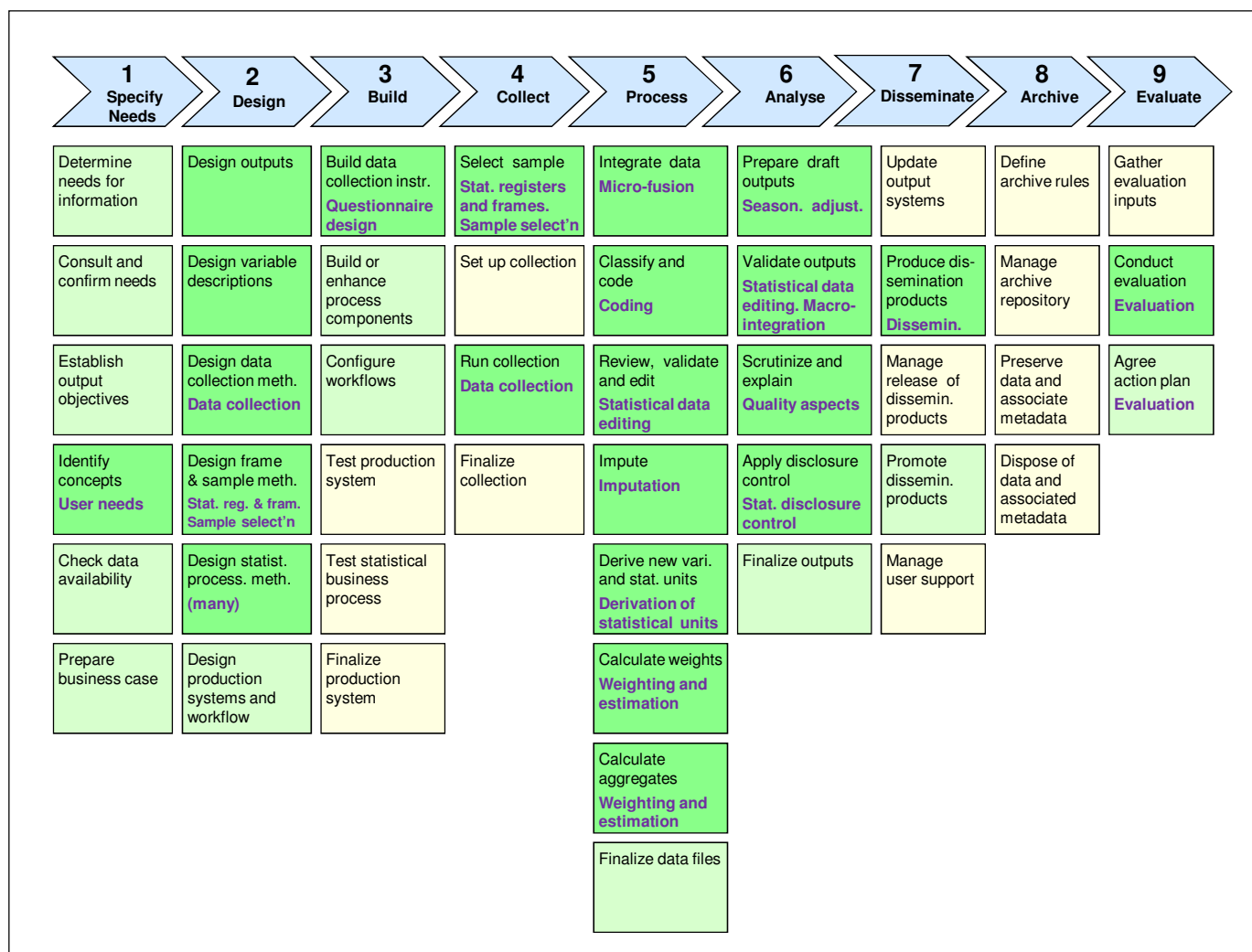


Figure 2. The GSBPM modified to describe methodological content and the handbook topics.

The topic “Statistical Registers and Frames” can be regarded in two different ways when it comes to the GSBPM. One of them was described above in section 2.1, i.e., the on-going building and maintaining of the statistical register and frames; this is a statistical process in its own right. The second way to consider this topic is in the context of a production round of a survey: then the topic provides information for sub-processes 2.4 *Design frame and sample methodology* and the sub-process 4.1 *Select sample*, which includes the establishment of the frame. The topic “Dynamics of the Business Population has contents that are related to the topic “Statistical Registers and Frames”.

Comments on some sub-processes that have no handbook module in Figure 2 follow. Sub-process 2.1 *Design outputs* is related to phase 1 *Specify Needs* (especially sub-process 1.4 *Identify concepts*) and also phases 6 *Analyse* and 7 *Disseminate*. Sub-process 2.2 *Design variable descriptions* have similar relationships and even more to sub-process 2.3 *Design data collection methodology* and sub-process 3.1 *Build data collection instrument*.

2.5 Remarks

The GSBPM is a practical reference when developing a standard set of methods, tools, and routines for the statistical office. Statistics Netherlands and Statistics Sweden, which both developed their versions before the joint international one, have long experience of the value of such a model as a fundamental part of statistics production, improvements, and developments. Statistics Sweden has an internal support system based on its GSBPM.

Communication is further facilitated and standardised through the more recent GSIM, Generic Statistical Information Model. This model is too recent to have been used in this handbook. GSIM is intended as a complement to the GSBPM. Information is easily found on the Internet, for example in the description by UNECE (2013a) about different activities for modernisation of statistical production and services; there is also a link to GSIM.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Eurostat (2009), *ESS Handbook for Quality Reports (EHQR)*. This handbook (planned to be revised soon) is accessible on the webpage of Eurostat, currently:

http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/EHQR_FINAL.pdf

UNECE (2009), Generic Statistical Business Process Model. Version 4.0 – April 2009 (prepared by the UNECE Secretariat). Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS).

UNECE (2011), Applying the Generic Statistical Business Process Model to business register maintenance. UNECE Conference of European Statisticians, Group of experts on Business Registers, Twelfth session, Paris, 14-15 September 2011.

UNECE (2013a), What's New from the High-Level Group? Working paper from the UNECE (prepared by S. Vale) to the Meeting on the Management of Statistical Information Systems (MSIS 2013).

UNECE (2013b), Generic Statistical Business Process Model. Version 5.0 – December 2013. The United Nations Economic Commission for Europe (UNECE). See:
<http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0>

Interconnections with other modules

8. Related themes described in other modules

(Restricted to modules on a high level)

1. General Observations – Different Types of Surveys
2. Overall Design – Overall Design
3. Repeated Surveys – Repeated Surveys
4. Statistical Registers and Frames – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. All nine GSBPM phases

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. The process steps on level two are shown in Figure 1 (and Figure 2 without number).

Administrative section

14. Module code

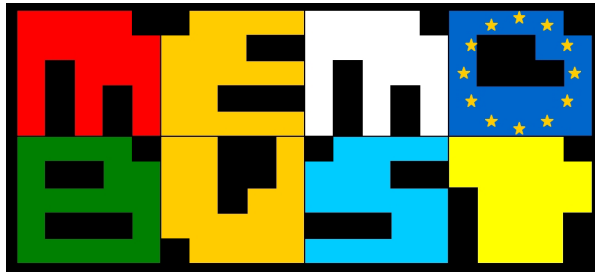
General Observations-T-GSBPM

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-02-2013	first version	Eva Elvers	Statistics Sweden
0.2	17-04-2013	extended, mapping	Eva Elvers	Statistics Sweden
0.2.5	16-11-2013	updates, figure	Eva Elvers	Statistics Sweden
0.2.6	26-11-2013	most of the EB comments	Eva Elvers	Statistics Sweden
0.2.7	26-11-2013	preliminary release		
0.2.8	20-01-2014	note: new GSBPM version	Eva Elvers	Statistics Sweden
0.3	10-02-2014	updates	Eva Elvers	Statistics Sweden
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:23



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Collection and Use of Secondary Data

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Research strategies with secondary data	4
2.2 On terminology and types of secondary sources	4
2.3 Consequences of using secondary data	7
2.4 Types of use of secondary data by NSIs.....	7
2.5 Some practical issues concerning collection of secondary data at NSIs	8
3. Design issues	8
3.1 Existence	9
3.2 Access.....	9
3.3 Usability (Fitness for use)	9
3.4 Coping with interruptions.....	12
3.5 Contact management	14
4. Available software tools.....	15
5. Decision tree of methods	15
6. Glossary.....	15
7. References	15
Interconnections with other modules.....	18
Administrative section.....	19

General section

1. Summary

National Statistical Institutes (NSIs) aim to produce undisputed and up-to-date statistics about their society. This requires up-to-date and reliable data. These could be data that the organisation itself collects (primary data) or data that are available in the outside world (secondary data). The latter can, for instance, be administrative sources maintained by other governmental organisations, and sources nowadays identified as ‘Big data’, such as data available on the internet and data generated by sensors. Mindful of the costs and response burden involved in the collection of primary data, more and more NSIs aim to maximise the use of secondary data for statistics production. The entire process of collecting already existing data is generally referred to as the collection of secondary data. This chapter discusses the advantages and disadvantages of this approach from an official statistics point of view.

In order to be in a position to use data from secondary sources, NSIs need to know which secondary sources exist with respect to their country and if they are allowed access them on a regular basis. Next, the ‘fitness for use’ of the data source for official statistics needs to be determined. There are many ways to determine this. The most important approaches focus on the metadata quality of the source, on the data quality of the input data, and on the data quality of the statistics produced. When a secondary data source is found suited for use, delivery agreements with the data provider need to be set up. It is considered good practice to assign an NSI-employee as the contact person for the source and the data provider. For important statistics that are dependent on the availability of the secondary data, ways to deal with any interruption or delay in the delivery need to be set up. These so-called fall-back scenarios may range from very simple actions, such as directly contacting the data provider, to the use of complex models that are able to cope with any data missing.

Apart from administrative data, some more recent work also focuses on the use of innovative secondary sources, so-called Big data, for statistics. Since a lot of these projects are still going on and these sources are not used for statistics yet, the focus of this chapter is limited to what is already known on the use of secondary sources for statistics.

2. General description

National Statistical Institutes (NSIs) that want to produce undisputed and up-to-date statistics need recent and reliable data. These could be data that the organisation itself collects, primary data, or data that is available in the outside world, so-called secondary data (Hox and Boeijs, 2005). Secondary data may be data gathered and maintained by other organisations for administrative purposes (Statistics Denmark, 1995; Wallgren and Wallgren, 2007), or data that is generated by an increasing number of electronic devices surrounding us and on the internet; so-called ‘Big Data’ (UN Global Pulse, 2012). The latter sources constitute a new and rapidly developing area of the use of secondary data for statistics (Glasson et al., 2013). However, since these sources are not used for statistics yet, the main focus of this chapter is on the – more established – use of administrative sources for statistics.

The remainder of chapter 2 is organised as follows. First, in section 2.1 we will discuss research strategies with secondary data. In section 2.2 we give a classification of secondary data types,

followed in 2.3 by an overview on the different types of use of secondary data by NSIs. We close this chapter by summing up the dependencies of secondary data use.

2.1 *Research strategies with secondary data*

The technique of acquiring and using secondary data sources is not unique to the field of official statistics. It evidently has multidisciplinary appeal, with extremely diverse academic fields drawing on the information included in secondary sources. All methods used belong to the academic discipline known as secondary research (Golden, 1976; Stewart and Kamins, 1993), which involves using existing data for a purpose different from the one for which they were originally collected.

In general, three different secondary research strategies can be discerned ('t Hart et al., 2005; Golden, 1976): content analysis, secondary analysis, and systematic review. The focus in content analysis is on extracting or summarising the content of various forms of human communication. Frequently used sources include newspapers, books, TV images, websites and paintings. A problem with content analysis is how to satisfactorily categorise and code what is often a large volume of unstructured data. Secondary analysis is about using quantitative data that were previously collected by other people for a different purpose. The general methods of secondary analysis differ very little from those used for primary data sources (Golden, 1976; Wallgren and Wallgren, 2007). Systematic review (sometimes referred to as meta-analysis) combines and investigates the output of multiple studies concerned with the same or a similar phenomenon.

Many NSIs may apply all three secondary research methods. However, without doubt the most commonly used method is secondary analysis, since usually the data content of secondary sources provides input for official statistics. Examples of secondary analysis from official statistics practice are processing of Value Added Tax data, from the tax office, for the short-term business statistics (Constanzo, 2011) and the use of administrative sources containing (human) population-related data for the Virtual Census. The other above mentioned two secondary research methods (content analysis and systematic review) may be less frequently used. A typical example of content analysis is a historical review of an NSI statistic or statistics. Examples of a systematic review are a publication in which time series of trade statistics are compared between various countries and an investigation into the relationship between cancer and nutrition by combining all data published on the subject in the scientific literature over the past 15 years.

2.2 *On terminology and types of secondary sources*

There is a wide range of terminology and definitions concerning 'secondary sources' and 'registers' which can be quite confusing. The first two terms we want to clarify is the distinction between a source and a register. SDMX (2009) defines a source as "a specific data set, metadata set, database or metadata repository from where data or metadata are available" and a register as a "data store where registered items are recorded and managed". The crucial point of a register here is that it is *managed*. The context of the SDMX(2009) definition of a register clarifies this further and explains that a (statistical) register is "a continuously updated list ...", which is also found in the definition of UNECE (2007). In summary, we follow the ideas of SDMX (2009) and use the term source as a general notion for a data set whereas we use the term register as a special case where data are stored and structured in such a way that they can be managed and continuously updated.

To clarify the term register further, we provide some more context. A selection of registers are specifically devoted to maintaining a population of objects by updating any changes in the properties of the objects. Examples are the so-called base registers (UNECE, 2007) that hold lists of objects that are used by public institutions (see below for more explanation) and a business register where statistical units with identifying variables are derived according to Eurostat recommendations. In addition to this however, there are also statistical registers (SDMX, 2009) where a number of data from different sources is integrated and continuously updated for statistical purposes. Thus, in the present paper, the term register is not synonymous to an updated list of objects, since in some registers also many variables are integrated, to be used for statistical output.

We will now further specify secondary data sources, since NSIs exploit a very diverse range of secondary sources. Examples of these are base registers, data on taxes, survey data from another survey oriented organisation in the country, price scanner data of supermarket products, and airline ticket prices from the internet. Some of these sources may be deemed to constitute an administrative source, but the distinction between administrative and other types of secondary data is unclear in some cases. Price data given on a website clearly do not constitute an administrative source, and neither are they maintained for administrative purposes.

From the viewpoint of NSIs information requirement, three main categories of secondary data sources can be discerned: statistical sources, administrative sources and organic sources (slightly modified from Daas and Arends-Toth, 2009). This categorisation is based on assessing the sources against their various characteristics. Figure 1 shows the various categories of secondary sources distinguished. In UNECE (2012) a more detailed list is included. The way by which the sources in each category can be integrated in the statistical process varies (Daas and Arends-Toth, 2009; Groves, 2011; UNECE, 2012) as will be explained below. Some examples are also included in figure 1 for clarification.

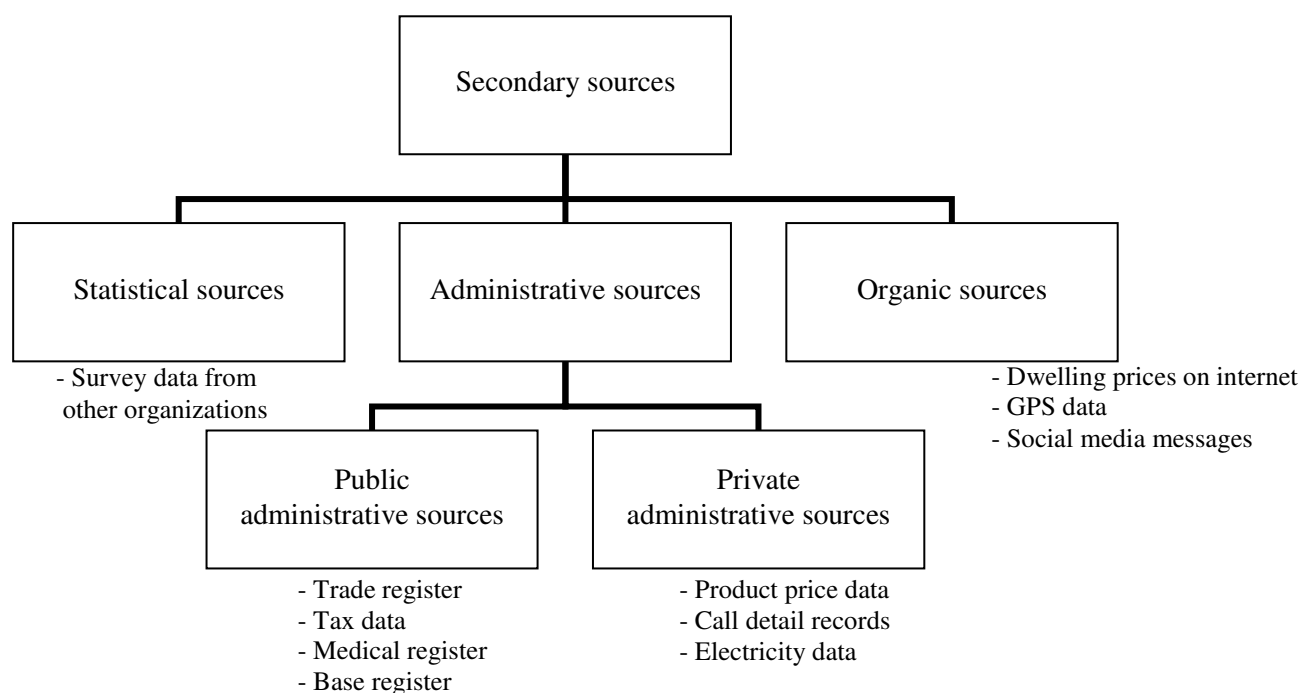


Figure 1. Distinguished categories of secondary sources with examples

The first main category concerns statistical sources. Statistical sources consist of statistical objects and statistical data, that can ‘easily’ be processed for official statistics. Among the statistical secondary sources used by NSIs are survey data collected by other (survey-oriented) organisations, such as those collected by market research organisations, or by government research bodies.

The second category are administrative sources, i.e., secondary sources that have an administrative purpose. Figure 1 shows that the administrative sources can be split into two subcategories namely public administrative sources and those from private (including companies) organisations. Examples of source in the first subcategory are the Trade Register, a National Medical Registration and the Population Register. Examples of the second subcategory are a database with prices of supermarket products, mobile phone call-detail records, and data collected by smart electricity and gas meters. They all serve an obvious administrative purpose.

A special group of administrative data from public organisation are so called ‘Base registers’. Base registers are special data sources that form the foundation of many government’s implementation tasks in the Nordic countries and in the Netherlands (UNECE, 2007). Base registers contain data that is frequently used by the government in policy, implementation and enforcement. The typical function of base registers is to keep stock of the population of objects at any given time. In addition, they have to maintain identification information to be used by other sources (UNECE, 2007). For instance there is a base register on individuals with data on address, age, sex etc. and a base register on legal units (‘firms’), with ownership, kind of economic activity. Since governmental organisations are obliged to use the data in base registers and report any suspected errors in the data, its use will improve the quality of the data. Storing data in a system of related base registers is expected to help improve quality (Wallgren and Wallgren, 2007). Next to the base registers, other (public) administrative sources may hold data that could be useful for some governmental organisations. When exploratory or feasibility studies reveal that using one of these sources might help to reduce response burden, the potential reduction in burden is substantial and that the data will be heavily used; it may be earmarked as a future base register.

Public administrative data can usually be related to a large proportion of the statistical target population. Administrative data from private organisations however, can often be related to only a subset of the target population, with the risk of being selective. Administrative data from private organisation may have to be complemented by other data in order to cover the target population. Examples of private administrative sources are call detail records of a mobile phone provider, electricity data of an energy company and product price data of a supermarket chain. For both subcategories of administrative sources it holds that the concept of the variables may be different from the target one, so derivation or estimation rules are needed to delineate the target variable.

The last group of secondary sources is composed of organic sources (Groves, 2011), indicating the fact that these sources contain data created in a more unstructured way. In fact, a considerable part of these sources can be identified as ‘Big Data’ (Glasson et al., 2013). Examples of organic sources are a dataset with dwelling prices collected from a website, satellite-based navigation system (GPS) data, and a collection of social media messages. For all these sources no immediate administrative use is foreseen nor do they cover an exactly defined population. The potential application of organic sources for statistics is the focus of several studies currently employed (Daas et al., 2013). Processing these sources poses a number of challenges. First of all, it is often difficult to link those data to a statistical

population, since no identification numbers are available and there is little to no auxiliary data available for those units (such as age, sex, etc.). One of the main advantages of these sources is their volume and their near real-time availability.

We now return from secondary sources that are maintained by external organisations to those maintained by NSIs. Statistical registers (UNECE, 2007), such as the Business Register (BR) and the Social Statistical Database (SSD), are *internal* NSI products. They are compiled from primary and secondary sources, and as such *cannot* be considered to be secondary data sources. A characteristic of statistical registers is that they contain an enumeration of statistical object, together with statistical data (properties) of those objects.

2.3 *Consequences of using secondary data*

NSIs that want to increase the use of secondary data sources for statistics usually aim to lower the response burden of respondents and/or the costs of data collection. Needless to say, the cost aspect is also affected by an NSIs secondary data acquisition expenses and the amount of work needed to transform these data to their requirements. Furthermore, some secondary sources tend to have data about a complete population, which enables the publication of extremely detailed statistics. Moreover, if integration of one or more secondary data sources is successful, new and detailed statistics can be published with no additional response burden (UNECE, 2007; Wallgren and Wallgren, 2007). The various ways in which secondary sources are used in statistics production (section 2.4) clearly show how advantageous the use of the types of data sources can be.

Downside of an increased use of secondary data, is that it makes NSIs become more dependent on:

- 1) the existence of and access to secondary sources;
- 2) the fitness for use (i.e., quality) of the secondary sources available;
- 3) the timely and stable delivery of secondary sources.

Problems in one or more of these dependencies can have serious implications on the production of statistical output. In the most extreme case an NSI might no longer be able to produce some of its statistics, when it solely depends on a single administrative source. The above mentioned three dependencies and the ways developed to cope with them are discussed in section 3.

2.4 *Types of use of secondary data by NSIs*

The benefits that secondary data sources offer makes them very interesting for statistics production. NSIs accordingly use secondary sources for the following statistical applications:

- 1) in statistics production as a replacement for primary data;
- 2) as a sample framework and source of auxiliary information in sample design (see also the topic “Statistical Registers and Frames”);
- 3) as a source of additional variables to be used for estimates;
- 4) as auxiliary information to support processing of primary data (e.g., data editing, imputation, calibration of estimates)
- 5) as input for statistical registers (such as the Business Register).

Also, the data in secondary sources may be ideal for some specific statistical applications, in particular when these data sources cover an almost complete population. These data sources can be used for:

- 6) detailed publications (such as regional statistics);
- 7) publications about special (infrequently occurring) events.

Secondary sources that cover multiple time periods and maintain a stable composition, over a relatively long period of time, are also very suited for:

- 8) detailed longitudinal studies.

The above mentioned uses make secondary data ideally suited for statistics production in our modern world, which demands statistics at a very detailed level without an increase in perceived response burden. Big data sources have the additional potential to produce very timely statistics (Glasson et al., 2013) and may enable the creation of so-called leading or even (nearly) real-time indicators.

2.5 Some practical issues concerning collection of secondary data at NSIs

When an NSI is using secondary data for statistics production some practical processing steps need to be taken. Firstly, the data needs to be transferred in a secure fashion from the data holder to the NSI. Data can, for instance, be transferred on a physical storage medium, such as a hard drive or DVD, to the institute or send electronically (web, email). If this is the case serious data protection measures need to be taken. Data should be encrypted and the decryption key should be send separately. Furthermore, it is good practice to store the data in a general file format, such as XML or CSV.

Secondly, the NSI needs to check whether the data received meets the quality standards agreed upon. This can be done by applying some elementary technical checks, such as whether the format is correct and the total number of columns agrees to the number expected. Thirdly, the data received needs to be uploaded into the data storage system of the NSI. When the data is uploaded and no problems have occurred it is good practice to check the input quality of the data, for instance the completeness of the records. This is described in more detail in section 3.3. This is also done in subsequent processing phases, for instance when different data sources are combined during the creation of statistical registers by micro-integration or other data integration methods. These processing steps however are beyond the scope of the this chapter.

3. Design issues

NSIs that use or start to use secondary data become more dependent on the availability of secondary data. Unavailability of a part of the source or the source as a whole can have serious implications on the statistical output. NSIs need to take measures to deal with the consequences of this dependency.

The remainder of this chapter is organised as follows. First, in section 3.1 a way to obtain an overview of existing secondary data sources in the country of interest is discussed. Next arrangements that enable structured access to secondary sources are listed. In section 3.3 data quality issues are discussed, followed by an overview of ways to deal with an interruption in the availability of secondary data. The chapter ends with guidelines on maintaining good relations with the data providers.

3.1 *Existence*

An NSI that wants to use secondary data needs to know what secondary sources are available in its country. The data protection law offers a good starting point. Countries that have set up a personal data protection act (DLA piper, 2013) generally have an organisation that registers all data sources and organisations that process data in which personal identifiers are included. The official authority responsible for data source registration is – very likely – able to provide a list of all data sources reported to them. For example, in the Netherlands, the website of the Dutch Data Protection Authority (www.dutchdpa.nl) has such a list available on their website. This list consists of all sources in which personal identifiers are included in the country and only lacks of i) sources that are exempted, such as membership and payroll records, and ii) databases used by the police and judicial authorities.

3.2 *Access*

To enable structural access to secondary data sources by an NSI, special arrangements may need to be made. The statistical offices of the Nordic countries have created an overview of their best practices that facilitate the large-scale use of data from secondary sources in their countries (UNECE, 2007; Statistics Finland, 2004). In summary, these are:

Legal basis: Legislation provides a key foundation for the use of secondary data sources for statistical purposes. Data protection arrangements must be part of these provisions.

Public approval: The general public must have no objection to the use of ‘their’ data for statistical purposes. The reputation of a statistical institute as a reliable and eminent user of secondary sources is an important factor in acquiring and preserving public consent.

Unified identification codes: It is vital that unified identification codes are used (for the various object types) across different sources. The identifiers enable fast data processing and give rise to fewer linkage errors. Sources without such identifiers can still be used, but costs are higher and their use will result in an increased number of errors (because of incorrect and missing links).

Reliable secondary data: The secondary sources used must contain reliable data covering as much of the target population as possible. The use of these sources by multiple official organisations and the population itself increases data reliability and decrease the chance of units missing from the target population.

Cooperation among administrative authorities: Effective liaison between the authorities involved in using and maintaining the sources helps in the development of a stable and reliable system of secondary sources. It is important that this is supported up to the highest management level in the organisations involved.

The reader needs to be aware that for the use of individual secondary sources specific agreements need to be made with the data provider regarding the delivery and other issues, such as the possibility for feedback or assistance.

3.3 *Usability (Fitness for use)*

NSIs will, very likely, use the data in a secondary source for a purpose different from that for which it was originally collected (see also the theme module “Data Collection – Techniques and Tools”). This may give rise to problems. For instance, a source may define an important variable, such as turnover,

(slightly) differently from the one used in official statistics leading to a reduced validity (Scholtus and Bakker, 2013). It is important that an NSI is able to access the fitness of use of a secondary source for official statistics, and to pinpoint the cause of the problem. These aspects are all related to the quality of secondary data. For an overview of the sources of error in secondary sources, the reader is referred to the paper by Zhang (2012).

In recent years quite a number of projects have been (partly) devoted to the study of the quality of secondary data used for statistics. Most noteworthy projects are the BLUE Enterprise and Trade statistics project (BLUE-ETS, 2013) and the ESSnet on the use of Administrative and Accounts data for Business statistics (ESSnet Admin Data, 2013). As such a whole range of possible ways to get grip on the ‘fitness of use’ of secondary sources is available. Main difference between the approaches developed is their focus. Three general approaches can be discerned which specifically focus on: 1) the quality of the input data of a secondary source (Daas et al., 2012), 2) the quality of the output of the statistics based on secondary data (Frost, 2011; Laitila et al., 2011; Burger et al., 2013) and 3) the metadata quality of secondary sources (Daas and Ossen, 2011).

Since the approaches suggested above complement each other, they – as a whole – constitute a more complete framework with implications of potential value for use in other contexts than first intended (Laitila, 2012). Below a short overview is provided of each of three general approaches discerned. The goal is to enable the reader to quickly decide which approach best covers his or her needs.

Input oriented data quality

When a secondary source enters an NSI, assessing its quality as early on as possible may be important for an NSI. When this is the case, the user has an *input oriented view* on the quality of secondary data.. Daas et al. (2013) have developed an evaluation procedure and a report card to structurally note the findings. The quality indicators used are grouped into five dimensions, these are: 1) Technical checks, 2) Integrability, 3) Accuracy, 4) Completeness, and a so-called 5) Time-related dimension. These dimensions contain indicators that specifically focus on: 1) the technical usability of the file and data in the file, 2) the extent to which the data source is capable of undergoing integration or of being integrated, 3) the extent to which data are correct, reliable and certified, 4) the degree to which a data source includes data describing the corresponding set of real-world objects and variables, and 5) the indicators that are time and/or stability related, respectively. To ease the use of the procedure, the indicators have been incorporated in the ‘dataquality’ package for the open source statistical programming environment R (R core team, 2014). Because the time required to thoroughly evaluate secondary data is a serious issue, a visualisation based approach (a ‘tableplot’) has also been developed; as a quick and general applicable alternative. This allows the creation of data ‘pictures’ of sources and subsequent deliveries, enabling a comparison of these ‘pictures’ for a selected number of variables over time. The reader is referred to the paper of Tennekes et al. (2013) for more details on this topic. The approach has been applied to various administrative sources used in business statistics in several countries.

Output oriented data quality

The ultimate intention of secondary data is its use for the production of statistical output. The quality of such output is obviously affected by the quality of the secondary data in the source, by the combination of sources used, and by the quality of the production process itself (Laitila, 2012). This

makes the assessment of the quality of the output based on secondary data a difficult task. In recent years, two ways to determine this have been independently developed.

The first one is described by Frost (2011). This work is based on the dimensions of quality proposed by Eurostat (Eurostat, 2003). The dimensions discerned are: a) Accuracy, b) Timeliness and punctuality, c) Comparability, d) Coherence, e) Cost and efficiency, and f) Use of administrative data. The dimensions each contain indicators that particularly focus on: a) the closeness between an estimated result and the unknown true value, b) the lapse of time between publication and the period to which the data refer and the time lag between actual and planned publication dates, c) the degree to which data can be compared over time and domain, d) the degree to which data that are derived from different sources or methods, but which refer to the same phenomenon, are similar, e) the cost of incorporating admin data into statistical systems, and the efficiency savings possible when using admin data in place of survey data, and f) background information relating to admin data inputs. The framework has been applied to various administrative sources used in business statistics in several countries.

The other approach is based on the framework developed in Sweden (Laitila et al., 2011). This general framework contains quality indicators divided into four groups. The groups discerned are i) Metadata, ii) Accuracy, iii) Integration with a base register, and iv) Integration with other data sources. The indicators in each group report evaluation findings on: i) the information available from the data provider, ii) the results of analysis and data editing of the source, iii) the results of integrating the source with the relevant base register, and iv) the results of integrating the source with relevant other primary and secondary sources. Evaluation starts with the metadata contents of the source followed by the accuracy of its content. The integration steps focus on the incorporation of the source into the statistical system by first relating it to a base register, followed by addresses the issue of how the source can be utilised for improving other relevant statistics produced by the NSI. In the end, the findings are summarised in a quality report card. This framework has been applied to several Swedish sources (Daas et al., 2013).

Metadata quality

Apart from the quality of the data, evaluating the metadata quality components of secondary sources is very important (Daas and Ossen, 2011). Next to the general Swedish approach described above – covering both metadata and data quality –, specific metadata quality specific alternatives are available for secondary sources. It is highly recommend that quality evaluation of secondary sources starts with the evaluation of metadata quality. Advantage is that it i) enables the identification of important issues very early on in the process that ii) not immediately a great deal of attention and work is put into the evaluation of quality of the data aspects. The latter is often the case in practice.

Metadata quality evaluation approaches have been described by Daas et al. (2009) and by Verschaeren (2012). In both approaches two different views on metadata quality are discerned, notably: 1) those essential for the delivery of the source and 2) the conceptual metadata quality indicators. The quality indicators in the first view are related to the stable delivery and the continuation of the access to the source by the NSI. The indicators in this view focus on the provider of the source, the relevance of the source, privacy, security and delivery issues, and procedures. In the second view the availability and comparison of conceptual metadata definition of the units, variables, and reporting period(s) in the source with those of the NSI are evaluated. Here the description of the metadata of the provider is

evaluated and compared to those of the NSI. In addition the inclusion of unique keys, and any data checks performed by the provider of the source is studied. The latter is very important process-related meta-information because it highly affects the quality of the secondary data source. Evaluation of both views is guided by and the findings are summarised in a specific metadata checklist. The checklist is regularly applied in the Netherlands (Daas et al., 2009).

An important addition included in the work of Verschaeren (2012) is the explicit mentioning of the keeping a repository of evaluation information on secondary data sources. This assures findings are structurally stored. For this approach a pre-evaluation checklist has been created. The list has been tested in Belgium and in the Netherlands.

Other alternatives

Apart from the overview provided above, there are also several interesting alternatives suggested by others. Two of them are particularly interesting and they will both be briefly mentioned here. The NSI of New Zealand has proposed a quality framework on administrative sources from a business statistics perspective (McKenzie, 2009). Both a preliminary assessment of data quality and a process management oriented way on quality are discerned. A very different way of dealing with data quality is to increase co-operation with the data provider who could, for instance, implement additional checks upon request of the NSI. Statistics Norway is pursuing this approach (Hendriks, 2012).

3.4 Coping with interruptions

When an NSI starts using secondary data for statistics production, it seriously needs to consider the effect of an interruption or delay in the delivery of the source. The problems that may occur and their effects on the statistics that use the source need to be identified and scored by using risk analysis. Depending on the importance of the statistics based on the secondary sources, the risk analysis may indicate the need for implementing measures to cope with the potential (temporary) loss of secondary data. The combined set of measures constitutes a so-called fall-back scenario. In all situations maintaining good contact with the data provider is essential.

Risk analysis

The standard process specification used by NSIs, such as those applied for the required availability of information systems (including databases), can also be used to assess the risk of unavailability of a secondary source. In the Netherlands, a template has been created to determine the need of developing a fall-back scenario for a given statistic (table 1). The risk assessment component of the template estimates whether there is any need to create a fall-back scenario. Among the components considered are the assessment of problems with the delivery of the source, the stability of the delivery, and the impact on the statistical output. If delivery problems are likely to occur, with severe consequences for the NSI, it is advised to draw up a fall-back scenario. In all other cases, the NSI manager responsible for the statistics produced needs to decide whether or not a fall-back scenario has to be developed.

Fall-back scenarios

It is unrealistic to prepare fall-back scenarios for all imaginable situations. In our experiences, fall-back scenarios are often tailored to specific situations. The best solution in any given situation will depend on what exactly has occurred, what part of the data is missing and the quality of the data available. The chosen solution *must* also address the costs and time available, which will usually be

short. It is therefore advised to draw up fall-back scenarios only for statistics for which the unavailability of secondary data will have serious consequences. The early detection of potential problems increases the chance of a satisfactory response. This is why active relationship management, contact with the data provider, is very important. For sources on which several statistics depends, more than one fall-back scenario may have to be drawn up.

Table 1. Dutch evaluation template for fall-back scenario

<i>Which statistics are involved?</i>
<ul style="list-style-type: none"> • Name • Division, sector, task force • Uses the following secondary sources: ...
<i>General information about each secondary source</i>
<ul style="list-style-type: none"> • Name of source • Name of data provider • Contact person at provider • NSI contact person for the source/provider • Other NSI contacts (if any) • What regular contacts are there between the data provider and NSI?
<i>Risk assessment</i>
<ul style="list-style-type: none"> • How great is the estimated risk of the data provider being unable to deliver the source? • What are the consequences for the NSI? • How stable is the delivery of the source?
<i>Process information of the statistic</i>
<ul style="list-style-type: none"> • Are there any alternative sources, or does any research exist which indicates that the data could be derived from a model if the source or any of the required variables are unavailable? • Possible fall-back scenarios: 1. wait; 2. model-based approach; 3. use alternative source
<i>Summary</i>
<ul style="list-style-type: none"> • Risk of untimely publication or non-publication of the statistic • Consequences for NSI • Available alternatives
<i>Meta-information checklist</i>
<ul style="list-style-type: none"> • Update frequency of the checklist • Date of last update • Drawn up by: • Signed (name and position)

No fall-back scenario has to be drawn up for a source that becomes permanently unavailable. In this case, a new statistical data collection process, needs to be started or re-organised in order to satisfy the statistical output obligation.

The transition period may be lengthy. External pressure and publication obligations may necessitate the introduction of other ‘creative’ temporary solutions in the meantime, such as a completely model-based figure, a nowcast, an expert ‘guess’, or even the use of the Delphi method. It goes without saying that the use of such temporarily solutions must be communicated clearly to the outside world. The emergency measure applied in the transition period can be viewed upon as a temporary fall-back scenario.

The following general approach is recommended for developing a scenario for dealing with the temporary unavailability of an important secondary source:

1. determine whether it is feasible, in terms of time and costs, for NSI-employees to obtain – preferably via alternative external sources – the missing data elsewhere;

2. apply a model-based approach if there is no alternative for the missing data and some of the data about the reporting period are available. Application is subject to the plausibility of the quality of the results provided by the model, so develop and test the model in advance;
3. notify the important users of the potential consequences of unavailability of the source;
4. postpone publication or decide not to publish at all if the above options are impossible.

Postponement is not an option for very important statistics, such as unemployment or economic growth. In such cases alternatives, for instance other sources or a model, *must* be available. Since it may take considerable time to develop a model, it should be created shortly after the need for a fall-back scenario was identified.

We provide an example to illustrate the advantage of having fall-back scenarios available. In the Netherlands quarterly turnover levels and changes are estimated for populations of enterprises, classified by kind of economic activity, for the Short Term Statistics. These estimates are obtained by combining sample survey data with Value Added Tax (VAT) data, provided by the Tax office, leading to observations for nearly all population units. Because there had been some irregularities in the delivery of the VAT data, a fall back scenario was developed. In this we distinguished two situations (a) the delivery problem is discovered very shortly before the planned publication date or (b) the delivery problem is known well in advance.

In situation (a) we use a model-based approach in which the growth rate is estimated as a weighted combination of growth rate of the sampled and non-sampled units in the population. For the non-sampled units normally quarterly VAT values are available. However, in case of delivery problems only the values from monthly VAT emitters are available for the first two months of the quarter. The VAT-data of those two months are subsequently used to estimate the growth rate of the non-sampled population combined with survey data obtained by directly contacting crucial enterprises and requesting their quarterly turnover. Crucial enterprises are those missing units that have the largest (historical) turnover, up to a certain threshold which is determined by the desired accuracy of the estimate.

In situation (b) we send out a small sample survey, for which stratified random sampling is used. The sampling design has been made in advance and can be used directly when needed.

3.5 *Contact management*

An increase in the use of secondary sources necessitates good relations with the suppliers of the sources: the data provider. This is an activity that is in the field of relationship management. A way to deal with this is appointing supplier managers for the most important data providers, such as the Tax Administration, the Chambers of Commerce and the owners of the Population Register. These managers are required both to provide and to gather information to and from the sources under the responsibility of their contacts. The duties also may include making and monitoring agreements, managing expectations and detecting new developments. For instance, an appointment for exploratory talks will be made with a view to establish the statistical usability of a potential source. Clear agreements must be drawn up with data providers for sources that an NSI decides to use, covering the delivery of the source (including metadata), the use of the data in the source, and the mutual obligations involved. The agreements must be recorded in a formal contract.

The supplier managers are involved in contacts with the data providers at a strategic level: to enhance the long-term relationship between the NSI and the data provider. It is good practice to also make the supplier manager the NSI's internal contact person for any questions and problems regarding the source, its delivery, and the data provider. As a consequence, nearly all contacts with the data provider are channelled through, or follow consultation with, the supplier manager.

A part of the tasks of the suppliers manager, especially those concerning contacts at operational level, concerning daily production issues, may be delegated to other representatives within the NSI. For instance, an NSI may appoint someone who monitors and verifies whether the agreements on delivery and other quality requirements are met. This representative contacts the data source holder in case the delivery of administrative data is too late or incomplete. At tactical level, consultations between data source provider and NSI may concern desired delivery schemes, and objects, variables and classifications within the data source. These consultations may, for instance, include annual meetings at a high (administrative) level, three-monthly user meetings, or two-monthly bilateral meetings of technical experts. The owners of the statistical process that uses the secondary data will usually join those consultation meetings; supplier managers will not necessarily attend all meetings of this kind. Needless to say, the supplier manager needs to be kept informed on the outcome of all meetings.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

BLUE-ETS (2013), Project description on the BLUE Enterprise and Trade Statistics website. (<http://www.blue-ets.eu>)

Burger, J., Davies, J., Lewis, D., van Delden, A., Daas, P., and Frost, J-M. (2013), *Quality guidance for mixed-source statistics*. Deliverable 6.3 of ESSnet Admin data, February 2013.

Costanzo, L., Di Bella, G., Hargreaves, E., Pereira, H. J., and Rodrigues, S. (2011), An Overview of the Use of Administrative Data for Business Statistics in Europe. Paper for the 58th Session of the International Statistical Institute, Dublin, Ireland.

Daas, P. J. H. and Arends-Tóth, J. (2012), Secondary Data Collection. Statistical Methods 201206, Statistics Netherlands, The Hague/Heerlen.

Daas, P. J. H. and Ossen, S. J. L. (2011), Metadata Quality Evaluation of Secondary Data Sources. *International Journal for Quality Research* 5, 57–66.

- Daas, P. J. H., Ossen, S. J. L., Tennekes, M., and Burger, J. (2012), Evaluation and visualisation of the quality of administrative sources used for statistics. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.
- Daas, P., Ossen, S., Vis-Visschers, R., and Arends-Tóth, J. (2009), Checklist for the Quality Evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
- Daas, P. J. H., Puts, M. J., Buelens, B., and van den Hurk, P. A. M. (2013), Big Data and Official Statistics. Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium.
- Daas, P. J. H., Roos, M., de Blois, C., Hoekstra, R., ten Bosch, O., and Ma, Y. (2011), New data sources for statistics: Experiences at Statistics Netherlands. Discussion paper 201109, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
- Daas, P. J. H., Tennekes, M., Ossen, S. J. L., Di Bella, G., Galiè, L., Laitila, T., Lennartsson, D., Nilsson, R., Wallgren, A., and Wallgren, B. (2013), *Guidelines on the usage of the prototype of the computerized version of QRCA, and Report on the overall evaluation results*. BLUE-ETS deliverable 8.2, March 2013.
- DLA piper (2013), *Data Protection Laws of the world*. Second edition, March 2013.
http://www.dlapiper.com/files/Uploads/Documents/Data_Protection_Laws_of_the_World_2013.pdf
- ESSnet Admin Data (2013), Project description on the web site of the ESSnet on the use of Administrative and Accounts data for Business statistics. (<http://essnet.admindata.eu>)
- Eurostat (2003), *Quality Assessment of Administrative Data for Statistical Purposes*. Working group on assessment of quality in statistics, Luxembourg, 2-3 October.
- Frost, J. M. (2011), Development of Quality Indicators for Business Statistics Involving Administrative Data. Paper for the 58th Session of the International Statistical Institute, Dublin, Ireland.
- Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., and Khan, A. (2013), What does “Big Data” mean for Official Statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services, March 10.
- Golden, M. P. (1976), *The research experience*. F.E. Peacock Publishers Inc., Itasca, Illinois, USA.
- Groves, R. M. (2011), Three Eras of Survey Research. *Public Opinion Quarterly* **75**, 861–871.
- Hendriks, C. (2012), Input Data Quality in Register-Based Statistics: The Norwegian Experience. Paper for the Joint Statistical Meeting, San Diego, U.S.A.
- Hox, J. J. and Boeijs, H. R. (2005), Data collection, Primary vs. Secondary. *Encyclopaedia of Social Measurement* Vol. 1, 593–599.
- Laitila, T. (2012), Quality of registers and accuracy of register statistics. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.

- Laitila, T, Wallgren, A., and Wallgren, B. (2011), Quality Assessment of Administrative Data. Research and Development – Methodology reports from Statistics Sweden, 2011:2, Stockholm/Örebro, Sweden.
- R Core Team (2014), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Scholtus, S. and Bakker, B. F. M. (2013), Estimating the validity of administrative and survey variables by means of structural equation models. Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium.
- SDMX (2009), Statistical Data and Metadata eXchange content-oriented guidelines, Annex 1: Cross-domain concepts. SDMX website.
http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf
- Statistics Denmark (1995), *Statistics on persons in Denmark, a register-based statistical system*. Office for Official Publications of the European Communities, Luxembourg.
- Statistics Finland (2004), *Use of registers and administrative data sources for statistical purposes*. Best practices of Statistics Finland, Handbook 45.
- Stewart, D. W. and Kamins, M. A. (1993), *Secondary research, information sources and methods*, second edition. Sage publications, Newbury Park, Ca., USA.
- McKenzie, R. (2009), Managing the quality of administrative data in the production of economic statistics. Paper for the 57th Session of the International Statistical Institute, Durban, South Africa.
- Tennekes, M., de Jonge, E., and Daas, P. J. H. (2013), Visualizing and Inspecting Large Datasets with Tableplots. *Journal of Data Science* **11**, 43–58.
- 't Hart, H., Boeijs, H., and Hox, J. (2005), *Research methods*, 7th impression. Boom, Amsterdam.
- UNECE (2007), *Register-based statistics in Nordic countries – review of best practices with focus on population and social statistics*. United Nations Publication, Geneva.
- UNECE (2012), *Using Administrative and Secondary Sources for Official Statistics. A Handbook of Principles and Practices*. (http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf)
- UN Global pulse (2012), *Big Data for Development: Challenges & Opportunities*. White paper, May. (<http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobaIPulseJune2012.pdf>)
- Verschaeren, F. (2012), Checking the Usefulness and Initial Quality of Administrative Data. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.
- Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology, John Wiley & Sons, Ltd., Chichester, England.
- Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistics Neerlandica* **66**, 41–63.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Registers and Frames – Main Module
2. Data Collection – Techniques and Tools

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

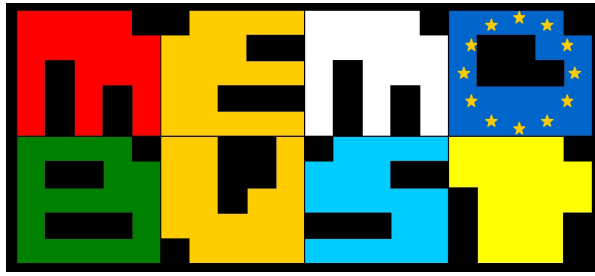
Data Collection-T-Secondary Data Collection

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	29-03-2013	first version	P. Daas, A. van Delden	Statistics Netherlands
0.1.1	22-04-2013	first revision	M. Murgia	ISTAT
0.2	31-05-2013	second revision	P. Daas, A. van Delden	Statistics Netherlands
0.2.1	04-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:51



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Response Process

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Response models for business surveys.....	3
2.2 Application of the response process model	7
3. Design issues	9
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

The statistic survey perspectives can be viewed as a design perspective and quality perspective (Groves et al., 2004). The design perspective leads from concepts through “constructs” and measurements to questions to become a process and one of its stages is the response process. The quality perspective makes numerous references to “error”. The sampling error, nonresponse error are just two examples. Measurement errors refer to the gap between what is called the ideal value and the obtained response, i.e., at the response process stage. Survey methodologists attribute deviations from perfect measurements to cognitive problems in the response process. Hence, these problems lie at the heart of the response process model. Originally, the model was developed to reflect aspects of households and individual surveys. Further development of cognitive research extended the model to fit the response process in business surveys. A merger of the two produced a Hybrid Response Process Model for Business Surveys, a complex and general model encompassing the entire response process in business surveys. Since it still did not fully and clearly address numerous aspects the model has recently been developed into the Multidimensional Integral Business Survey Response Process Model.

The response process models can serve as a framework for the evaluation of business surveys (Giesen, 2007). The linkage between model steps and observations of real respondent behaviour when dealing with survey requests, provides the structure which can help to analyse this complex activity. This is a way to spot problems and try to fix them for the future. Furthermore, considering the data collection instrument and the response burden connected with answering items it contains, response process steps make it possible to establish at which stage the burden is especially heavy and what can be done to ease it. This can improve the questionnaire and even influence its design. The division of the response process into separate stages was the foundation of cognitive methods for pretesting survey questions. Cognitive interviewing, understood as an extension of the standard interviewing process of eliciting answers to questions, studies processes distinguished in the response process model (Willis, 2004). The foundation of the response process for establishment surveys, which is more complex and contains more steps, adequately allows to split survey evaluation into the response process steps. When the data collection process and the response burden are assessed using different methods (Giesen, 2007) and the findings are linked with the response process stages it is possible to establish the nature of the problems, whether cognitive or logistic, and consequently, adopt the results to improve the data quality or ease the response burden.

2. General description

2.1 Response models for business surveys

The starting point is the respondent’s task in the interview. The cognitive analysis of the task provides the basis for a description of operations the respondent must go through to arrive at an answer to a survey question. The widely adopted model for answering questions in interviews was introduced by Tourangeau (1984) and consisted of four basic consecutive steps:

1. Comprehension – first, understanding the meaning of the question.
2. Retrieval – recalling the relevant information.

3. Judgment – formulating an answer based on recalled information.
4. Communication – formatting the answer to fit the demands.

The psychological aspect of the question-answer process and its social dimension are accounted for in the general response model mentioned above. However, potential sources of the measurement error exist even before these four cognitive steps. Eisenhower et al. (1991) introduce another step at the top of the list, namely “Encoding”. Addition of another step was motivated by the fact that before the four steps of the model take place memory must be formed from experiences of the respondent. The earlier model was mostly suited to individual and household surveys, because it relied on social interaction of interviewing and memory engagement. By comparing the differences between responses in household survey and establishment surveys, Edwards and Cantor (1991) developed the response model for establishment surveys. The major difference between those models results from the fact that establishments often use information systems or records, not memory, to obtain knowledge to a question. Hence, the *record formation* step in a business survey is an equivalent of the cognitive *encoding* step in a household survey is. Similarly, *retrieval* from memory is analogous to the *record look-up* process in establishments. The decision which source is to be used – records or memory – calls for yet another step: the *source decision*. The cognitive activities of comprehension, judgment and communication apply directly to the establishment response model. Finally, the model consists of six steps:

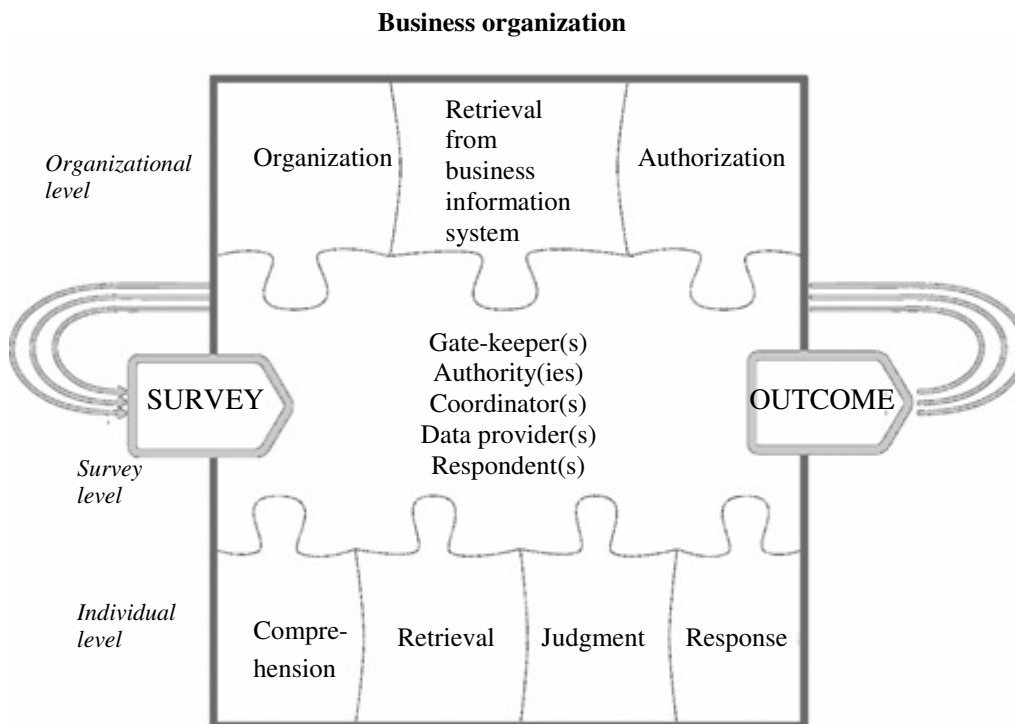
1. Encoding/Record formation.
2. Comprehension.
3. Source decision.
4. Retrieval/Record look-up.
5. Judgment.
6. Communication.

Exploratory research on reporting to statistical surveys and its findings produced the Hybrid Response Process Model for Establishment Surveys (Sudman et al., 2000; Willimack and Nichols, 2001). This model extends and revises the previous models by explicitly distinguishing organisational and cognitive steps of the response process in establishment surveys. Singling out the consecutive phases in the survey request tasks performed by an establishment links the cognitive and organisational factors of the process. The combination of cognitive and organisational levels results from a qualitative study of large establishments conducted by Sudman et al. (2000). The model was slightly modified and complemented by Willimack and Nichols (2001). The organisational context creates a framework for cognitive factors. The complete model includes the following steps:

1. Encoding in memory/record formation.
2. Selection and identification of the respondent or respondents.
3. Assessment of priorities.
4. Comprehension of the data request.
5. Retrieval of relevant information from memory and/or existing company records.

6. Judgment of the adequacy of the response.
7. Communication of the response.
8. Release of the data.

While cognitive factors remain valid, the additional steps are motivated by the complex nature of the response process in establishments, perceived as living organisms with goals other than releasing information for statistical purposes. In other words, the organisational steps in the hybrid model can be treated as integral processes, which characterise a business as an organism, while the individual steps connect the four step model with personal abilities associated with comprehension, retrieval, judgment and communication; the result is the Multidimensional Integral Business Survey Response Process Model (MISBR) proposed by Bavdaž (2010). This model integrates previous findings with new research results. The model addresses the two composite layers of the response process in establishments – the organisational layer and individual layer. Between the two layers the model distinguishes the survey layer, which provides a link between them. The illustration beneath provides the author's visual representation of the model.



reprinted by permission of the author (Bavdaž, 2010)

The *Organisational layer* includes a complex list of factors, which influence the consecutive steps of the response process. Respondent selection and assessment of priorities are an integral part of the organisational layer have their own significance as far as organisational priorities and individual priorities are concerned. Individual priorities and organisational priorities may not always be unified. Typical organisational factors influencing the response process are: tradition, customary practices, established procedures and the location of information. Individual and organisational mixture of factors include competing tasks and formation and delivery of requested data. Business policy to

surveys and the individual attitude to tasks concerning these surveys are both influential factors, which connect the organisational and individual level. At the organisational level the model also distinguishes retrieval from business records and authorisation of the business response. Retrieval is based on the business information system. How the system is organised depends on two kinds of factors: internal and external. External factors include: legal obligations, standards and benchmarking practices. These factors are imposed more or less from outside of an organisation. Internal factors, on the other hand, depend on management needs. The record formation process is conditioned by the kind of business activity and its environment. It is also related to the problem of data availability. As a result, response forms a kind of continuum: from exact values through various levels of estimation to nonresponse in extreme cases.

The *Individual layer* moves the response process from the organisational level to the individual level, since participants of the process are individuals, who act according to their own cognitive processes. The stages of the individual response process, i.e., comprehension, retrieval, judgment and communication, are linked to the organisational level. The multidimensional integral business response process distinguishes three types of knowledge needed in the response process: the knowledge of the business reality, the knowledge of record formation, the knowledge of business records. Comprehension of the business reality involves matching survey variables with business activity and determining its relevance for survey questions. Retrieval is closely connected with business records and therefore the knowledge of business records is a key element, provided the required data are stored in business systems. In case they cannot be obtained, the business reality can be a helpful factor. Judgment, in turn, refers to the compilation of possessed information and the record formation process to properly link the data with business concepts. During the communication step, the business knowledge from records must be edited and categorised to suit the format required by the measuring instrument.

The *survey layer* accounts for the response process during surveys and refers to the general implementation of the survey response as well as repeated response to the same surveys. The layer can be used to conceptualise the influence of various elements of a survey on the response process. Distinguishing this level enables the observation how survey design components influence the response process. For example one of the observed dimensions at this level is the impact of repeated administrations of the survey to the same respondent on the organisation of the response. The other example of the dimension can be the impact of respondent's contact with the survey staff on response. At the survey layer the focus is on repeated administrations of the survey response. Additionally, the layer allows observation of a contagious effect transmitting the experience from one business survey to other business surveys (Bavdaž, 2010).

In addition to distinguishing organisational and individual levels of the business survey response process, the model also assigns different roles to people taking part in the process. All those people participate in the process at the organisational level, but at the individual level they have their own internal cognitive processes. Their roles and their influence on the response process exceeds the four-steps of the cognitive model (comprehension, retrieval, judgment and response) since their participation may only be episodic, at various points of the process, and may not affect the later understanding of questions or the organisation of the response. The model distinguishes the following roles: the gate-keeper (a person or a unit that brings information into an organisation or sends information from an organisation to the surrounding environment), the supervisor or people with

authority, the data provider and the respondent. The completion of the response process may even require the participation of persons from outside of an organisation or contacts with a survey agency. The *survey* level draws attention to the fact that repeating the same activities leads to routine performance of tasks, which may be done only partially or superficially. On the other hand, the repetitive character of reporting procedures may even eliminate the need for a supervisor by progressively supervision in consecutive rounds of recurrent surveys.

2.2 *Application of the response process model*

The widely adopted response process model developed by Tourangeau (1984) created a framework for cognitive methods of questionnaire pretesting in household and social surveys. The aim of these methods is to improve questions and to reduce measurement errors. The study of the response process model steps supports the development of rules for questionnaire design, but the main goal of cognitive methods is to evaluate survey questions and change them whenever necessary (Willis, 2004). The development of response process models for establishment surveys turns the attention to the complexity of the response and the burden associated with it. A better understanding of the process of establishment's statistical reporting may reduce the response burden (Sudman et al., 2000). Establishment activities at each step of the process and interactivity between them may increase or reduce the burden, and consequently result in item non-response and influence data quality (Hak et al., 2003). The evaluation of questionnaires used in the field for data collection and the detection and understanding of the problems connected with them can be based on the extended hybrid response process model for business surveys (Giesen, 2007). Research on the response process model provides results for data users and data collectors (Willimack and Nichols, 2010). Conclusions for data users include, among others, the awareness of possible cases of non-availability of the required data in the context of complex and burdensome nature of business surveys. Data collectors can use it as a basis to improve data collection instruments and to facilitate data collection process.

Moving down the extended hybrid model (Sudman et al., 2000; Willimack and Nichols, 2001) the consecutive steps can be briefly characterised as follows:

Encoding in memory/record formation step links two aspects of the process: cognitive and organisational. Two approaches are possible depending on the type of required information: categorical data or figures. In the first case, data can be *usually* retrieved from memory; in the latter case it is *usually* necessary to consult transactional systems. In this case memory is needed, too, to recall the knowledge of company systems. The greater an establishment is, the more complex the acquisition of information may be. What is important, however, is that such data actually exist in the systems, though it is not a sufficient condition. Studies show that businesses keep their data according to:

- management needs,
- regulatory compliance,
- established standards.

The influence of data collectors on record formation would be very desirable and could facilitate the *retrieval* step (Willimack and Nichols, 2010). Another application from empirical observations of this

step in establishments can be the relaxation of requirements concerning burdensome items of the survey or items for which information may not exist (Giesen, 2007).

Selection and identification of the respondent or respondents

Researches draw attention to the fact that the selection of a proper respondent can reduce the measurement error (Edwards and Cantor, 1991; Willimack and Nichols, 2010). The step is singled out on account of its further consequences for cognitive steps. The respondent can be more of a coordinator, whose task may focus on compiling collected pieces of information. Since it is very likely that data from many users are required to answer survey questions, questionnaires should enable respondents to forward different parts of the questionnaire to different users (Giesen, 2007). In the case of electronic collection instruments, features which facilitate the distribution of questionnaires among users of an organisation can also decrease the response burden by involving multiple users in the response process.

Assessment of priorities

Tasks in establishments have their priorities. Statistical obligations are ranked low on the list of priorities. They are defined as “Other government data requests” (Willimack and Nichols, 2010). Government reporting duties generate costs to establishments. Factors which respondents need to pay attention to – and are therefore worth to underline in the design of elements in the survey related to response – include:

- mandatory status of the survey,
- clear due date, explicitly given, according to the standard date format of the country,
- advance notice of new surveys.

The mandatory status of the request is a feature implicitly distinguished by respondents (Willimack, 1999) and therefore worth stressing. Feedback from a statistical agency underlining the importance of the supplied data is recommended as an incentive to respondents (Giesen, 2007).

Comprehension of the data request

Comprehension is a typical cognitive step. Understanding varies among respondents, which emphasises the importance of respondent selection. The key factor is the knowledge of business reality. Respondents fit the meaning of a given concept to standards used in business practice such as accounting standards. There are several additional factors, which complicate the response process in the case of electronic reporting as opposed to paper questionnaires (Morrison, 2005). Electronic instruments require a friendly user interface and the user-centred design, which can improve the understanding of the instrument and contribute to a positive image of electronic reporting.

Retrieval of relevant information from memory and/or existing company record

As mentioned above, the *record formation* stage is connected with the physical availability of the requested data. This is only a first step. Another problem involves data retrieval from company records, which may be difficult either owing to the complexity and the subject scope of questions or because of the organisational complexity of an establishment. At a more specific level, two questions should be asked: first, to what extent do survey concepts match business practice? Any deviations in this respect can influence comprehension. Secondly, who has the ability to access company data?

Another problem is connected with compiling individual pieces of information into one response item. As can be seen, the overlap between the steps of *respondent selection* and *retrieval* requires cooperation between company employees. This should be considered when designing collection instruments: namely, they should facilitate the distribution of a part or parts of the questionnaire to ease the burden of response. Given respondents' familiarity with spread sheets, questionnaires can become more user-friendly when they are organised like spread sheet tools (Willimack, 2010). Record formation factors also influence data availability. Several levels of data availability in business systems are reflected at various levels of the response outcome (Bavdaž, 2010). In some cases, answering questionnaire items may require estimation. The response outcome can vary from approximate values to item non-response. The recommendation for data collection instruments is to explicitly inform when estimation is acceptable or to add an field where estimated values can be entered (Giesen, 2007).

Judgment of the adequacy of the response

The collected information are assessed to determine if they meet the requirement criteria. At this stage data can be submitted to various operations such as summation, categorisation. Figures may represent an estimated value if exact data could not be acquired. Studies stress the role of questionnaire instructions as tools to judge the correctness of prepared data and their continuity, which means that procedures established previously are also valid in future periods, which also means that errors made earlier are carried over to the future. In the case of business surveys the prevalent data collection mode is the self-administered questionnaire. Electronic data collection instruments contain edit checks. Edit messages help to form a judgment about the validity of the response (Morrison, 2005). Built-in edit rules can encourage the respondent to review the data for accuracy or provide an explanation when the rule is not satisfied. The module "Questionnaire Design – Editing During Data Collection" discusses issues connected with editing within the questionnaire.

Communication of the response

Communicating the response means matching prepared data to fit the options of the measuring instrument. Electronic questionnaires ushered in "editing" at the data collection stage. Data consistency may require some correctional operations. Respondents are generally positive about electronic reporting (Willimack, 2010), which may be due to the common use of spread sheets, no matter what skills users have. The burden of communicating the response may also be seen by respondents as unwarranted as opposed to the burden connected with retrieval (Hak et al., 2003). The user-centred design can address many issues in order to facilitate the communication of the response.

Release of the data

Studies indicate that this step may require authority. What literature refers to as "social desirability" can also be observed in establishment surveys. The company's desire to comply with external obligations and the concern to project a good public image may entail an internal policy, whereby any information leaving the company must first be approved by the management. Another factor is the confidentiality of business activity, which raises the question of trust towards a statistical agency.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bavdaž, M. (2010), The multidimensional Integral Business Survey Response Model. *Survey Methodology* **36**, 81–93.
- Edwards, W. S. and Cantor, D. (1991), Toward a Response Model in Establishment Surveys. In: P. P. Biemer et al. (eds.), *Measurement Error in Surveys*, John Wiley & Sons, New York, 211–233.
- Eisenhower, D., Mathiowetz, N.A., and Morganstein, D. (1991), Recall Error: Sources and Bias Reduction Techniques. In: P. P. Biemer et al. (eds.), *Measurement Errors in Surveys*, John Wiley & Sons, New York, 127–144.
- Giesen, D. (2007), The Response Process Model as a Tool for Evaluating Business Surveys. *Proceedings of the Third International Conference on Establishment Surveys (ICES-3), 18-21 June, Montreal, Canada*, American Statistical Association, Alexandria, VA, 871–880.
- Groves, R. R., Fowler Jr., F. J., Couper, M. P., Lepkowski, M. J., Singer, E., and Tourangeau, R. (2004), *Survey Methodology*. John Wiley & Sons.
- Hak, T., Willimack, D., and Anderson, A. (2003), Response Process and Burden in Establishment Surveys. *Proceedings of the 2003 Joint Statistical Meetings - Section on Government Statistics*, 1724–1730.
- Morrison, R. L., Anderson, A. E., and Charles, F. (2005), The Effect of Data Collection Software on the Cognitive Survey Response Process. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Sudman, S., Willimack, D. K., Nichols, E., and Mesenbourg, T. L. (2000), Exploratory Research at the U.S. Census Bureau on the Survey Response Process in Large Companies. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 327–337.
- Tourangeau, R. (1984), *Cognitive science and survey methods: a cognitive perspective*. In: T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, National Academy Press, Washington, DC.
- Willimack, D. K., Nichols, E., and Sudman, S. (1999), Understanding the questionnaire in business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 889–894.

- Willimack, D. K. and Nichols, E. (2001), Building an Alternative Response Process model for Business Survey. *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9.
- Willimack, D. and Nichols, E. (2010), Hybrid Response Process Model for Business Surveys. *Journal of Official Statistics* **26**, 3–24.
- Willis, G. B. (2004), Cognitive Interviewing Revisited: A Useful Technique, in Theory? In: Presser, S. et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, Wiley, New York, Chapter 2.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Editing During Data Collection
2. Data Collection – Main Module
3. Response – Response Burden

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

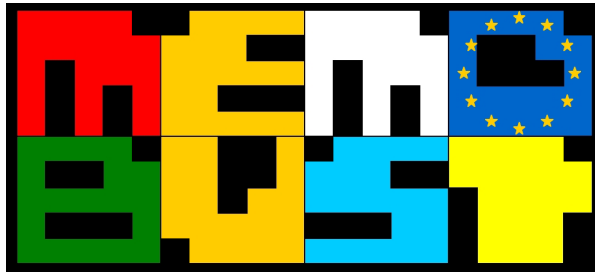
Response-T-Response Process

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	16-01-2012	first version	Paweł Lańduch	GUS
0.2	11-09-2012	second version	Paweł Lańduch	GUS
0.3	14-06-2013	third version	Paweł Lańduch	GUS
0.4	17-09-2013	fourth version	Paweł Lańduch	GUS
0.4.1	04-10-2013	preliminary release		
0.5	18-02-2014	minor revisions according to EB review	Paweł Lańduch	GUS
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:51



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Response Burden

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 The essence of response burden and its typology.....	3
2.2 Factors affecting response burden	8
2.3 Measurement of response burden.....	12
2.4 Reducing response burden.....	23
3. Design issues	28
4. Available software tools.....	28
5. Decision tree of methods.....	28
6. Glossary.....	28
7. References	28
Interconnections with other modules.....	32
Administrative section.....	33

General section¹

1. Summary

One of the main purposes of modern statistics is to ensure high quality data release necessary to satisfy expectations of their users and enable them to take effective political decisions. Statisticians struggle with this important problem mainly by seeking how to minimise response burden – one of the main barriers hampering the completion of this task. The burden can result both from the methodological design and survey management and from the respondent or technical support.

In this module we present the essence of response burden, analyse fundamental concepts related to this problem (with some original recommendations) and list the main types of difficulties which arise depending on the approach adopted. The importance and causes of the burden are discussed in detail. We also characterise the most important methods of measuring burden (both actual and perceived, also in the complex form) and their effects. In this context, we assess the efficiency of the burden reduction methods by referring to the assumptions of the Standard Cost Model. Practical examples of observed difficulties are presented. Basic and selected special methods to minimise these difficulties and international recommendations are also discussed.

2. General description

The problem of response burden is one of the main challenges facing modern statistics and a subject of interest to international organisations. It is among the key points in planning strategies of development of statistical methodology and improvement of data quality.

2.1 *The essence of response burden and its typology*

2.1.1 *The concept and awareness of response burden*

Response burden is a negative effect of the growing demand for data about the economic situation of businesses and – following this trend – a wide scope of detailed statistical surveys. Moreover, as noted by Jones (2012), these surveys should keep pace with quick and intensive economic changes. Therefore, several alternative ways of data collection are usually used (censuses or sample surveys, administrative data sources, electronic data interchange, published documents, etc.). So, in any way, businesses have to provide various data, which can generate additional burden and incur costs.

A good example of such burden recognition can be the EU Project on Baseline Measurement and Reduction of Administrative Costs (2010), which has provided credible estimates of administrative burden caused by 13 priority areas, as identified in the EU Action Programme to reduce administrative burden. The total administrative burden in the EU in years 2005 – 2007 is estimated at €102 billion².

¹The Authors thank Mrs. Katarzyna Maciejewska, Mrs. Agnieszka Kubasik, Mr. Adam Budziński and Mr. Andrzej Graf from the Statistical Office in Poznań (Poland) as well as Mrs. Deirdre Giesen (Central Bureau of Statistics, Netherlands), Mr. Magnar Lillegård (Statistics Norway), Mr. Johan Erikson (Statistics Sweden) and the anonymous reviewer of the Editorial Board for interesting comments and suggestions, which constituted a significant contribution to this module.

² Within this project an original measurement in representative samples of EU Member States was carried out and also the results of national administrative burden measurement efforts in a number of EU Member States were drawn upon. The results from these two groups of countries were then extrapolated to the EU as a whole. The baseline date for the measurement carried out by this project was July 2007. Reductions achieved between 2005 and 2007 are not taken into account in the measurement or burden reduction figures of this project.

Burden caused by statistics is estimated at €552 million. This is only 0.5 percent of the total burden. However, statistics is one of the three priority areas that cause the highest irritation.

It seems paradoxical that the relatively small burden imposed by statistics should cause so much irritation. However, statistical burden, unlike the burden caused by most other information obligations, is usually based on samples. Consequently, even though the total level of burden caused by statistics is low, the individual level of burden experienced by sampled businesses can still be relatively high. Moreover, statistical burden is unevenly distributed among businesses, i.e., typically the larger businesses are, the more surveys they get. Also, it has often been reported that respondents to business surveys often doubt the usefulness of statistical reporting requested of them (both to themselves and to society).

Of course, response burden can be unevenly distributed. That is, such burden is especially noticeable in the case of business surveys and afflicts mainly large firms, which are usually subjects of a number permanent and exhaustive surveys and obligations. On the other hand, the EU Project on Baseline Measurement and Reduction of Administrative Costs (2010) concludes that small companies suffer more from administrative burden than larger businesses (when administrative burden is expressed as the relative cost per employee or related to turnover). This is because of economies of scale (larger companies can invest in specialised staff and reporting systems).

Considering the effort required to satisfy the demand for data and relatively little time devoted to this task, which is given low priority in relation to the main activity of companies, reported data are likely to contain more gaps and errors, as companies become increasingly unwilling to cooperate. Sometimes, these gaps can also be the result of partial or total refusal to respond, which can be motivated by various circumstances (e.g., lack of necessary time or qualified staff, difficulties in finding or estimating required data, general reluctance, etc.). Some problems in this regard can also be linked to the way surveys are designed by methodologists and implemented by statisticians (e.g., proper collection of data from other sources). Thus, NSIs should also be concerned about response burden in their own self-interest, as it seems that excessive burden can cause problems with data quality (e.g., unit non-response) and affect the efficiency of data collection (e.g., the need to remind respondents, the need to re-contact respondents for editing, etc.). **All circumstances and factors negatively affecting the quality and cost of collecting statistical data directly from respondents or other external sources (e.g., administrative registers) are regarded as *response burden*.** It is the essence of the discussion presented, e.g., by Haraldsen et al. (2013).

In order to have a comprehensive understanding of response burden, we need to identify its causes, influencing factors, effects and be aware of possible threats and methods that can help reduce inconveniences for respondents, statisticians, analysts and data users, which result from poor data quality or the work of respondents. All these issues are discussed below.

2.1.2 Classification of response burden problems by type

The concept of response burden is far from straightforward. There are many classifications depending on the point of view on the nature of such burden adopted by a given researcher. Listed below are the most important ones.

As was suggested by Willeboordse (1998, pp. 113–114), the concept of response burden can be interpreted in various ways, which are usually presented as four dichotomies. The following contrasts can be considered:

- **objective vs. subjective burden** – objective response burden refers directly to the actual cost of completing questionnaires by respondents; subjective burden reflects their perception. Which of the two burdens is “heavier” depends to a large extent on the perceived usefulness of statistics resulting from respondents’ efforts. The distinction is in particular relevant when one compares the response burden of large and smaller businesses. While the latter carry a much larger objective burden, the former tend to be the heaviest complainers. Their subjective burden is higher, because they often do not make use of statistical data;
- **gross vs. net burden** – resulting from the quantification of response burden: net objective burden takes into account the “benefits” enjoyed by respondents for their contribution, gross response burden ignores them;
- **imposed vs. accepted burden** – imposed response burden assumes that all respondents sampled will fully and consciously complete the questionnaire with sufficiently accurate data; accepted response burden takes a more realistic approach: only responding businesses are accounted for, at the real completion cost.
- **maximalist vs. minimalist burden** – it is worth noting that completing a questionnaire often requires the respondent to look for and check other files, read the introductory letter and methodological hints, make necessary additional computations, etc. Thus, the actual completion time can be significantly shorter than the time needed to perform all related actions necessary for a proper completion of the questionnaire.

Taking into account the various concepts mentioned above, the following question arises: which choices should apply when monitoring response burden, either by estimates or by direct measurement. Although the general rule should be that “different concepts (apply) for different purposes”, in most circumstances the following choices from the three aforementioned alternatives will be preferred (cf. Willeboordse, 1998, p. 115):

- **objective burden** – subjective burden is in some respects more relevant (e.g., as a measure of acceptance and willingness to cooperate) but it is much more difficult to measure;
- **gross burden** – net burden would require the quantification of the value of data published, which is even more difficult. Moreover, this value would differ per respondent;
- **accepted burden**, since it is more realistic than imposed burden. Still, for internal NSI use, there is one disadvantage: because only responding businesses are taken into account, increasing non-response rates can have a positive effect on response burden. To avoid such undesirable “rewards” and, consequently, a less alert attitude towards declining response rates, survey managers should be confronted with burden figures, which include hypothetical non-response burden as well; hence, the term “accepted burden” can, in fact, denote the acceptable level of response burden;
- **maximalist burden**, as being much more realistic.

According to another approach represented and developed by Hedlin et al. (2005), the concept of response burden can be divided into **actual** and **perceived** burden. Actual burden can be reflected by 'hard' measures of duration of response preparation and costs. For example, we can consider the time taken to complete a survey, the number of tasks performed, the number of staff involved in the task or the costs to the business in terms of resources allocated to the survey. The concept of perceived burden was initially developed owing to an observation that traditional measurement does not take into account factors which may affect burden and which are rather subjective, such as the amount of effort required by the respondent and stress induced by sensitive questions. This dichotomy was conceptualised by the aforementioned handbook by Willeboordse (1998). That is, quantifiable actual burdens are regarded as objective and qualitative perceived burden can be defined as subjective.

It is a commonly observed fact that high response burden usually leads to significantly lower survey quality. Indeed, given many possible causes of excessive response burden, which will be presented later in this subsection, it can result in the following attitudes of respondents:

- refusal to participate in the survey; thus, no data from it will be available,
- refusal to provide some data (item non-response),
- provision of data of too low quality, e.g., presenting rough figures, errors in estimation or computation, etc.,
- in the case of similar surveys, some data can be mechanically copied from one questionnaire to another without special concern for their methodological correctness;
- deliberate provision of false data (an extreme situation).

Excessive response burden can also contribute to a growing level of incoherence and incomparability between some variables (e.g., concerning financial aspects), whose quality is especially sensitive to response burden (cf. Młodak, 2013).

The aforementioned problems are reflected in the quality of final survey results. The higher response burden is, the more effort should be made by the statistician to ensure acceptable quality of the published results of a survey. That is, the costs (financial and personal) of conducting imputation, estimation, using alternative data sources, etc. are higher. In extreme cases even high investment outlays in this respect may not produce expected effects. The reduction of response burden is, therefore, one of the key problems of modern official statistics.

Berglund et al. (2013) noted that there is a correlation between actual and perceived response burden. That is, businesses which complain that the questionnaire is burdensome actually use more time to collect the required information and to fill in the questionnaire than businesses which claim that the questionnaire is not difficult to complete. Moreover, actual and perceived burden seen together are also highly correlated with the number of corrected values in the questionnaire. It confirms our earlier observations.

Dale and Haraldsen (2007) show the necessity and usefulness of measurements of perceived burden for individual surveys. They note, however, that the quality of its recognition depends on the number of surveys directed to one respondent. They point out a significant difference between the way the issue of response quality is treated in studies of perceived response burden and in the Standard Cost Model (SCM). The SCM focuses on regulations concerning statistical financial costs of actions which

have to be taken by businesses to meet the requirements. In other words, SCM ignores perceived (subjective) burden. Moreover, SCM is generally based on a strategic collection of units, whereas Perceived Response Burden Study (PRB) uses a statistical sample. Hence, statistical calculations cannot explain many results obtained by SCM. On the other hand, SCM – although it is generally very expensive and time consuming – provides much more detailed and precise information, whereas PRB can be used to collect more representative information, which can be easily generalised and is simply less expensive. The SCM model will be presented in detail in subsection 2.3.2. Some approaches to the observation of perceived burden are described in subsection 2.3.1.

To complete the presentation of basic concepts related to response burden we should also mention the approach developed by Fisher and Kydoniefs (2001), who assume that burden is a combination of the following factors: **respondent burden** (factors connected with behavioural and attitudinal attributes of respondents, which affect the survey, e.g., belief in the usefulness of the survey), **design burden** (all aspects of the survey environment that are not directly associated with the respondent, e.g., incorrect sampling, frequency of contact, etc.) and **interaction burden** (a product of the relationship between respondent burden and design burden, e.g., requirement concerning memory and effort to be made, familiarity of the respondent with IT methods and tools, etc.). They argue that the perception of burden can be affected by these factors. So, the categories of actual and perceived burden can provide a good basis for a classification of response problems and an important factor in their quantification.

Haraldsen (2004) noted, however, that perceived burden is influenced by respondents' ability to answer, by the survey design and by the combination of these elements. Thus, the Fisher and Kydoniefs (2001) model does not distinguish between causes of burden and the perception of burden. In section 2.2. we will discuss a theory of causes of burden.

This subsection raises an obvious question: which system of classification of response burden can be recommended as an optimal solution for official statistics? As far as the authors of this module can tell, there are no formal documents specifying such recommendations at present. However, considering the connections between various attempts presented above, one can propose a compromise solution in this respect based on characteristic features of official statistics. What follows is our attempt at formulating such recommendations.

First, it should be remembered that response burden has two dimensions: quantitative (e.g., time and money spent) and qualitative (mainly perceived), depending on subjective opinions of respondents. We should also recognise at which stage of the survey design and implementation such burdens occur and what their nature is.

A good starting point for our recommendation will be a division of burdens into actual and perceived ones. Each of them can result from factors related to the respondent, design or interaction (for example, the difficulty of filling the questionnaire can be assessed both in terms of how much time it takes or by the respondent's subjective judgements of the level of difficulty – e.g., easy, rather easy, rather difficult, very difficult). Although this trichotomy, introduced by Fisher and Kydoniefs (2001), is applied mainly to perceived burden, it seems obvious that also these subcategories can be – in some circumstances – quantified. So, more universality is required. Within each of these subcategories one should make further distinctions depending on whether the burdens are quantifiable or not. Thus, within each subcategory one can distinguish gross burden, for the broad category of actual burden and actually observed burden (e.g., by a post-survey based on PRB questions – see Section 2.3.1) for the

broad category of perceived burden. Moreover, accepted and maximalist options can be applied for each category of the higher level of our classification (they cover both measurable and subjective factors). Finally, our classification can be visualised in the form presented in Figure 1. The category “observed burden” refers to the level of perceived burden observed as a result of relevant study, e.g., the PRB.

The presented proposal can help precisely systematise and present a variety of factors affecting response burden and design actions intended to reduce them.

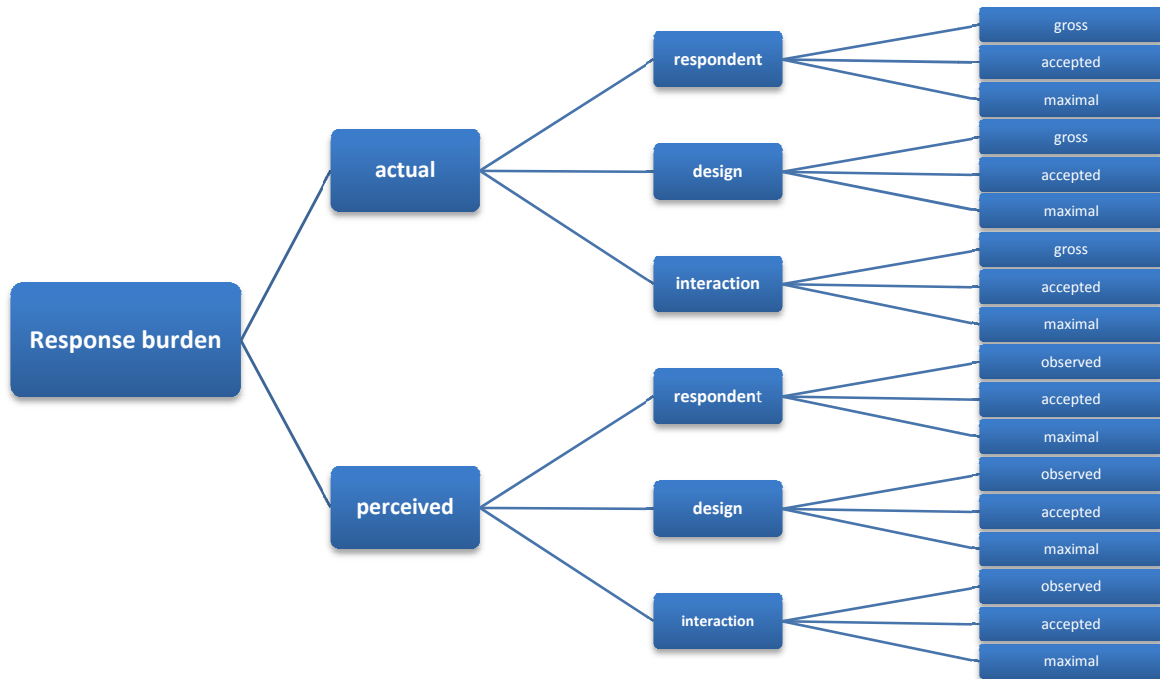


Figure 1. Suggested universal classification of response burden. (Source: Own elaboration.)

2.2 Factors affecting response burden

As we have mentioned earlier, the distinction between causes and the perception of burden is very important to understand the main problem connected with response burden. To overcome weaknesses described in Section 2.1., Haraldsen (2004) proposes replacing the idea of response burden with the term “*causes of response burden*”, i.e., which refers to what happens at the interface between the survey instrument and the respondent’s ability and willingness to respond. According to his idea, response burden and gratifications are regarded as the result of the encounter between the survey design and the respondent.

That is, perceived burden and gratifications are affected by **survey properties** and **respondent characteristics**. These factors coincide roughly with design and respondent burden according to the methodology introduced by Fisher and Kydonieffs (2001), but it is now assumed that they mutually interact. Moreover, according to Haraldsen (2004), the focus is not on the exact amount of response burden, but on whether the burden outweighs advantages and other positive aspects of the survey.

Survey properties are divided into *instrument features* and *data collection*. According to Haraldsen (2004) the former category includes, e.g.,

- the number of questions, determining the amount of information to be collected by the respondent,
- the questionnaire content, i.e., wording, requirements for information and response formats,
- the flow of questions and different elements within them, their logical ordering saves respondents' time, e.g., related questions are presented together and the respondent can use one data source to quickly complete all relevant items,
- the questionnaire layout: clear, logical, visually attractive structure and graphical arrangement of particular components of the form.

In this context Haraldsen (2004) pays special attention to the usefulness of computerised questionnaires discussing their advantages and drawbacks. These problems are presented in detail in the topic "Questionnaire Design" of this Handbook.

The data collection procedure consists of the following elements:

- the contact mode, including the type of contact form used, control of the respondent and response formats,
- the recruiting strategy, including the creation of incentives and motivation for respondents,
- administrative tasks performed before completing the questionnaire, during the process of completion responses and afterwards,
- security measures, i.e., tools and methods ensuring required confidentiality of individual responses of respondents.

As regards respondent characteristics, Haraldsen (2004) describes three main features of personality which determine the respondent's attitude to the survey. The first one is **interest in the topic of the survey**. If the topic is of no interest to the respondent, they will find no personal benefit in participating in the survey and, consequently, will either refuse to respond or provide only cursory answers. The second factor is **the competence of the respondent**. One should make sure that the respondent the survey is addressed to is fully competent to answer the questions properly. Otherwise, they can give "rounded" and "selective" answers instead of careful step-by-step processing. The last but not least feature is **availability**. That is, the quality of responses depends significantly on the amount of time and concentration the respondent is willing to devote to completing the questionnaire. In this context, one should take into account not only formal and technical possibilities of respondents, but also their personal features, such as patience, efficiency, etc. In the case of e-questionnaires, familiarity with Internet technology is also required.

Hedlin et al. (2005) discusses six main factors which are determinants of the level of response burden. They are as follows:

Survey organisation/sponsor is the first information which is taken into account by a potential respondent when they assess perceived response burden and decide whether to give a response. Usually, surveys conducted by agencies of official statistics or government (or self-government) inspire greater confidence than others and, hence, elicit higher response rates. The main reason for this is that these surveys are conducted according to special regulations (e.g., Official Statistics Act in Poland) which require interviewers and statisticians to respect rules of confidentiality and reliability of

collected data. Moreover, respondents feel that their contributions to such surveys will be used for the good of society (and hence also for their own good). In contrast, non-public statistical investigators, such as polling agencies, market research companies, individual scientists, students (e.g., who need relevant data for their diploma theses), etc. are formally not obliged to respect the norms of statistical ethic described in the legal acts (although they should also follow them to ensure high quality of collected and published information) and often make many errors while preparing and conducting the surveys. So, respondents often have no sense of security and usefulness of data which they provide in such situations. Even the anonymity of questionnaires is sometimes perceived negatively – as a method with a high risk of manipulating results. Respondents who express such concerns sometimes knowingly provide false responses and inform the interviewer about it. On the other hand, however, thanks to anonymity, respondents are inclined to be more truthful than when their answers are not anonymised. In general, from non-government surveys, academic ones usually enjoy higher response rates than, e.g., commercial ones. Hence, government support for surveys sent to businesses is desirable.

The second factor is **publicity**. That is, social attitudes to surveys can foster a better “climate” and “atmosphere” of motivation and willingness to respond. Based on their literature review Hedlin et al. (2005) show that significantly lower response rates can result from the specific character of a survey (e.g., addressing survey correspondence directly to specific persons rather than to respective enterprises or households, asking potentially difficult questions, etc.) or from political and economic conditions which contribute to a more reluctant participation in surveys. That is, the ‘public climate’ surrounding a large, repeated and well-known survey (e.g., the national census) may give rise to an atmosphere of motivation and willingness to respond rather than a specific, single survey. It is well-known that the respondent’s opinion about the usefulness, advantages and convenience of participating in a survey is determined by many factors, such as the current political situation, trust in institutions, economic conditions (especially the standard of living), etc. Loosveldt and Storms (2004) use the general term ‘survey-taking climate’ covering all circumstances affecting the attitude of respondents. Such an attitude has a great impact on the final quality of surveys and the usefulness of their results. In general, the negative attitude leads to an increase in the probability of refusals.

Loosveldt and Storms (2004) present their methods of assessing respondents’ attitudes based on a special drop-off questionnaire concerning respondents’ attitude towards a conducted survey and compare its results with the doorstep reaction of respondents (i.e., during direct contact). We will describe them broadly in Section 2.3.

One of the most important factors in this typology is the **implementation strategy**. It refers to a combination of factors, such as the initial contact and re-contacts with respondents, low cost of return of information and the clarity of the questionnaire and ease of its completion. Respondents also want to avoid double collection of data: providing the same information that has already been collected in another survey (possibly merely using a different structure of classification) is perceived as a waste of people’s time and effort and is often regarded as irritating. For example, a cover letter explaining the objectives and usefulness of the survey, sent prior to (or together with) the main survey questionnaire can persuade the respondent to participate; a kind reminder indicates the respondent’s importance for the interviewer. Also, the first direct contact of the respondent with the interviewer may affect the scope and quality of the received response. If the interviewer is nervous or awkward, the respondent may perceive the survey as very burdensome. Nowadays, when electronic means of communication

play a key role in contacting respondents, the initial contact, which should demonstrate special attention paid by the interviewer to the respondent and their responses, increases the sense of their importance and contributes to the growth of response quality. The last remark is also connected with *follow-up communication*, which should be undertaken to clearly appreciate the effort and expenses that have gone into gathering complete and high-quality statistical data. This presence of this stage usually increases the response rate. Measures aimed at *reducing the cost* of response for the respondent, e.g., the use of electronic transmission or pre-paid envelopes (in the case of paper-based surveys), contribute to a positive reaction to the survey. Another factor that matters is the *questionnaire appearance*. A questionnaire may be perceived as not very user-friendly if it's inconvenient (i.e., it is printed on a large piece of paper – in traditional surveys – or displayed in a small window or contains too small fonts – in the case of e-forms), graphically inconsistent (which leads to initial confusion), too complex (contains a row-column layout requiring additional effort to combine rows and columns), and if technical elements (i.e., marks and symbols used during processing) are too prominent and when instructions are too complicated. According to the old Roman adage “longus iter per praecepta, breve et efficax per exempla”³, it is better to replace, whenever possible, *long description in the instructions or notes* with clear examples. Especially, if additional or advisory information is presented on a separate card or incorporated in the question, respondents will avoid having to look back and forth through the questionnaire for the explanation, which would be strongly discouraging; any complication in this respect contributes to an increase in survey non-response.

The level of non-response may also be connected with the *questionnaire length*. This feature is usually negatively correlated with the level of response rate. Forms that are too lengthy can discourage the respondent from completing them, because this requires more effort. On the other hand, some respondents may actually appreciate the effort made by survey authors in preparing a comprehensive questionnaire and feel the importance of the survey. It is therefore necessary to find a healthy balance in this respect.

The content of the questionnaire, i.e., the *question comprehension* is also important. Asking troublesome or difficult questions can discourage the respondent – especially if they are not convinced of the usefulness of gathering such data or data security. This may be the case with financial or strategic data or questions containing many options (categories, rating scales, etc.). On the other hand, however, providing a greater number of possible answer options can actually decrease the perceived response (cognitive) burden by helping respondents to produce more informed answers.

The sixth major factor is the *mode of data collection*. A lot depends on respondents' preferences – increasingly more respondents prefer the more efficient methods of answering (online questionnaire, e-mail or automated phone, etc.) than traditional ones (such as paper forms). It is worth noting that response burden is proportional to the burden experienced by the respondent. In other words, the more complex the surveyed issue is and the more effort is required of a respondent to prepare an answer, the greater the resulting error (and burden). Two elements play an important role here: the interview method (paper, phone – CATI, e-questionnaire – CAII, personal – CAPI, etc.) and the questionnaire design. A good questionnaire design reflecting key connections within and between data and their validation is the factor leading to a significant reduction in response error, but it is often achieved at

³ The long road goes through advices, the short and efficient one – through examples (Latin).

the expense of a higher burden imposed on the respondent. Therefore, when designing a questionnaire it seems reasonable to follow the rule of the “golden mean”, finding a balance between the degree of necessary data verification and the level of questionnaire complexity. Paradoxically, a questionnaire that is too sophisticated and requires too much duplicate information may discourage a respondent and, therefore, negatively affect the completeness and quality of collected data. However, in surveys concerning sensitive issues (such as financial information, planned economic strategies, etc.), the use of face-to-face contact produces better results than an on-line questionnaire⁴. Of course, modes of data collection can vary, i.e., at various stages of the survey the mode can be changed. This strategy can be very useful, e.g., in a situation, when the respondent has given no response using the basic method. So, the researcher can try to obtain response by other means, such as phone, fax or traditional registered mail. Web-based data collection reduces the amount of paperwork and the cost of processing and improves timelines and quality of collected data. One should, however, take into account technical possibilities of the respondent and the extent to which e-questionnaires can be read using the respondent’s current IT tools. In some cases, the response may actually entail additional expenses for the purchase of equipment and software or even web access. Thus, traditional methods cannot be completely dropped.

One more area of difficulties concerning response burden is connected with discrepancies between the time when survey data are transferred to NSIs and rules applied in accounting systems and the timetable of wage and salary payments in different economic entities. For instance, in Poland data concerning the previous month must be submitted by the 5th business day. As a result, it is difficult to obtain data from accounting systems, where most recent transactions are not recorded because of delays in submitting invoices. This inconvenience significantly increases response burden.

The next problem concerns wages. Owing to certain regulations, some companies pay salaries by the 10th business day for work done in the previous month. Consequently, on the day of reporting, wages are not accounted for. The lack of required data is one of the most evident examples of response burden and forces companies to invest extra time and effort into preparing estimated data to fulfil the reporting obligation, which increases gross burden.

The lack of clear-cut and uniform definitions of concepts can also lead to a misunderstanding of ideas and force companies to contact statistical agencies conducting surveys to seek clarification of all doubts concerning ambiguous concepts.

Another Polish example of inconvenience concerns a very burdensome survey – enterprises employing over 49 people are obliged to submit monthly reports and the obligation automatically continues in the following year if the number of employees at the end of the previous year (on the last day of November) isn’t lower.

2.3 *Measurement of response burden*

In this subsection we will present the most important methods of observation and quantification of response burden. First, we will describe the main indicators enabling the assessment of actual and perceived burden. These burdens can be recognised on the basis of special surveys including both

⁴ For example, a student of one of the authors of this module, as part of her diploma thesis, has conducted a poll of strategies used by businesses in Kalisz (Poland) concerning employment of disabled persons. Despite a lot of effort made in preparing the online questionnaire no selected entity responded and thus she had to contact each of them face-to-face.

measurable quantities and subjective (i.e., categorical) observations. Next, the fundamental model for the assessment of actual burden, used within the European Statistical System, i.e., the Standard Cost Model, is characterised. Finally, we try to formulate a universal recommendation in this respect.

2.3.1 Indicators of response burden

The problem of measuring response burdens can be perceived as related to the observation of respondents' attitudes, costs and errors. The former one is much more difficult to accomplish owing to the subjective nature of this problem. Dale and Haraldsen (2007) analyse the methodology of PRB survey and suggest formulating two PRB core questions, which can be used to recognise whether respondents perceived the target survey as burdensome or not. If they did, they will need to answer another two questions specifying reasons and their perception. These answers will provide minimum knowledge about the perceived and actual burden, indicate where problems occur and how one can try to overcome them. Dale and Haraldsen point out that the actual burden is usually measured by the time necessary to fill the questionnaire and introduce two more questions concerning the time needed to collect required information and one to assess the time necessary just to fill the questionnaire. This approach takes into account the fact that some businesses could have multiple respondents and hence it provides a complete indication of the amount of time spent by the business (total) and by particular respondents. A complete collection of proposed questions is presented in Table 1.

Table 1. The PRB Core Question Set, for monitoring changes over time.

Dimension	Indicator	Question	Response categories
Perceived burden	Perception of time	Did you think it was quick or time consuming to collect the information to complete the questionnaire?	Very quick, Quite quick, Neither quick nor time consuming, Quite time consuming, Very time consuming
	Perception of burden	Did you find it easy or burdensome to fill in the questionnaire?	Very easy, Quite easy, Neither easy nor burdensome, Quite burdensome, Very burdensome
Actual burden	Time to collect information	How much time did you spend collecting the information to complete the questionnaire?	Number of hours, Number of minutes, Did not spend any time on this at all
		How much time do you think <u>the business</u> spent on collecting the information to complete the questionnaire?	Number of hours, Number of minutes, Did not spend any time on this at all
	Time to complete questionnaire	How much time did you spend on actually filling in the questionnaire?	Number of hours, Number of minutes

Dimension	Indicator	Question	Response categories
Perceived causes of burden	Reason for time consuming	What were the main reasons that you found it time consuming?	Had to collect information from different sources, Needed help from others in order to answer some of the questions, Had to wait for information that was available at different times, Other reasons, please specify
	Conditions for burden	What conditions contributed to making the questionnaire burdensome to fill in?	The high number of questions, Messy presentations made the questionnaire hard to read, Unclear terms and explanations of terms, Questions that asked for complicated or lengthy calculations, Available information did not match the information asked for, Difficult to decide which response alternative was the correct answer, Other reasons, please specify
Motivation	Usefulness for own business	Do you think that the statistics from this questionnaire are useful or useless to your business?	Very useful, Fairly useful, Neither useful nor useless, Fairly useless, Very useless, Don't know
	Usefulness for society	Do you think that the statistics from this questionnaire are useful or useless to society?	Very useful, Fairly useful, Neither useful nor useless, Fairly useless, Very useless, Don't know

Source: Dale and Haraldsen (2007).

It is worth noting that the question about conditions for burden provides a number of specific options. In contrast, the answers to the question about usefulness are very general and do not include any possible aspects of usefulness. It would, therefore, be useful to formulate a set of more informative answers in the future.

Dale and Haraldsen (2007) also describe a procedure focused on core questions that can be recorded in order to monitor how response burden changes over time. They present a more analytical approach that is designed to explain what causes response burdens, what effect these burdens have on the response quality and what can be done to reduce response burden. According to the study, there are

three key reasons why statistical organisations would want to carry out response burden surveys: to monitor perceived response burden over time, to evaluate changes that have been made to the questions and/or questionnaire and to evaluate changes that have been planned or made in the mode of data collection. In order to monitor perceived response burden over time, if there are no other changes to the survey, the core version of the PRB question set is recommended, otherwise (i.e., in the case of a mode switch, i.e., adding or removing several questions, changing several questions or redesigning the whole questionnaire) the authors propose a longer, analytical version of the aforementioned set. A PRB survey is also recommended before as well as after the changes. This will enable the institution which conducts the survey to measure the impact on perceived response burden. The document also provides examples of visual design for paper and web questionnaires. The model constructed by Hedlin et al. (2005) is used to identify a socio-psychological, causal model and to discuss how different components of this model could be measured and analysed. In addition, the authors present an overview of the sampling in a PRB.

Of course, if several people participate in providing information or completing a questionnaire, the situation is slightly more complicated. In this case, Haraldsen et al. (2013) suggest a stepwise variant of the survey: if the main respondent declared that other people provide assistance in preparing necessary data, the respondent is asked to specify the amount of time spent on pre-collection of relevant information, the number of supporters and the total amount of time they devoted to collecting data/completing questionnaire.

The simplest method of modelling of the respondent's decision whether or not to participate in the survey is the leverage-salient theorem (cf. Haraldsen et al., 2013). The theorem assumes that the respondent's attitude results from the interaction of several factors and their final balance. Thus, various survey aspects or participation arguments are visualised as hooking weights of different size on the leverage and the distance from the seesaw fulcrum to a given weight represents the importance of a relevant aspect to the respondent, while the size of the weight represents how salient this aspect is made.

Moreover, it is noteworthy to mention at this point a paper by Loosveldt and Storms (2004). Their special contribution (also mentioned briefly in section 2.2) is an original "General Attitude Towards Survey Scale" consisting of seven statements expressing attitudes about a survey. They are so universal that they can also be effectively used for purposes of business statistics. They are as follows:

- Surveys like this are a waste of time for people participating in it.
- By means of surveys like this one can express their opinion.
- Results of surveys like this are useful to make policy decisions.
- Surveys like this are an invasion of people's privacy.
- Everyone is obliged to cooperate with surveys like this.
- Results of surveys like this are mostly correct.
- With surveys like this the government gets a good picture of what's going on in the population.

Each of the statements can be evaluated using a five point response scale: 1 – completely agree, 2 – agree, 3 – neither agree nor disagree, 4 – disagree and 5 – completely disagree. To obtain a higher score for a more positive attitude, this scale was finally reversed. The individual respondent's score was computed as a mean of responses for particular questions; as a result, comparisons between participants and refusals are positive.

Of course, it is much more difficult to study attitudes of refusers. However, one can obtain at least partial information on their attitudes by re-contacting them (e.g., by CATI). It is very important in business statistics, where direct contact with respondents is especially intensive. Hence, respondents' reactions during direct contact are not affected by specific experiences of an interview and it is easy to register reactions of all types of respondents. In this case, information concerning negative reactions about time, interest, knowledge, privacy, research, etc. was collected. Loosveldt and Storms (2004) conclude that the measurement using the former manner will be less biased in a positive direction than in a face-to-face interview, although the latter one can provide more information about opinions of respondents and refusers.

Vorglimler, Bartsch and Spengler (2012) analyse the problem of administrative burden for businesses caused by statistical obligations in Germany and a solution leading to overcoming most difficulties in this field. For this purpose a special barometer of burdens has been developed. It is based on the Standard Cost Model and enables the measurement of statistical burden over time both with and without the influence of short-term economic effects. Of course, this barometer is also a good tool to observe effects of actions taken in order to reduce burdens.

Response burden can also be measured indirectly by relevant indicators of response. That is, the high level of non-response may suggest – if no other significant circumstances occur – discouragement of respondents to reply due to the observed (in previous rounds of the survey) or expected large effort required to collect relevant data. More precisely, response rates provide a general picture of the scale of observed problems (i.e., **unit response rate** – the ratio of the number of units for which data for some variables have been collected to the total number of units from which data are to be collected and **item response rate** – the ratio of the number of units which have provided data for a given data item to the total number of units from which data are to be collected or to the number of units that have provided information at least for some data items). Moreover, there are also other specific response rates, e.g., design-weighted response rates or size-weighted response rates. Remind that Hedlin et al. (2005) noted that the lower response rate in a given survey conducted on a relatively small population could be explained by the lack of 'census climate' during this study (the publicity factor – see Section 2.2). A discussion and recommendation concerning the complex assessment of the non-response problem can be found in the document by Eurostat (2009).

2.3.2 *Standard Cost Model*

The Standard Cost Model (SCM) provides a simplified, consistent way of estimating administrative costs imposed on businesses by regulations. The aim of this method is to reduce administrative burdens in the business environment by adopting a policy based on costs of regulations.

The advantage of this method is the possibility to measure burdens at different levels of the legal system – by analysing a single regulation or its segments, evaluating selected areas of legislation or performing a baseline measurement of all legislation in a given country. Another benefit is the

opportunity to assess existing regulations or results of new or amended laws, which came into force. Furthermore, the SCM approach is suitable for ex-post measurement of implemented regulations as well as for ex-ante examination of anticipated administrative burdens. Thanks to this approach, it is possible to assess the consequences of new regulations before their implementation.

Administrative cost (burdens)

Businesses have to comply with many administrative requirements and obligations imposed by law. Most of them are to do with the reporting obligation. We can consider statistical reporting as a kind of information obligation imposed on businesses to provide information and data on economic activity to the public sector.

Because of the increasing demand for new or more detailed information, it is very important to constantly make an effort to examine existing and future costs of surveys not to impose unnecessary burdens.

The SCM can also be applied to estimate the cost of statistical reporting incurred by businesses.

Components

SCM splits regulations into detailed components (cost and quantity parameters), which can be measured.

The cost parameters used in the SCM measurement include:

Time

Number of hours/minutes it takes a business to perform an activity.

Tariff

Internal cost (hourly pay for employees plus overhead and non-wage costs per hour).

External cost (hourly rate for external services, which perform administrative activities).

The Quantity parameters used in the SCM measurement include:

Population

This refers to the number of businesses to which the regulations apply.

Frequency

The number of times that a business delivers required data per year.

Acquisitions

In addition, certain necessary expenditure may be included, for example stationery or postage costs.

Structure

Represented by the following figure.

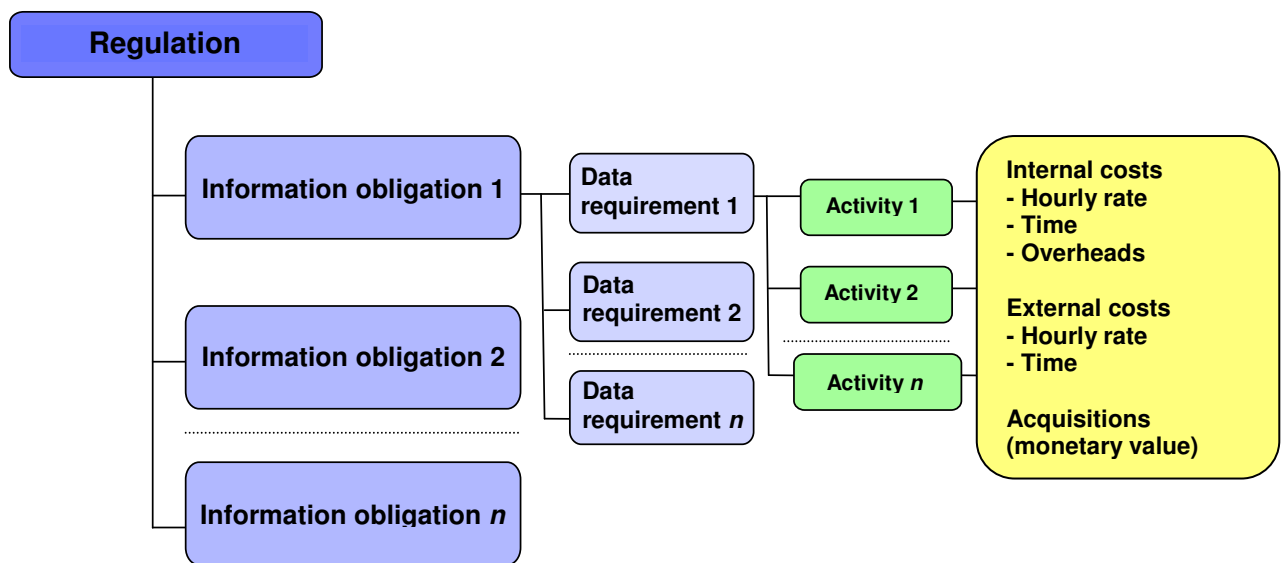


Figure 2. Structure of the Standard Cost Model. (Source: ISCM, 2003.)

How to get all this information

All components to measure administrative burdens can be obtained during interviews with a small (deliberately chosen according to relevant characteristics) number of businesses which are subject to a specific reporting regulation. Using the above parameters, we have to ask how much time and money they spend to perform each administrative activity that is required to fulfil a given information obligation. After collecting the data, we can perform the next step by standardising the amount of time and money spent on performing each activity within each segment of business and calculating costs incurred by businesses as a result of the imposed regulations.

Standard Cost Model – computational pattern

The cost parameters combined with quantity parameters enable us to estimate the total cost. The burdens are calculated by multiplying Price and Quantity.

$$\text{Price} = \text{Time} \times \text{Tariff}$$

$$\text{Quantity} = \text{Population} \times \text{Frequency}$$

Combining these elements give the basic SCM formula:

$$\text{Activity Cost} = \text{Price} \times \text{Quantity} = (\text{tariff} \times \text{time}) \times (\text{population} \times \text{frequency})$$

SCM example

For example, an administrative activity takes 3 hours to complete (time) and the hourly cost of one member of staff in the business completing it is £10 (tariff). The price is therefore $3 \times £10 = £30$. If this requirement applied to 100,000 businesses (population), each of which had to comply twice a year (frequency), the quantity would be 200,000. Hence the total cost of the activity would be $200,000 \times £30 = £6,000,000$.

Source: *Measuring Administrative Costs: UK Standard Cost Model Manual*

SCM implementation

The SCM was implemented in the United Kingdom (UK). According to the Prime Minister's Instructions on the Control of Statistical Survey, on behalf of UK government departments, the measurement of burdens was made by a consultancy company – Price Waterhouse Cooper (PWC). SCM involved examining, by a face-to-face questionnaire, a small group of businesses of varying type and size within the sample. The aim of the survey was to find out how much time and money businesses spend on each activity for each obligation imposed by law.

The method of collecting information on response burdens used by PWC posed a problem connected with ensuring adequate information concerning statistical surveys conducted by the Office for National Statistics (ONS). Hence, for purposes of ONS a paper questionnaire was developed with 17 questions and more detailed breakdowns, among others on types of activity and external costs. Some of them were required for the SCM calculations, but some referred to perceived burdens. The pilot SCM focused on nine surveys conducted during the period 2006 to 2009. On the basis of information gathered some values were calculated according to the following formulas:

Overall survey cost=(sum of weighted cost per questionnaire x survey frequency) × an uplift factor for re-contacting businesses

Mean cost per questionnaire=(sum of the weighted cost per questionnaire / survey sample size) × an uplift factor for re-contacting businesses

and therefore respondent burden cost (SCM formula) was estimated as:

Respondent Burden Cost=(weighted mean cost per questionnaire + uplift for re-contacting business) × number of questionnaires in survey sample × survey frequency

where:

cost per questionnaire=[(internal cost + overhead - adjustment for business-as-usual + external cost]

The adjustment for business-as-usual was employed when the information for a statistical survey was already held for the business's purposes, according to following rules: when "all information was already held, the adjustment was a 90% reduction; if some information was already held – a 40% reduction, and if none, no adjustment was made.

The weighted mean response burden per questionnaire was the average compliance cost of the business that corresponds to the review survey sample, where the review sample design is taken into account." (Frost et al., 2010)

As a result of using the SCM model some figures appeared representing the cost incurred by businesses to fulfil governmental obligations. Also some findings were formulated concerning this method of calculation.

First editions of the SCM showed that there was a necessity to redesign the questionnaire and to introduce a lot of changes. Despite the reduction in the number of main questions – from 17 to 10 – the paper method of gathering information was still debatable. It took respondents a long time to provide all required information concerning burdens on a statistical survey, even more than the participation in the survey which was assessed. Furthermore, some questions posed a significant problem to respondents. The main difficulty was to break down their time into parts corresponding to

different actions. The method also appeared to be impractical and ineffective for some respondents involved. For example, taking into account ‘business-as-usual adjustment’ resulted in some cases in non-realistic amounts of time devoted to completing the questionnaire.

Additionally, this method seemed to be very burdensome not only for businesses but also for data producers. The level of information gathered using SCM was significant and hugely resource-intensive – the process of calculating respondent burden required too much time and work. Consequently, the advantages – usefulness of information – were out of proportion to the effort put into its gathering and compiling.

Taking into account experiences from the pilot study of SCM, ONS leans towards a less resource intensive and simplified model. Variables which pose a problem in precise estimation and evaluation should not be taken into calculation to avoid measurement error. Also, implementation of the SCM approach should not saddle respondents with further burdensome requests for information. As a result, only two main pieces of data are necessary to implement SCM: time spent to complete the questionnaire and external costs resulting from the participation in statistical surveys. The adjustment of gathered data concerning the time taken to re-contact businesses to verify responses should be made by data producers. This way of proceeding should be more proportionate and robust for parties involved.

ONS also formulates an opinion that the approach to measuring burdens should focus on observing changes in burdens over time rather than measuring their actual level.

Summary of the SCM

The Standard Cost Model is a tool enabling us to work systematically towards reducing the response burdens for businesses by:

- creating awareness among statisticians about the level of response burden
- constantly monitoring response burdens
- setting out a strategy of reducing existing burdens
- minimising the response burdens in future undertakings (surveys)
- simulating ex-ante the burdens effects of new surveys in order not to impose unnecessary response burdens and design solutions where costs and benefits are more carefully balanced.

We should remember that SCM was developed to give only an indication of administrative burdens, and it is not intended to give detailed or exhaustive information.

So it is very important to compare the quantitative aspect to the qualitative one. Two dimensions of burdens – objective (concerning the actual cost) and perceived (concerning the willingness to cooperate) are the basis for an overall assessment of response burdens.

The implementation of SCM by ONS provides a very important lesson, which can be useful for statisticians within ESSnet. Firstly, a survey on response burden should not be the source of more burdens and obligations for survey participants and, secondly, it seems more important to monitor changing levels of burdens over time rather than calculate their actual costs.

Further information on administrative burdens is also available at the SCM Networks website⁵: www.administrative-burdens.com

2.3.3 *Recommendations for measuring response burden*

At the end of this subsection, we wish to formulate some guidelines about the use of the models of response burden measurement in the practice of official statistics. First of all, we will make a short overview of the existing literature presenting most important problems observed in NSIs in this context.

The models and solutions presented in previous subsections are elements of a general concept called Cost Benefit Analysis, introduced by Prest and Turvey (1965) and discussed by Haraldsen et al. (2013). It treats respondent burden as an effect of participating in a survey, which can generate both costs and benefits for users and institutions conducting the survey. That is, the cost of a survey is divided into respondent burden costs and survey organisation costs. The benefits can be viewed both in terms of user perception and as a change in quality. So, this model and measurement of all its parts can be recommended as a widely applicable solution that helps to perceive various aspects of response burden in a complex way – as part of a statistical survey strategy.

Rainer (2008) argues that the system used in most NSIs is highly desirable to document the burden caused and to monitor the effects of the efforts and measures taken in order to meet the reduction goals. The experience of statistical institutions in various EU member states shows that the actual response burden caused by official statistics is quite low compared to the total administrative burden. Thus, the real problem with response burden is that there is no strict correlation between a reduction in actual and perceived burden⁶. Rainer formulates some principles which could constitute conceptual guidelines for establishing a measurement instrument of the actual response burden at the EU-level; these guidelines are based on the currently applied practices, especially in Austria (the “response burden barometer” was mainly developed in cooperation with the Austrian Economic Chamber and the results have been published in an annual article in the bulletin of Statistics Austria and on the homepage of Statistics Austria since 2004). To avoid recall problems, they postulate performing response burden measurement right after the response action. Rainer suggests that the measurement should cover obligatory as well as voluntary data collection from businesses. He believes that voluntary reporting is treated by NSIs in the same way as obligatory reporting in terms of contact and reminder procedures; thus, since a specific survey might be obligatory in some member states while not obligatory in others, the voluntary factor of a survey seems to be necessary.

Giesen and Raymond-Blaess (2011a) provide the final deliverable of Work package 2 of BLUE-Enterprise and Trade Statistics (BLUE-ETS), which concerns the measurement and reduction of response burden at National Statistical Institutes (NSIs). It involved a survey of 45 NSIs from all European and some non-European NSIs. On the basis of this study one can observe that most NSIs do not seem to have a central place where knowledge of various response burden reduction actions and response burden measurement methods is coordinated. Giesen and Raymond-Blaess also note that

⁵ A booklet – *The Standard Cost Model – a framework for defining and quantifying administrative burdens for businesses* was published in August 2004. This manual contains a detailed description of the Standard Cost Model method and how to apply it.

⁶ Although – as we remember from section 2.1.2 – there is usually a correlation between actual and perceived burden (cf. also Berglund et al., 2013).

there is a large variation in the extent to which NSIs have implemented actions that can reduce response burden in their business surveys. It is difficult to conduct research on how actions aimed at response burden reduction actually affect three crucial aspects: response burden, data quality and the costs of producing statistics as well as how actions aimed at response burden reduction may have different effects for different businesses, depending on characteristics such as size class, industry or previous experiences with responding. According to Giesen and Raymond-Blaess (2011a), Eurostat should initiate the development and implementation of a standardised methodology for response burden measurement, research concerning business data collection methodology must move on from qualitative, explorative research to quantitative and preferably experimental research designs, effects of actions intended to reduce response burden should be monitored, reviewed, documented and published and burden reduction measurement and burden reduction actions should be coordinated within NSIs. Using data collected during this survey, Giesen and Raymond-Blaess (2011a) discuss problems connected with a systematic development of knowledge about efficient and effective methodologies for response burden reduction in business surveys. Continuing this matter, Giesen, Bavdaž and Haraldsen (2011) show that most NSIs conduct measurement of response burden using various methodological approaches but most of them have some kind of response burden measurement. In their opinion NSIs should move towards standardisation in order to provide good quality and comparable response burden data; they also discuss some issues that need to be solved in order to accomplish standardisation. These conclusions confirm the problems indicated by Rainer (2008).

Summarising, it is obvious that each NSI should have a central unit coordinating the measurement of response burden and equipped with the relevant knowledge to overcome difficulties. But the question remains how to construct the design of response burden measurement. It seems that an optimal solution is to do this after response collection is finished. Both actual and perceived burden should be quantified. For each group of burdens it should be indicated whether the survey is obligatory or voluntary. The actual burden measurement should be a combination of SCM and response indicators and quantities expressed in the relevant row of Table 1. Perceived (subjective) burden can be measured on the basis of a special survey with questions similar to those listed in the relevant cells of Table 1. It is also a good idea to include a question about the usefulness of the survey for the development of the country and regions, i.e., whether surveys of this kind help the central and local government to obtain a good picture of issues they are interested in.

In general, a complex measure of response burden of a given survey can be presented in the following form

$$\mu = \frac{\theta}{\Theta} + \sum_{i=1}^p \sum_{j=1}^{p_i} q_{ij} \varphi_{ij},$$

where θ is the cost of conducting a survey for businesses, obtained using the SCM model, Θ denotes the total cost of conducting this survey, q_{ij} is the value of j -th category in the i -th question concerning perceived burden (i.e., we assume that i -th question as p_i options of answer ordered from 0 to $p_i - 1$ in inverse relation to their burdensome character; for example answers to the question from Table 1: *Did you find it easy or burdensome to fill in the questionnaire?* have $p_i = 5$ and will be quantified as follows: 0 – very easy, 1 – quite easy, 2 – neither easy nor burdensome, 3 – quite burdensome, 4 – very burdensome) and φ_{ij} is the percentage of a given answer in surveyed businesses. The measure μ takes

values from $[0, \infty)$. Of course, the situation where $\mu = 0$ is impossible in practice (otherwise, e.g., all businesses would have no cost of filling the questionnaire – which is nonsense). The greater the measure, the higher response burden. This approach seems to be more efficient than the simple solution proposed by Haraldsen et al. (2013), who suggested assigning values to the response categories of PRB Questions (Table 1) according to a scheme, where a neutral answer receives the value 0 and burdensome ones are assigned negative values, e.g., -1 – very burdensome, -0.5 – quite burdensome, 0 – neither/nor option, 0.5 – easy or quick and 1 – very easy or quick and averaging the responses to the questions. This model can conceal difference between particular components and does not account for some important factors affecting response burden.

2.4 *Reducing response burden*

To minimise the problems concerning response burden as much as possible, NSIs should implement complex strategies involving a permanent overview of all business surveys and domains they cover, controlling data quality, recognition and reductions of threats, etc. To do it, efficient policies of National Statistical Institutes aimed at reducing response burden are necessary. In this section we will present a review of fundamental methods and forms of conducting such activities.

2.4.1 *Basic instruments and factors affecting reduction of response burden*

According to Willeboordse (1998), there are several instruments for carrying out such policies:

- **co-ordination, concentration or integration of data collection,**
- **rationalisation of the number of questionnaires and institutions where they should be reported;** optimally – one should try to construct universal solutions useful for all institutions involved – each of them could find data which it is interested in, one respondent should communicate only with one authority/department. Of course, there may be good reasons to deviate from this ideal approach. Still, even when respondents are tackled from different places in the organisation, contacts can be streamlined by appointing an account manager who is responsible for a harmonised approach to a particular (group of) respondents. Integration of questionnaires and clustering of surveys may not only reduce (the perception of) burden, but also contribute to the consistency of reported data and thus to the quality of statistics,
- **coordinated delimitation of sampling frames** (drawing samples for all such surveys from one unequivocal source, i.e., a centrally maintained business register; moreover, different surveys should apply the same type of statistical unit, as well as a uniform method and moment of determining their respective sampling frames from the business register),
- **coordinated sampling** (control of response burden achieved by a coordinated selection of samples). Without any internal coordination within the statistical agency it might happen that some businesses receive more forms than others, although these businesses are comparable in terms of size, activity, etc. A powerful tool to spread the response burden is a combination of a centrally maintained business register and a comprehensive computer program for coordinated sampling,
- **Electronic Data Interchange – EDI** (survey statisticians comply with accounting practices and also stimulate centralisation of data collection operations),

- **information on response burden** (NSIs should try to inform respondents in advance about surveys they will be involved in. Ideally, NSIs should send a comprehensive list of these surveys, including an average completion time of the questionnaire, at the beginning of each year. Of course, such a frank attitude is only possible with a very well planned and centrally organised surveying strategy, while all of the above mentioned issues should have been completed or at least be underway, e.g., the Database of Statistical Obligations in Poland),
- **policies applying at the level of individual surveys.** In this case the following aspects should be taken into account: *number of respondents* (using samples that are as small as possible and making a maximum use of auxiliary information. The use of advanced sampling techniques and high quality sampling frames as well as specialised databases and results of other surveys contributes to this goals), *units* (the observation unit should be defined in such a way that the respondent can recognise himself as a real transactor in the economy rather than an artificial construct; this can be accomplished by stressing the requirements of autonomy and data availability in operational unit definitions, while accepting a certain degree of heterogeneity), *concepts and definitions of variables* (questionnaires should be designed in such a way that they can be completed directly from book keeping records, and that it is, again, up to the statistician to bridge the gap between questionnaire concepts and statistical output concepts), *number and details of variables* (the contents of questionnaires should be alternated: once the “maximum” questionnaire is designed, one should seriously consider whether it is really necessary to apply it full size for each respondent during each reporting period), *accuracy of variables* (for smaller units the burden may be relieved by collecting data in ranges rather than discrete values, without a notable effect on the quality of statistical data), *tailor-made questionnaires* (when a survey covers distinct SIC-areas, accounting practices and vocabulary may differ among branches. This may require different questionnaires for different groups of respondents), *relevance of questions and explanatory notes* (if time and effort needed to read and understand questions, introductory letters and explanatory notes is excessive for respondents, it is recommended that questionnaires be tailored to homogeneous groups of respondents using, e.g., data from a previous survey), *feedback of results* (it is necessary to find out whether survey results provided to the respondent come up to their expectations and whether the effort put into preparing relevant data might have a positive effect on the perception of burden; if properly introduced, respondents may consider such a question as an indication that the NSI is aware of their problems and tries to do something about it; besides, outliers might be given after-care by advising them how to reduce the completion time).

Hedlin (2011) observes that the main factors affecting the total reduction in actual burden are as follows: **use of registers** (administrative databases can be a good source of a lot of information, which should eliminate the necessity of collecting it in surveys; it is commonly perceived as the first option to think of when reducing response burden), **the number of respondents** (to reduce sample size, either in every period of the survey or in some periods and using design-based methods of efficient sampling and estimation, e.g., a domain estimator is also recommended as the second option when the use of registers is impossible or insufficient), **time per question** (question text and questionnaire design-related response burden can potentially be reduced without any loss of exactitude), **the number of questions per questionnaire** (one should avoid using similar questions in the same or other surveys), **the range of questions in different survey rounds** (rather than collecting the full data set in

every period of the survey, some questions in some periods can simply be skipped, also time series analysis may be useful to impute some data), **the frequency of the questionnaire** (reduce frequency of questionnaires, for example from four times to three times a year in a repeated survey), **general survey tasks** (opening the envelope, logging onto the website, retrieving a web questionnaire, storing responses in an archive, communicating the response, etc.), **the frequency of re-contacts for the same questionnaire** (minimise re-contacts for editing and follow-up purposes), **spreading actual response burden out** (a more even distribution of response burden over businesses is highly desirable even if the total burden remains the same; one can spread questionnaire requests evenly over the population or spread questions evenly by dividing up items in a survey in question sets and not putting more than one question set to any one business; both approaches can be combined).

Hedlin (2011) focuses on how to reduce actual response burden by means of sampling and estimation. There are, in principle, two main data sources in surveys: data that the survey organisation collects and data that are collected by another organisation for purposes other than the survey. An obvious way to reduce actual response burden would be to cut down on the information output and, hence, the need for data input from respondents. Whether this is feasible or not is a pertinent question to ask; however, we focus on burden reduction measures that largely maintain information output. An overview of survey results for survey design actions that can reduce actual burden (based on the survey of NSIs described by Giesen and Raymond-Blaess, 2011b) shows that the use of administrative or register data, reduction in sample size and reduction in the number of items are particularly common measures implemented by NSIs. However, burden reduction actions may also reduce the quality of survey estimates. For example, replacing a survey with register-based statistics may lead to a loss in validity. Sometimes it is possible to estimate the size of the loss by running the survey while simultaneously producing register-based statistics.

To be closer to current practice and to account for problems mentioned in sections 2.1. to 2.3, a strategy to reduce response burden should contain the following actions:

- changing deadlines for submitting statistical reports to reduce response burden and to improve survey quality and completeness – adjusting them to book keeping regulations.
- obtaining data from administrative registers, where data are submitted by companies because of reporting obligations imposed by law. Such registers are maintained by government institutions and contain data about social insurance (employment data) or revenues (income and tax data).
- using other (administrative) sources to ensure current information on enterprises to update business registers (phone numbers, e-mail addresses, postal addresses)
- creating regulations to support statisticians in their work with enterprises that consistently refuse to report information.
- simplifying, providing explanations to reports, variables and concepts. The more complex such explanations are, the higher the rate of incorrect data.
- adjusting survey assumptions to the requirements of accounting systems and regulations to enable the transfer of data directly from accounting systems.

2.4.2 Some practical solutions concerning response burden policy

We will now present several examples of special strategies within the response burden policy concerning the recognition and reduction of burdens applied in various countries.

Bolin and Thyrestrand (2011) describe tasks carried out by the Survey Help Desk in Statistics Sweden, which monitors response burden and tries to help heavily burdened enterprises, that is enterprises that are in a particularly difficult situation, with many surveys and limited possibilities to respond to them. This organisational unit researches the situation of such enterprises and tries to find ways to ease the burden for them in accordance with the scope of non-response and sample scheme. To support this undertaking, the Survey Help Desk relies on the Swedish Register of Data Providers. The purpose of this database is to measure and analyse the burden at an aggregated level and to be able to give information to each individual enterprise about the surveys they are participating in.

Goddeeris and Bruynooghe (2011) provide an overview of the simplification process and its results and the use of the XBRL based web survey used in many countries. XBRL (eXtensible Business Reporting Language) is an open standard based on the electronic collection and transfer of business economic data via the internet. The use of the XBRL technology has made it possible to develop programs that automatically search all the data for Structural Business Statistics in the accountancy data of the enterprise and organise these data in an XBRL file that can be uploaded (see the topic “Data Collection” of this Handbook and <http://www.xbrl.org/>).

Yancheva and Iskrova (2011) present objectives, assumptions and results of the Bulgarian project aimed at reducing administrative burden in business statistics in that country. It is conducted on the basis of the Information System of “Business Statistics” (ISBS), which provides an online collection of annual reports of all economically active enterprises, containing a set of accounting and statistical questionnaires. The key result of the project was the implementation of the single entry point for reporting fiscal and statistical information, which involved defining the scope and content of data that have to be submitted, ensuring that definitions and concepts used in the reports are identical for institutions which collect business data (in Bulgaria there are two), introducing amendments in the legal acts related to fiscal and statistical obligations of business, developing the concept of SBS data warehouse to ensure the common use of data that fits the specific purposes of each institution, creating the Information System of ‘Business Statistics, developing and launching a public awareness campaign and training sessions for accountants and business associations and, finally, promoting electronic data submission instead of paper based one. These tasks were performed by specialised experts from statistical and financial institutions.

Oswald and Stanton (2011), on the basis of experiences of the United States of America, suggest reducing instruction/explanatory materials and item redundancy, distributing subsets of items strategically across units using available data or imputation to complete analyses and automating field completion by means of relevant optimisation techniques, which are based on the dependencies between data and restrictions imposed on them.

Finally, we would like to mention some solutions adopted in Poland. To ensure effective knowledge and control of survey implementation one should have a central database indicating for each economic entity which surveys it is actually involved in. In Poland this information is stored in the Database of Statistical Obligations, containing a list of reports that each statistical unit should submit to the Central Statistical Office. The timeliness of fulfilling these obligations is also systematically monitored.

2.4.3 International recommendations for response burden policy

When describing the problem of reducing response burden, we should also present most important international recommendations which will be useful for statisticians and users of statistical data. A starting point in this context can be the document prepared by Eurostat and ESS (2011), Principle 9 concerning non-excessive burden on respondents was formulated. The document states that *“the reporting burden should be proportionate to the needs of the users and should not be excessive for respondents. The statistical authority monitors the response burden and sets targets for its reduction over time”*. This principle states that *“the range and detail of European statistics demands is limited to what is absolutely necessary”*. That is, the reporting burden should be spread as widely as possible over survey populations by applying appropriate sampling techniques. To collect information from businesses, their accounts and electronic means should be used where possible. This should improve data transfer and their quality. If exact figures are not readily available, best estimates and approximations are accepted. The document also underlies the role of administrative data sources, which can be used to avoid duplicating requests for information and keep the number of surveys to a minimum. Thus, data sharing within statistical authorities has to be harmonised and generalised.

Eurostat (2009) states that the procedures of treating respondent burden should include among others: assessment of annual respondent burden in financial terms and/or hours, the definition of respondent burden reduction targets, recent efforts made to reduce respondent burden, answers to questions of whether the range and detail of data collected by surveys is limited to what is absolutely necessary, whether administrative and other survey sources are used to the fullest extent possible, whether electronic means are used to facilitate data collection, whether best estimates and approximations are accepted when exact details are not readily available and whether reporting burden on individual respondents is limited to the extent possible by minimising the overlap with other surveys. Also, one should consider the scope of data collected from businesses – it should be verified whether such data are readily available from their accounts. These elements are necessary components of any efficient and comprehensive report on response burden.

In the document by Eurostat (2009), it was pointed out that the difference between costs on the one hand and benefits in terms of output data quality on the other should also include respondent participation understood as a cost (to respondents) that has to be balanced against the benefits of the data thus provided.

On the basis of his statements described in subsection 2.4.1 of this module, Hedlin (2011) formulates the following main recommendations for internationally consistent policy aimed at burden reduction:

“1. Eurostat should initiate the development and implementation of a standardised methodology for the measurement of response burden caused by official business surveys. The standardised methodology may include multiple indicators and a minimum version of the measurement, to accommodate NSIs differences regarding the purposes of and resources for response burden measurement. Standards are needed to ensure that basic comparisons can be made over time and between NSIs. To make informed decisions on the minimal requirements for standardised response burden measurement research is needed that assesses to which extent different aspects of response burden are relevant for the quality and costs of data collection.

2. Research concerning business data collection methodology must move on from qualitative, explorative research to quantitative and preferably experimental research designs. Research into data

collection methodology for business surveys is relatively young and has so far been mainly qualitative. These studies have provided many valuable insights in how data collection design characteristics that are under control of the survey organisation can affect response burden, data quality and the costs of data collection. However, quantitative studies that provide information about the significance and the magnitude of these effects are lacking. This information is essential for NSIs to efficiently plan their resources and optimize their data collection.

3. Effects of actions intended to reduce response burden should be monitored, reviewed, documented and published. When NSIs plan actions to improve their data collection, they should include a plan to make statistically sound comparisons between alternative (or old and new) methodologies. As these kinds of studies are very scarce and most NSIs face similar challenges, effort should be put in making the results of these studies known to the international community of statistical agencies and survey methodologists.

4. Burden reduction measurement and burden reduction actions should be coordinated within NSIs. Within NSIs the knowledge on response burden measurement and response burden reduction actions seems to be rather fragmented and scattered. Statistics Canada and Statistics New Zealand are examples of what seem to be the current best practices concerning the organisation of response burden measurement and response burden reduction at NSIs. Both agencies have dedicated staff, an Ombudsman and a Respondents Advocate respectively, to coordinate the response burden work.”

Thus, careful and permanent monitoring and treatment of the response burden and taking efficient actions aimed at reduction of them is one of key tasks of the modern statistics.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Berglund, F., Haraldsen, G., and Kleven, Ø. (2013), Causes and Consequences of Actual and Perceived Response Burden Based on Norwegian Data. In: Giesen, D., Bavdaž, M., and Bolkopp, I. (eds.), *Comparative report on integration of case study results related to reduction of response burden and motivation of businesses for accurate reporting*, BLUE-Enterprise and Trade Statistics, BLUE-ETS, European Commission, European Research Area, 7th Framework Programme, 26–32.

- Bolin, E. and Thyrestrand, S. (2011), Approaches to increase motivation and help the most burdened enterprises. In: Giesen, D. and Raymond-Blaess, V. (eds.), *Response burden measurement and reduction in official business statistics. A literature review of national statistical institutes' practices and experiences*, Deliverable 2.1 BLUE-Enterprise and Trade Statistics, Statistics Netherlands, Heerlen, 161–166.
- Dale, T. and Haraldsen, G. (eds.) (2007), *Handbook for Monitoring and Evaluating Business Survey Response Burdens*. European Commission, Eurostat.
- DETI (2009), *Annual Report on Statistical Surveys to Businesses – Compliance and Quality Improvement Plan*. Department of Enterprise, Trade and Investment, UK government, December 2009, http://www.detini.gov.uk/deti_2008_report_to_ministers-2.pdf.
- Eurostat and ESS (2011), *European Statistics Code of Practice for the national and community statistical authorities*.
- Eurostat (2009), ESS Standard for Quality Reports. Eurostat Methodologies and Working Papers, Office for Official Publications of the European Communities, Luxembourg.
- Fisher, S. and Kydonieffs, C. (2001), Using a Theoretical Model of Response Burden (RB) to Identify Sources of Burden in Surveys. Paper presented at the 12th International Workshop on Household Survey Non-response, Oslo, Norway, September 12 – 14.
- Frost, J.-M., Green, S., Jones, J., and Williams, D. (2010), *Measuring Respondent Burden to Statistical Surveys*. Office for National Statistics – Methodology Directorate.
- Giesen, D., Bavdaž, M., and Haraldsen, G. (2011), Response Burden Measurement: Current Diversity and Proposal for Moving towards Standardization. In: Giesen, D. and Raymond-Blaess, V. (eds.), *Response burden measurement and reduction in official business statistics. A literature review of national statistical institutes' practices and experiences*, Deliverable 2.1 BLUE-Enterprise and Trade Statistics, Statistics Netherlands, Heerlen, 125–134.
- Giesen, D. and Raymond-Blaess, V. (2011a), National Statistical Institutes' response burden reduction measures: first survey results. In: Giesen, D. and Raymond-Blaess, V. (eds.), *Response burden measurement and reduction in official business statistics. A literature review of national statistical institutes' practices and experiences*, Deliverable 2.1 BLUE-Enterprise and Trade Statistics, Statistics Netherlands, Heerlen, 139–149.
- Giesen, D. and Raymond-Blaess, V. (2011b), Overview of research design and results. In: Giesen, D. and Raymond-Blaess, V. (eds.), *Response burden measurement and reduction in official business statistics. A literature review of national statistical institutes' practices and experiences*, Deliverable 2.1 BLUE-Enterprise and Trade Statistics, Statistics Netherlands, Heerlen, 9–14.
- Goddeeris, O. and Bruynooghe, K. (2011), Administrative Simplification of the Structural Business. In: Giesen, D. and Raymond-Blaess, V. (eds.), *Response burden measurement and reduction in official business statistics. A literature review of national statistical institutes' practices and experiences*, Deliverable 2.1 BLUE-Enterprise and Trade Statistics, Statistics Netherlands, Heerlen, 167–176.

- Haraldsen, G. (2004), Identifying and Reducing Response Burdens in Internet Business Surveys. *Journal of Official Statistics* **20**, 393–410.
- Haraldsen, G., Jones, J., Giesen, D., and Zhang, L.-C. (2013), Understanding and Coping with Response Burden. In: Snijders, G., Haraldsen, G., Jones, J., and Willimack, D. K. (eds.), *Designing and Conducting Business Surveys*, Wiley Series in Survey Methodology, John Wiley & Sons Inc., Hoboken, New Jersey.
- Hedlin, D. (2011), Reducing actual response burden by survey design. In: Giesen, D. (ed.), *Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes*, Deliverable 2.2 BLUE-Enterprise and Trade Statistics, 25–32.
- Hedlin, D., Dale, T., Haraldsen, G., and Jones, J. (eds.) (2005), *Developing Methods for Assessing Perceived Response Burden*. Research report, Statistics Sweden, Stockholm, Statistics Norway, Oslo, and Office for National Statistics, London.
- ISCM (2003), *International Standard Cost Model Manual, Measuring and reducing administrative burdens for businesses*. SCM Network to reduce administrative burdens.
http://www.administrative-burdens.com/filesystem/2005/11/international_scm_manual_final_178.doc
- Jones, J. (2012), Response Burden: Introductory Overview Lecture. Fourth International Conference on Establishment Surveys, Survey Methods for Businesses Farms and Institutions, June 11th – 14th, 2012, Montreal, Canada (<http://www.amstat.org/meetings/ices/2012/papers/302289.pdf>).
- Loosveldt, G. and Storms, V. (2004), Measuring Respondent's Attitude Towards Survey. 6th International conference on social science methodology, Amsterdam, 17th – 20th August, 2004, ed. by the Centre for Sociological Research, Katholieke Universiteit Leuven, Belgium. Available at http://konference.fdvinfo.net/rc33/2004/Data/PDF/stream_08-11.pdf.
- Młodak, A. (2013), Coherence and comparability as criteria of quality assessment in business statistics. *Statistics in Transition – new series* **14**, 287–318.
- Oswald, F. and Stanton, J. (2011), *Reducing Response Burden*. Rice University and Syracuse University, U.S.A. (<http://www.slideshare.net/jmstanto/reducing-response-burden>)
- Prest, A. R. and Turvey, R. (1965), Cost – benefit analysis: A survey. *The Economic Journal* **75**, 683–735.
- Rainer, N. (2008), Measuring response burden under EU-context: Some principles for a management tool at the EU-level. Paper presented at the 94th Directors-General of the National Statistical Institutes (DGINS) Conference, Vilnius, September 25-26, 2008.
- Vorgrimler, D., Bartsch, G., and Spengler, F. (2012), Measuring cost efficiency and response burden in statistical surveys. Federal Statistical Office of Germany (Destatis), European Conference of Quality in Official Statistics, 29th May – 1st June 2012, Athens, Greece, available at http://www.q2012.gr/articlefiles/sessions/35.1_Vorgrimler_response_burden_Q_2012.pdf.
- Willeboordse, A. (ed.) (1998), *Handbook on the Design and Implementation of Business Surveys*. Eurostat, Luxembourg.

Yancheva, D. and Iskrova, K. (2011), Reducing the administrative burden for the business in Bulgaria: Single Entry Point for Reporting Fiscal and Statistical Information. In: *Proceedings from BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, Statistics Netherlands, Heerlen, March 22 & 23, 2011, 189–198.

Interconnections with other modules

8. Related themes described in other modules

1. User Needs – Specification of User Needs for Business Statistics
2. Overall Design – Overall Design
3. Questionnaire Design – Main Module
4. Sample Selection – Main Module
5. Data Collection – Main Module
6. Data Collection – Collection and Use of Secondary Data
7. Imputation – Main Module
8. Weighting and Estimation – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

1. Cost computation
2. Effective sample selection algorithms
3. Imputation algorithms
4. Weighting algorithms

11. GSBPM phases explicitly referred to in this module

1. GSBPM Phases 4.1 and 5.2–5.6

12. Tools explicitly referred to in this module

1. CAII
2. CAPI
3. CATI
4. EDI
5. Reporting portals, e-questionnaires

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

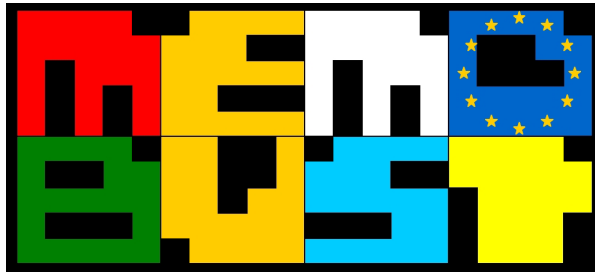
Response-T-Response Burden

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	6-6-2011	first version	Monika Natkowska Andrzej Młodak	GUS (PL)
0.2	31-08-2012	revised version	Monika Natkowska Andrzej Młodak	GUS (PL)
0.3	14-02-2013	third version	Monika Natkowska Andrzej Młodak	GUS (PL)
0.4	26-04-2013	fourth version	Monika Natkowska Andrzej Młodak	GUS (PL)
0.5	07-11-2013	fifth version	Monika Natkowska Andrzej Młodak	GUS (PL)
0.5.5	27-01-2014	corrected version according to EB-review	Monika Natkowska Andrzej Młodak	GUS (PL)
0.5.6	28-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:52



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Data Fusion at Micro Level

Contents

General section	3
1. Summary	3
2. General description.....	3
2.1 Data fusion	3
3. Design issues	9
4. Available software tools	9
5. Decision tree of methods	9
6. Glossary	9
7. References	9
Interconnections with other modules.....	11
Administrative section.....	12

General section

1. Summary

This module gives a general overview of problems and methods concerning the integration of several data sources for statistical purposes. It is focused on the case of integration at micro level, which means to integrate data sources composed of units (input: micro) in order to obtain still a data set composed of units (output: micro).

2. General description

There are more statistical data produced in today's society than ever before. These data are analysed and cross-referenced for innumerable reasons. In the case of National Statistical Institutes (NSIs) the joint analysis of two or more statistical and administrative sources is a result of a rational organisation of all available informative sources and, among all, it allows the reduction of survey costs, the response burden and to enrich the information already held on such units by means of adding new data from other sources enabling for instance the analysis of relationships among variables observed in different data sources. Nevertheless, the integration process must deal with many different problems.

This module gives a general overview of problems and methods concerning the integration of several data sources for statistical purposes. Integration is made at micro level, which means the integration of data sources composed of units (input: micro) with the aim of obtaining still a data set composed of units (output: micro).

The problems discussed in this section essentially deal with two questions: how to fuse different data sources and how to manage consistency problems.

The section is mainly based on the documents produced in the two European projects funded by Eurostat on data integration: the ESSnet *Integration of Surveys and Administrative Data* carried out during 2006-2007 (ISAD, 2006), and the Essnet on *Data Integration* during 2010-2011 (DI, 2011).

2.1 Data fusion

An important element to take into account when different data sets are fused concerns if they are composed of

- 1) (almost) the same units;
- 2) different units.

The first case is typical of integration between registers and sample surveys, while the second typically happens when integration is related to sample surveys.

This distinction is important since different methods are required. Essentially, in the first case we resort to statistical classification methods and in this context they are referred to as record linkage procedures, while in the second we mainly resort to imputation methods that are usually referred to as statistical matching techniques.

2.1.1 Record linkage/Object Matching

Record linkage (also known as *object matching*) consists in identifying pairs of records coming from different data sets, which belong to the same entity, on the basis of the agreement between common

variables (name, address, telephone,...). It may happen that the same units have different values for the common variables in the two data sets, for instance because of a change in the telephone number or because some error affects data (Herzog et al. 2007).

In general, key variables are compared in order to understand whether a pair of observations from the two files is either a match or an unmatched.

The results of the comparisons may be used in different ways resulting in different record linkage/object matching methods that can be classified as:

- 1) deterministic approaches;
- 2) probabilistic approaches.

Deterministic approaches are characterised by the use of formal decision rules. In this framework, some algorithms are developed for linking data, for details see the modules “Micro-Fusion – Unweighted Matching of Object Characteristics” and “Micro-Fusion – Weighted Matching of Object Characteristics”.

Probabilistic approaches make an explicit use of probabilities for deciding when a given pair of records is actually a match given the results of the comparison of the key variables. The probabilities allow to quantify the degree of uncertainty in a match/unmatched pair of observations and may help the researcher to take decisions within a formalised probabilistic setting allowing in some cases the estimation of errors associated to the performed action.

The procedure proposed by Fellegi and Sunter (1969) is one of the reference techniques for probabilistic record linkage. They deal with the problem of linkage by using a latent model, where the latent variable describes the two populations of matches and unmatcheds. For each pair of observations, the probabilities of belonging to the two populations are computed according to the values obtained by the comparison of the key variables. Intermediate situations where the pairs cannot be classified with a high probability in one of the two populations may arise. For these units a clerical review is needed. The advantage of using a probabilistic approach is that it is allowed to estimate the errors associated to the decision taken in the classification step, and they can be used to establish a proper methodological framework for a statistical procedure to decide links, or can be used when assessing the quality of estimates obtained with integrated data that are affected by a further source of uncertainty due to the linkage process. The latter is a problem that still requires further investigations, for details see Di Consiglio and Tuoto (2013) and references therein.

The probability distributions of the results of the comparisons of the key variables for respectively match and nonmatch populations are essential elements of the probabilistic record linkage approach. They are generally not known and an estimation step is needed.

It is worthwhile to remark a peculiarity of the linkage procedures as previously formalised. The number of matches is naturally much less than the number of unmatcheds. Without loss of generality, let n_A be the number of observations in the first file A and n_B the number of observations in the second file B to be linked, where $n_A \leq n_B$. In the most favorable situation there are n_A matches out of the $n_A \times n_B$ pairs of units. A very low proportion of matches with respect to all the pairs of units may result in a not reliable estimation result because the unmatcheds tend to overwhelm the information coming from the rare population of matches. This problem is alleviated by using blocking procedures, that is to split records into groups (blocks) and comparing only the units belonging to the same group. When data

sets are large, this task is also particularly important from an operational point of view since it can be computationally unfeasible to make a large number of comparisons. On the other hand restricting the matches within the block may be dangerous because of the exclusion of some possible matches. Suggestions for choosing the blocking procedures can be found in ISAD (2006).

For more details on probabilistic record linkage and the Fellegi-Sunter procedure see the modules “Micro-Fusion – Probabilistic Record Linkage” and “Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage”.

2.1.2 Statistical matching

Statistical matching (also named *data fusion* or *synthetic matching*) refers to a series of methods whose objective is the integration of two (or more) data sources referring to the same target population. The data sources are characterised by the fact they all share a subset of variables (common variables) and, at the same time, each source observes distinctly other subsets of variables. Moreover, there is a negligible chance that data in different sources observe the same units (disjoint sets of units).

In the simplest case of two samples (data sources A and B), the classical statistical matching framework is represented in Figure 1:

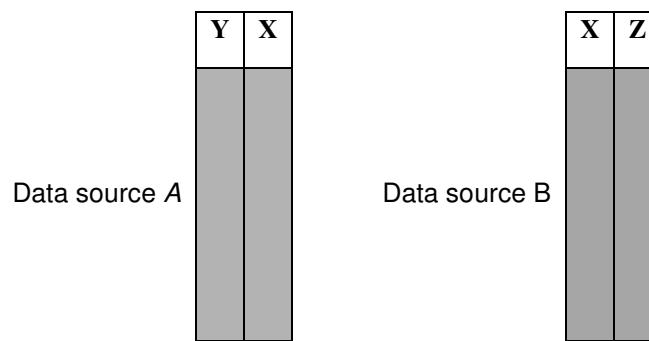


Figure 1. The statistical matching setting

The common variables are denoted by X , the set of variables Y are observed only in A but not in B, and Z are observed in B but not in A (Y and Z are not jointly observed).

Statistical matching methods aim at integrating the two sources in order to study the relationship existing among the two sets of variables not jointly observed, i.e., Y and Z or, more in general, to study how X , Y and Z are related.

In the *micro approach*, the statistical matching objective is the construction of a complete “synthetic” file, that is a file where X , Y and Z are jointly present. The term synthetic refers to the fact that this file is not the result of a direct observation of all the variables on a set of units belonging to the population of interest, but it is obtained exploiting information in the observed distinct files. For example, in the case of the data sets as in the previous figure, a synthetic file is the one in Figure 2, where the file A is filled in with the synthetic values \tilde{Z} by exploiting the joint information regarding X and Z observed in B.

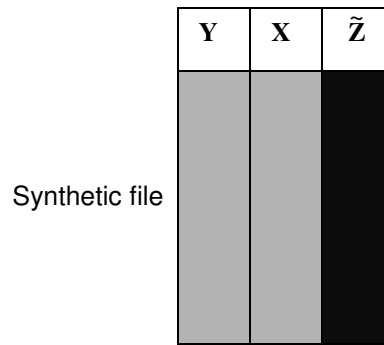


Figure 2. Statistical matching at micro level

The file is generally obtained through imputation techniques that may resort to parametric models, nonparametric models, and a mixture of them. For parametric imputations an important role is assigned to regression models, in this case the regression of Z (dependent variable) vs the covariate X is estimated on B . The prediction \tilde{Z} is obtained by applying the estimated model to the X values in A . The most frequently used nonparametric models refer to the family of hot-deck imputation methods. By considering the situation in Figure 2 the value \tilde{Z} for a given unit is obtained by taking the Z value observed in the most similar observation in the data set B , where the similarity is computed with respect to the X values. Details of those imputation methods are given in the module “Micro-Fusion – Statistical Matching Methods”. In this framework, whatever matching procedure is used, the results will be based on the conditional independence assumption (CIA) of Y and Z given X , which means that the results can be considered acceptable only if, roughly speaking, the information in X is so rich that it can explain the relationships between Y and Z . Another way of looking at the CIA is that the probability distribution of Y conditionally on Z and X depends only on X . Broadly speaking, it is not important to look at the value of Z to be imputed to the observation having Y , once you know its X value. For instance, let us suppose that the variable Y denotes the income, let Z be the level of consumption and let X represent the geographical area. The CIA states that, for example, in order to impute a consumption level to a certain unit, it is not important to know the level of income once you know the geographical area where the unit belongs to.

CIA is a strong assumption that cannot be always assumed and unfortunately it cannot be tested with the available data since Y and Z are not jointly observed. In order to avoid the CIA the use of auxiliary information may be useful and, for this reason in any statistical matching process, enough time should be devoted to search for any kind of additional information. In order to avoid CIA the auxiliary information should be about the variables not jointly observed. It can be in the form of a third file where either (X,Y,Z) or (Y,Z) are jointly observed (e.g., outdated data), or as plausible values of the inestimable parameters of either $(Y,Z|X)$, or (Y,Z) . Information is useful also when it is not exactly about the Y , Z and X but it refers to proxy variables (e.g., outdated data where numerical variables are observed as categorical/ordinal by collecting only ranges). This kind of information is generally sufficient to determine a model without assuming the CIA. Specific methods involving the use of auxiliary information are described in D’Orazio et al. (2006).

Another kind of auxiliary information is that provided by logical rules relating the variables Y and Z (usually named edit rules in the editing procedures, see the topic “Statistical Data Editing”). One

example of logical rule is that it is generally not acceptable that a ten-year-old person is married. This kind of auxiliary information is not generally sufficient to determine a unique model alternative to the CIA, however it can be useful to restrict the possible statistical models compatible with the data at hand. Their use is important to increase the quality of the matching procedure.

For more details on statistical matching see the modules “Micro-Fusion – Statistical Matching” and “Micro-Fusion – Statistical Matching Methods”.

2.1.3 *Micro-integration*

A problem that must be dealt with when integration of different data sources is performed is that of consistency. Procedures in order to make coherent and consistent data at micro level are generally unavoidable. The set of tasks with this purpose are named micro-integration (see Bakker 2011).

A definition of micro-integration is given in Bakker (2011): “*Micro-integration is the method that aims at improving the data quality in combined sources by searching and correcting for the errors on unit level.*”

The term “error” should be understood in a broad sense, Bakker refers to measurement and representation errors.

Representation errors exist if the target population is incompletely described by the data (e.g., over-coverage, under-coverage, ...).

Measurement errors exist if characteristics of the population elements are not correctly described. These errors may have different causes. By using information from different sources, these errors can be detected and corrected. *Harmonisation* is the correction on a conceptual level (for instance harmonising the definition of the variables), while for the correction on data level Bakker uses correction for measurement errors (also known as data reconciliation).

Representation and harmonisation problems are dealt with case by case, there are no general algorithms for this purpose. Since the focus of the module is on general statistical techniques, we do not discuss the so far mentioned problem, the interested reader may refer to Van der Laan (2000).

As far as data reconciliation is concerned, some techniques can be applied. In the module “Micro-Fusion – Reconciling Conflicting Microdata” it is discussed the problems arising when linked records do not satisfy edit-rules and an adjustment step is necessary to integrate the different pieces of information, (the data sources and the edit-constraints), to obtain consistent integrated microdata. One possible strategy is to adjust data in order to satisfy edit-rules. The module “Micro-Fusion – Prorating” describes a ratio adjustment method for balance edits. It solves the possible inconsistencies for each constraint separately by distributing the differences between the total and the items composing the total. The main advantages are that is easy to interpret and to apply.

More refined methods are introduced in literature. In the module “Micro-Fusion – Minimum Adjustment Methods” the data reconciliation task is formalised as a constrained minimisation problem, that is to find the final imputed values such that 1) they differ as little as possible from the observed data and such that 2) they satisfy the edit-rules. The evaluation of changes are computed according to different distances: least squares, weighted least squares and Kullback-Leibler. The procedure is presented according to two different settings: 1) one data set is considered more reliable

than the other, 2) data sets to be integrated are considered equally reliable. The constraints considered in the minimisation problem are linear.

The method described in the module “Micro-Fusion – Generalised Ratio Adjustments” aims to make the adjustments as uniform as possible, and in contrary to the other methods, the method can result in adjustments to variables that are not involved in the constraints. This may be useful to preserve relations between variables that are not connected by edit rules.

Procedures that aim at reaching consistency at output level by modifying data at micro level are still classified as micro-integration procedures. According to this definition, a problem that frequently arises is that of consistency of published figures (for instance frequency tables) when a survey is enriched with register data. An example can be useful to understand the problem. Let us suppose we have a situation like that depicted in Figure 3,

Units	X_1, \dots, X_j	Y_1, \dots, Y_k	w
1			
.			
n			
.			
N			

Figure 3. Micro-integration of a register and a survey.

where grey cells represents the available information, variables $X=(X_1, \dots, X_j)$ are the variables from the register and are observed for all the units in the population composed of N units, variables $Y=(y_1, \dots, y_k)$ are the variables collected in the survey and observed only on a subset of units of the population (sample size n) and finally $w=(w_1, \dots, w_n)$ are the sampling weights associated to each sample unit. Estimates for the parameters (totals, frequency tables) of the variables X can be obtained by using the sampling weights w . The estimates should be consistent with the value computed according to the registered data. This is usually accomplished by using calibration procedures, i.e., by changing as little as possible the values of w such that the two parameters are the same. We notice that also in this case there is a change at micro level (sampling weights) in order to have consistent outputs. The problem becomes more difficult when some of the variables X in the register are not used in the calibration procedure because for instance the treatment of a high number of variables in the calibration is not feasible. In this case an iterative procedure named *consistent repeated weighting* is proposed by (Houbiers et al., 2003; Kroese et al., 2000; Renssen et al., 2001; Houbiers, 2004). It is based on the repeated application of the regression estimator and generates a new set of weights for each table that is estimated. All the tables considered will be consistent.

An alternative approach to reach consistency of the estimates in the latter situation is named *mass imputation*. It consists in imputing all the variables y_1, \dots, y_k in order to obtain a final rectangular data set. Whitridge and Kovar (1990) discuss the practical advantages of such a procedure, in fact the

estimates obtained with such a completed data set are naturally consistent, however Kroese et al., (2000) warn about the fact of having in practice enough degrees of freedom to get a model for imputing data such that all the possible relationships are obtained.

3. Design issues

4. Available software tools

Workpackage 3 of the Essnet on Data Integration includes a thorough discussion on the available software tools (see Scanu, 2008b, Chapter 2).

Some specific references are reported in the following.

StatMatch is an R-package for statistical matching (D’Orazio, 2011) freely available on the website <http://cran.r-project.org/>.

Relais is a freely available software developed by Istat for probabilistic record linkage, downloadable from <http://www.istat.it/it/strumenti/metodi-e-software/software/relais>.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Bakker, B. F. M. (2011), Micro-Integration: State of the art. In: *Report WP1: State-of-the-art on Statistical Methodologies for Data Integration*, ESSNET on Data Integration, available at <http://www.cros-portal.eu/content/wp1-state-art>.

D’Orazio, M. (2011), *Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment*. Vignette for the application of the R package StatMatch, available on CRAN and at <http://www.cros-portal.eu/content/wp3-development-common-software-tools>.

D’Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical matching: theory and practice*. John Wiley, Chichester.

Di Consiglio, L. and Tuoto, T. (2013), Challenges in estimation on probabilistically linked data. Proceedings of NTTTS 2013, available at http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_112.pdf.

DI (2011), Essnet on Data Integration, <http://www.cros-portal.eu/content/data-integration-1>.

Fellegi, I. P. and Sunter, A.B. (1969), A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1210.

Herzog, T., Scheuren, F., and Winkler, W. (2007), *Data Quality and Record Linkage Techniques*. Springer.

- Houbiers, M., Knottnerus, P., Kroese, A. H., Renssen, R. H., and Snijders, V. (2003), Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005, Statistics Netherlands, Voorburg.
- Houbiers, M. (2004), Towards a Social Statistical Database and unified estimates at Statistics Netherlands. *Journal of Official Statistics* **20**, 55–75.
- ISAD (2006), ESSnet on Integration of Survey and Administrative Data, <http://www.cros-portal.eu/content/isad-finished>.
- Kroese, A. H., Renssen, R. H., and Trijssenaar, M. (2000), Weighting or imputation: constructing a consistent set of estimates based on data from different sources. In: P. G. Al and B. F. M. Bakker (eds.), Special Issue: Re-engineering social statistics by micro-integration of different sources, *Netherlands Official Statistics* **15**, Summer, 23–31.
- Pannekoek, J. (2011), Models and algorithms for micro-integration. In: *Report on WP2: Methodological developments*, ESSNET on Data Integration, available at <http://www.cros-portal.eu/content/wp2-development-methods>.
- Renssen, R. H., Kroese, A. H., and Willeboordse, A. (2001), Aligning estimates by repeated weighting. Research paper 491-01-TMO, Statistics Netherlands, Voorburg/Heerlen.
- Van der Laan, P. (2000), Integrating Administrative Registers and Household Surveys. In: P. G. Al and B. F. M. Bakker (eds.), Special Issue: Re-engineering social statistics by micro-integration of different sources, *Netherlands Official Statistics* **15**, Summer, 7–15.
- Whitridge, P. and Kovar, J. G. (1990), Use of mass imputation to estimate for subsample variables. In: *Proc. Bus. Econ. Statist. Sect.*, American Statistical Association, Washington, DC, 132–137.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Object Matching (Record Linkage)
2. Micro-Fusion – Probabilistic Record Linkage
3. Micro-Fusion – Statistical Matching
4. Statistical Data Editing – Main Module
5. Imputation – Main Module
6. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

1. Micro-Fusion – Unweighted Matching of Object Characteristics
2. Micro-Fusion – Weighted Matching of Object Characteristics
3. Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage
4. Micro-Fusion – Statistical Matching Methods
5. Micro-Fusion – Reconciling Conflicting Microdata
6. Micro-Fusion – Prorating
7. Micro-Fusion – Minimum Adjustment Methods
8. Micro-Fusion – Generalised Ratio Adjustments

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.1: Integrate data

Administrative section

14. Module code

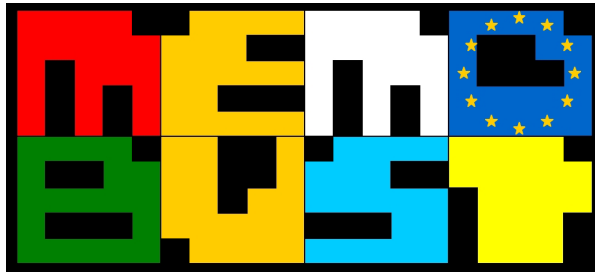
Micro-Fusion-T-Data Fusion at Micro Level

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-03-2012	first version	Marco Di Zio	Istat (Italy)
0.2	06-04-2012	second version	Marco Di Zio	Istat (Italy)
0.3	11-11-2013	revision based on EB comments for preliminary release. Final template is used	Marco Di Zio	Istat (Italy)
0.3.1	18-11-2013	preliminary release		
0.4	19-12-2013	final release based on EB comments	Marco Di Zio	Istat (Italy)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:56



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Object Matching (Record Linkage)

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Purpose of matching.....	3
2.2 What is matching?	4
2.3 Overview of object matching	5
2.4 Matching errors	6
2.5 Why is matching complex?	7
2.6 Matching applications	7
3. Design issues	8
4. Available software tools.....	8
5. Decision tree of methods	8
6. Glossary.....	9
7. References	9
Interconnections with other modules.....	10
Administrative section.....	12

General section

1. Summary

The aim of object matching (more commonly known as record linkage or as record matching) is to match the same units that are represented by records in two different files. This is to be contrasted with synthetic (or statistical) matching where the aim is to match similar, but usually different, units. Depending on the kind and quality of the information available a suitable matching method should be identified. In case object identifiers of good quality are available in both files, it is quite straightforward to use these to find the records matching on this key. Complications may arise when such object identifiers are not present. In that case one should investigate if object characteristics are present in both files that can be used for finding matches. Several methods exist that deal with this situation. The aim of the present module is to provide a context and overview of the various matching methods, and to give pointers to the specialised method modules in this handbook dealing with these methods.

2. General description

2.1 Purpose of matching

The increasing demand for timely, detailed and high-quality statistics combined with the obligation to use existing registries as much as possible makes it necessary to find alternative ways to produce statistics, such as by matching information from different files. Registries, for example, are not designed to produce statistics. To produce the desired statistics anyway, it is necessary to match registries and survey data to create more usable data sets. In this context, longitudinal data must also be taken into account. On the output side, there is more of a need to present events in their mutual relationships and not only as separate statistics. Matching of files makes it possible to publish over broader themes and to develop new output.

Data matching contributes, for example, to the following:

- Faster publishing of new output;
- Better quality of data through, for example, mutual confrontation;
- Reduction of the survey pressure and therefore lower costs for the respondents;
- Reduction of the costs of the NSI because it no longer needs to conduct surveys in a particular areas.

Data matching therefore supports the main goals of the NSI, such as creating new output, generate less survey burden, make better use of administrative sources and operate more efficiently.

Recent information on matching can be found among in Herzog et al. (2007) and the documents of the ESSnet project on Integration:

<http://www.cros-portal.eu/content/data-integration-1>

and

<http://www.cros-portal.eu/content/work-packages-and-executive-summary>.

Willenborg and Heerschap (2012) was used as a source of the present module (as well as of several of the modules on matching in this handbook).

2.2 What is matching?

Matching is about combining information from two or more records (each representing units in a target population), which are believed to relate to the same unit (or object), such as a person, business or region (see Newcombe, 1988). Normally in the matching process, two similar records, present in two different files (known as matching files) are combined, based on various criteria and preconditions. It should be stressed that this type of matching is different from that in statistical matching, where the aim is to match objects that are similar but not identical. Statistical matching therefore, although in execution being very close to the type of matching considered here, is more akin to imputation. (See the theme module “Micro-Fusion – Statistical Matching” in the handbook.)

The most direct case of matching concerns object identity matching. Here one attempts to join objects represented in different data files using identifiers for the objects. For this purpose, a matching key is used consisting of several (key) variables that both files have in common. The matching criterion can then be: ‘exactly the same scores on the matching key’. This is a relatively simple (but important) situation that often exists in practice.

Object identifiers suitable for object identifier matching are not always available in matching situations. However, it may be the case that object characteristics are present in the files to be matched, that allow certain objects (records) to be matched. As these characteristics are not key values that identify objects uniquely, it is possible that for a given object there are several candidates. In this case the matching takes place in two steps:

1. It is determined which records are *matching candidates*, potential matches, so to speak, and
2. From all possible matching candidates, the *best subset* is selected, which satisfies certain criteria (preconditions), for example, that no single record is matched with two or more records.

It is possible to simply indicate which objects are matching candidates or not, or it may be possible to differentiate in the strength of being matching candidates, using matching weights to express the strength of the matching. Matching candidates that have more characteristics in common are than stronger matches than those with less. These matching weights may also be probabilities, derived from a probabilistic matching model.

The decision to match or not to match objects (thus determining which matching candidates are considered matches) is generally made by a matching programme. If the matching takes place interactively or manually, a matching specialist takes these decisions.

In Figure 1 a schematic view of the matching process is presented. It indicates that in case two files are matched, first matching criteria have to be identified, including the choice of matching variables, when two objects are considered matches or not (in case object identifiers are used) or how to calculate the strengths of possible matches (matching candidates), expressed in matching weights. The matching that is then carried out yields matching candidates. From these the final matches are determined. It generally yields three subgroups of the group of matching candidates: those matching candidates that are considered as matches, those matching candidates that are considered as non-matches, and those

matching candidates for which it is not so clear whether they match or not (the doubtful cases). The first group is the one that will be used for further analysis. The second group is not. In case the third group is big, it may be that the matching is repeated, this time with (slightly) different matching criteria, in the hope that the yield (the size of first group, the matches) may be higher.

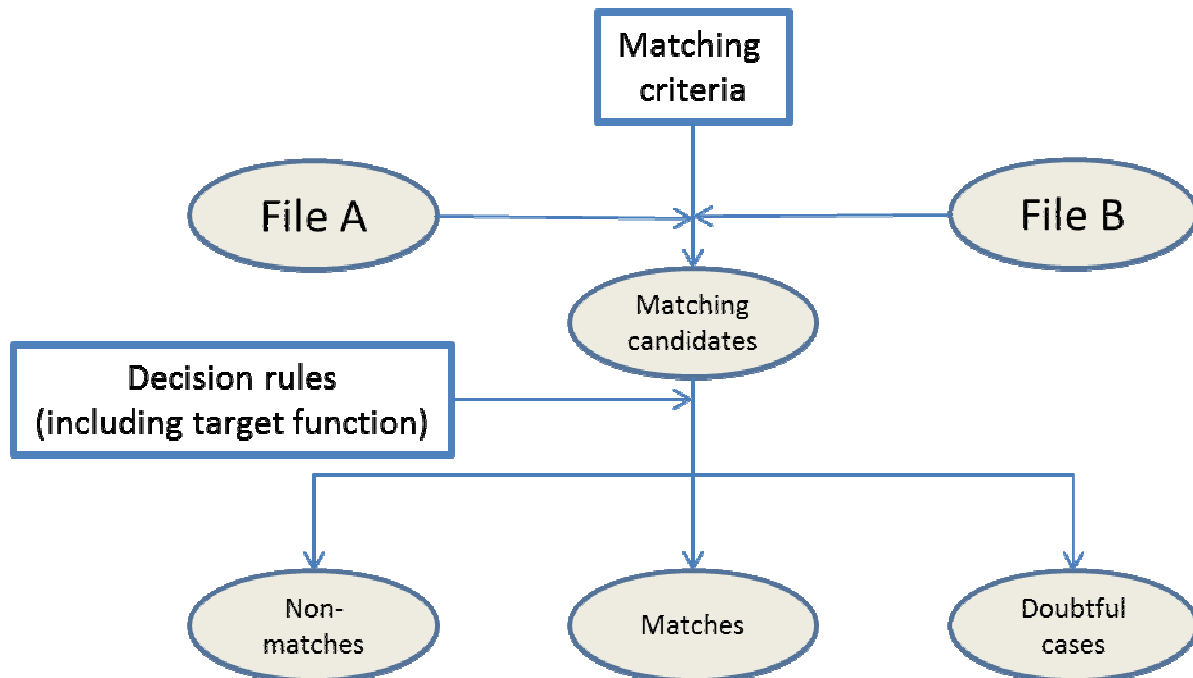


Figure 1. Main ingredients in matching.

In the next section we consider various matching methods in a bit more detail. But its main objective is to refer to the various modules in the handbook that deal with these methods in more detail.

2.3 Overview of object matching

In the handbook several matching methods are discussed focussing on matching identical objects.

The first one is uses object identifiers for matching. In this case for the objects to be matched object identifiers (also known as keys) are available. They have the property that they uniquely identify objects. They are ideal for matching objects, provided they are free of error. This is a matching method that is typical for, but not limited to, databases, where it is known as ‘joining’. This method is important as it is used frequently in practice. It is the simplest of the matching methods that we address in this report. For more information see the method module “Micro-Fusion – Object Identifier Matching”.

In practice object identifiers are not always available. But characteristics of objects may be available for matching. That brings us to the next form of object matching, namely that which uses object characteristics of objects. In fact, there is no single method for this kind of matching. We distinguish between two types of methods. The one uses no matching weights and the other does to distinguish in the strength of potential matches.

The first group of methods of methods dealing with object characteristics does not use matching weights to differentiate between the strength of matches. It is elaborated in the method module “Micro-Fusion – Unweighted Matching of Object Characteristics”. The second group of methods dealing with object characteristics is uses matching weights to express differences in strengths of potential matches The matching weights use to express the strength of potential matches can be calculated in various ways, depending on the problem at hand. One can use a metric (or distance function) or measure of dissimilarity to quantify how object characteristics differ. This class of matching methods is elaborated in the method module “Micro-Fusion – Weighted Matching of Object Characteristics”.

A special case of weighted matching is probabilistic record linkage. In this case the matching weights are derived from a probabilistic matching model. More details on this type of matching can be found in this handbook in the theme module “Micro-Fusion – Probabilistic Record Linkage”.

A special case of probabilistic matching that deserves special attention is a classical method proposed by Fellegi and Sunter (1969) and refined by Jaro (1989). In the handbook it is discussed in the method module “Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage”.

2.4 Matching errors

The matching of two files may lead to errors for various reasons (see also Section 2.5). After the matching candidates have been identified and the matches selected from them, two kinds of errors may result:

- Mismatches: records that are matched, but are not actually associated with the same objects.
- Missed matches: records that are not matched, but that are actually associated with the same objects.

Table 1 contains an overview of the various matching errors including the various names that are being in the literature used to indicate them.

Table 1. Object matching errors

	Objects associated with same unit	Objects associated with different units
Objects matched	<ul style="list-style-type: none"> - good result - rightly matched - correct match 	<ul style="list-style-type: none"> - mismatch - false positive match - type I error - erroneously matched
Objects not matched	<ul style="list-style-type: none"> - missed match - false negative match - type II error - erroneously not matched 	<ul style="list-style-type: none"> - good result - rightly not matched - correct unmatched

In practice, it is usually unknown whether a match of two records is correct, or a mismatch, or when two records should have been matched because they pertain to the same object (missed match). Nevertheless, it is useful to distinguish these errors.

2.5 *Why is matching complex?*

At first glance, the matching of files seems to be a simple task. In practice, however, this is seldom the case, especially in the context of business statistics. The following causes contribute to this circumstance:

- The *quality and the structure of the data* in the files to be matched. It will seldom be the case that the data provided, and therefore also matching variable data, do not contain ‘noise’. During processing, for example, observation and processing errors, such as typing errors, can occur. Consequently, it is possible that records that actually do correspond do not match, or vice versa. With respect to the structure of the data provided, it is possible, for example, for the scores of the matching variables to be good in both records, while they are represented in such a way that it is difficult to compare these with one other via automation. All of these aspects make the pre-processing stage important. This is where both the quality and the structure of the data can be adapted and improved, insofar as is necessary for matching.
- The *units of files to be matched may differ*, but still can be derived from one another. Consider, for example, a file with Business Units that must be linked with a file with Enterprise Groups. In this context, a matching table should be used that sets out the relationship between both units.
- The use of *different domains or classification divisions* for the matching variables. Here as well, it is desirable for the matching process that the domains or classifications are compatible.
- The *time dimension*. The matching variables or units are dynamic and were observed at different moments in time. This could be the case, for example, for businesses. In the time between two different observations, which are saved in the two different files, the enterprise may have split or merged, while it still has the same identifier or matching variable. In the matching process, this would seem to refer to the same enterprise, while in reality, the enterprise may not be the same anymore.

2.6 *Matching applications*

Examples of matching applications in the statistical process are the following:

- **Micro-fusion.** In this process, different pieces of data are confronted with each other, and a variety of differences about businesses may become apparent. The aim is then to explain and eliminate these differences. Confronting the data is only possible after the files have been matched. See the various modules in the handbook on micro-fusion, in particular those dealing with differences in the data and how to reconcile them.
- **Input matching.** Starting with the building of a statistical frame. Usually, a combination of sources is needed to compile such a frame or ‘backbone’, for example, the General Business Register. In the Netherlands, for example, matched data from the Chamber of Commerce and Tax

Administration are used. For more information on this see the modules of the topic “Statistical Registers and Frames” in the handbook.

- **Statistical matching.** Statistical (or synthetic) matching is concerned with filling in missing values in a file, and an auxiliary file is used for this purpose. Information from *similar* objects is used to fill in the missing values. So the goal of statistical matching is to match similar objects, not (necessarily) identical ones. The method can be viewed as an imputation method. See the theme module “Micro-Fusion – Statistical Matching”.
- **Allocation of CATI interviewers to sample elements.** The matching is carried out for the purpose of interviewing businesses, say. Here, the problem is deciding which interviewer should call which business at what time. The matching between interviewer and business to be called is done in several steps. First the deployment of the interviewers is scheduled. When they are at work they get telephone numbers of businesses assigned that they should call for CATI interviews. For more information on this see the theme module “Data Collection – CATI Allocation”.
- **Coding.** In this process, descriptions given by respondents in their own words are matched with codes from a classification. One of the problems here involves matching of words, while knowing that the respondents could have potentially made spelling or grammatical errors or used synonyms, hyponyms or hypernyms. See the modules of the topic “Coding” in this handbook for more information on this subject.

3. Design issues

4. Available software tools

Data matching is virtually impossible without the use of a specialised software package. Some examples of matching software tools are the following:

- **Trillium** (Harte-Hanks; www.Trilliumsoftware.com).
- **SSA NAME3** (Search Software America; www.searchsoftware.com).
- **IQ-Matcher** (Intech Solutions; <http://www.intechsolutions.com.au>).
- **Other matching tools** include: GDriver (US Census Bureau/Winkler), Relais (Istat), LinkageWiz, Tailor (a record linkage toolbox), NameSearch from Intelligent Search Technology, PA Oyster Engine, Fril, OxLink and Alta.

5. Decision tree of methods

Figure 2 presents a decision tree for the application of the various matching methods considered in the handbook.

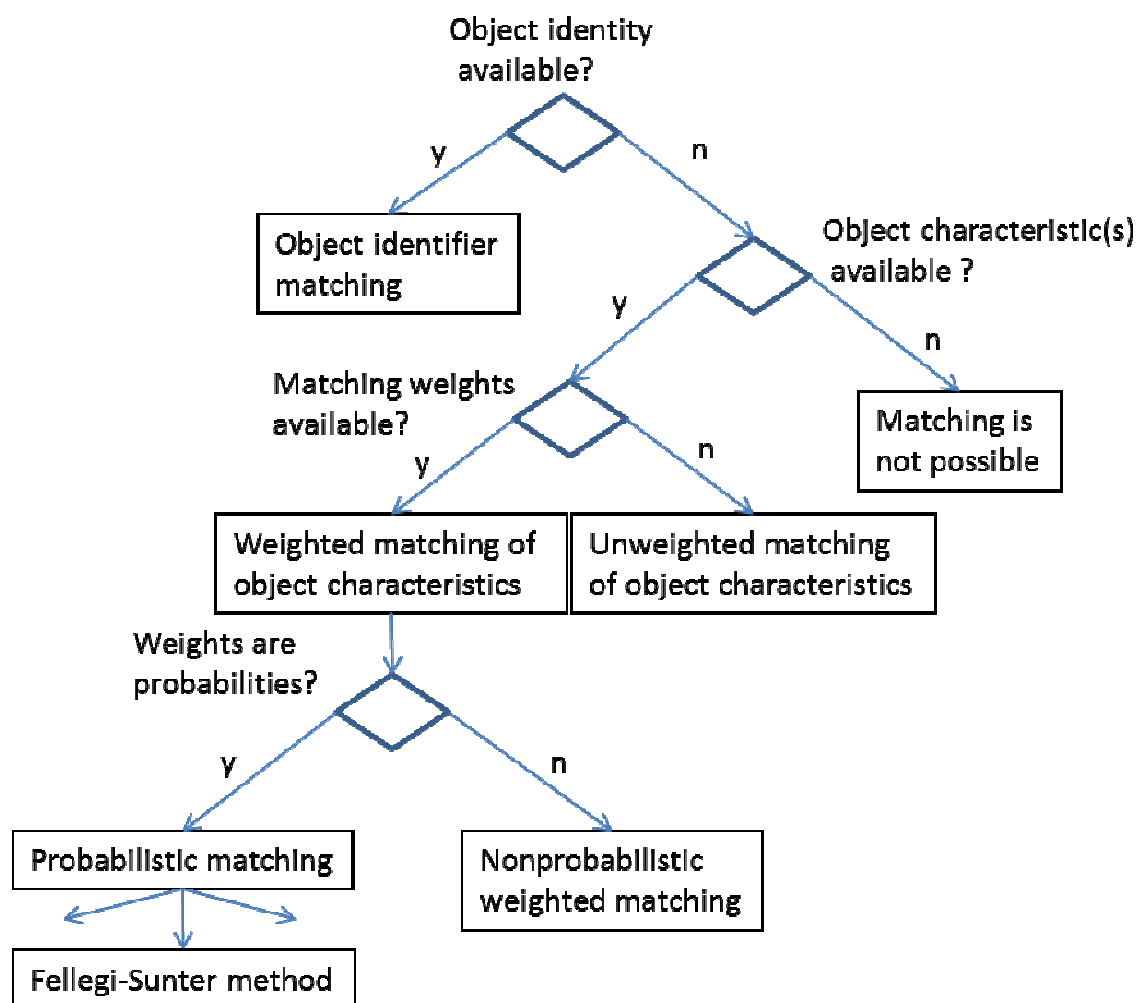


Figure 2. Overview of different matching methods.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Fellegi, I. P. and Sunter, A. B. (1969), A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1200.
- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007), *Data quality and record linkage techniques*. Springer.
- Jaro, M. A. (1989), Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* **84**, 414–420.
- Newcombe, H. B. (1988), *Handbook of record linkage*. Oxford University Press.
- Willenborg, L. and Heerschap, N. (2012), *Matching*. Contribution to the Methods Series, Statistics Netherlands, The Hague.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Registers and Frames – Main Module
2. Data Collection – CATI Allocation
3. Micro-Fusion – Data Fusion at Micro Level
4. Micro-Fusion – Probabilistic Record Linkage
5. Micro-Fusion – Statistical Matching
6. Coding – Main Module
7. Imputation – Main Module
8. Dissemination – Dissemination of Business Statistics

9. Methods explicitly referred to in this module

1. Micro-Fusion – Object Identifier Matching
2. Micro-Fusion – Unweighted Matching of Object Characteristics
3. Micro-Fusion – Weighted Matching of Object Characteristics
4. Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage
5. Micro-Fusion – Statistical Matching Methods

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5.1 Micro-integration.
2. Phase 5.2 Coding.

12. Tools explicitly referred to in this module

1. Alta.
2. Fril.
3. GDriver.
4. IQ-Matcher.
5. Linkage Wiz.
6. NameSearch.
7. Oxlink.

8. PA Oyster Engine.
9. Relais.
10. SSA Name3.
11. Tailor.
12. Trillium.

13. Process steps explicitly referred to in this module

1. Integration / micro-aggregation of information
2. Coding
3. Allocation of sample units to interviewers
4. Dissemination of information
5. Statistical (synthetic) matching

Administrative section

14. Module code

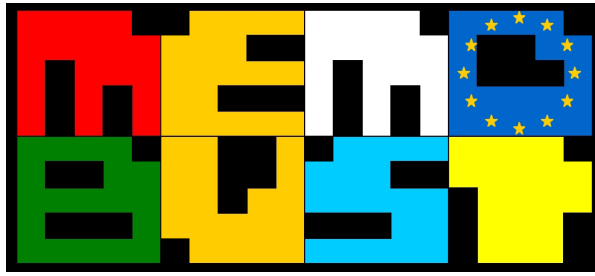
Micro-Fusion-T-Object Matching

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	30-06-2012	first version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.2	02-07-2012	second version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.3	11-07-2013	third version	Leon Willenborg	CBS (Netherlands)
0.4	09-08-2013	new version (using review comments)	Leon Willenborg	CBS (Netherlands)
0.4.1	21-08-2013	minor revisions	Leon Willenborg	CBS (Netherlands)
0.5	03-11-2013	new version (using EB review comments)	Leon Willenborg	CBS (Netherlands)
0.5.1	18-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:56



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Object Identifier Matching

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Two steps.....	3
3. Preparatory phase	4
4. Examples – not tool specific.....	4
4.1 First example	4
4.2 Second example.....	4
4.3 Third example.....	5
5. Examples – tool specific.....	5
6. Glossary.....	5
7. References	5
Specific section.....	6
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

The matching of records in two data sets is considered, on the basis of common object identifiers (key variables). The scores on these object identifiers are assumed to be of good quality in both data sets, though they need not be perfect. To understand the context of this type of matching in the handbook, the reader is referred to the theme module “Micro-Fusion – Object Matching (Record Linkage)”. The present module is based on Willenborg and Heerschap (2012).

2. General description of the method

Matching based on an object identifier variable is the simplest way to match. Both matching data sets contain the same unique object identifier that is used as the matching key. The assumption is that the quality of the object identifier is sufficiently high; otherwise this matching method cannot be used effectively. Although we talk about an object identifier, it may actually exist of more than one variable, these are referred to as ‘key variables’.

The basic principle is that a match is made if and only if a record from one dataset has exactly the same object identifier (key) value as another record from the second dataset. This type of matches is standard in databases, where it is called a ‘join’, or an ‘equijoin’. See, e.g., Date (2000), Elmasri and Navathe (2004), or another book on relational databases.

In object identifier matching there are two input data sets that need to be matched. It would be perfectly acceptable if both data sets play interchangeable roles. In practice, however, often there is a primary input data set. The idea is to ‘enrich’ the records of this data set with values from the second input data set through matching. So the records in the primary input data set act as receptors and those of the second input data set, as donors. In this situation the roles of both input sets is not symmetric anymore.

Exact matching, or ‘joining’ as it is defined above, describes an ideal situation, in the sense that there are no errors in the object identifiers. In practice this ideal situation may not exist because some object identifier values are in fact erroneous, for instance because they were wrongly copied from another source. This is what makes the present method nontrivial, as the ideal case is very simple, conceptually. If the data sets are big there may be computational problems when matching the two files. In this case it may be partition the data sets into blocks that are manageable. When matching the records in a particular block of one file, only the records of a specific block in the other file is considered to find matching pairs. This blocking may result in missed matches.

2.1 *Two steps*

The assumption underlying the object identifier matching method in the ideal case is that the matching keys used in both data sets are error free. In practice, however, they need not be perfect. It is sufficient if they are of good quality. This allows that enough records can be matched, although there is a chance that mismatches or missed matches will occur.

First step: Records from both data sets are matched on the basis of exact equality of the object identifier scores. In this version of the method it is assumed that each record of the first data set has at most one match in the second dataset.

Second step: If some records of the first data set are not matched, this may be due to errors in the object identifier values. In a second step it is attempted to match any of the remaining records using the object identifier only.

The errors in the object identifiers may be due to typing errors: a wrong character was typed, two neighbouring characters were wrongfully interchanged, a character was wrongfully not typed (or deleted), or an extra character was wrongfully typed, etc. With this in mind it could be possible to correct for a missed match. This is attempted in this second step. The idea is to look among the missed matches and find pairs that are close in terms of the Levenshtein (or Damerau-Levenshtein) distance. See also Example 4.2 below. The distance is discussed in the method module on weighted matching of object characteristics in the handbook. If some records of the first data set do not match in the second step, they can still be part of the output data set, where the added variables are missing. Whether this is allowable depends on the variant of the method that is used.

3. Preparatory phase

The quality of the object identifier scores in both data sets should be assessed, to see if the Primary key matching method is applicable. If this seems to be the case, the first step can be attempted. Depending on the number of unmatched records one has to decide what to do next. Go ahead with the method or not. And if so, choose a suitable metric, depending on the variables in the object identifier.

4. Examples – not tool specific

Most of the examples below refer to Statistics Netherlands, but the issue at stake in each case can be generalised.

4.1 First example

The matching of enterprises from two surveys, which are both based on the General Business Register. In both data sets, the unit – the enterprise – is identified by an eight-digit business identification number (a BEID). The BEID is the object identifier on which matching takes place. If the BEIDs in both data sets are the same, then a match is made; if the BEIDs are not the same, then the units are not matched. For example, no account is taken of the fact that, during the processing procedure for the individual statistics, errors could have crept into the BEIDs. This check is also often difficult because, in many cases, there are no more object characteristics present, such as names and addresses.

4.2 Second example

Suppose that BEID is used as the object identifier, and you also have a complete list with BEIDs with at least some information about the businesses concerned. If a BEID is found that does not seem to be correct, then you could look in the neighbourhood of this number in the list. The idea here is that a mistake was made when copying the number, for example, two digits were interchanged, or a 5 was replaced by a 6 (or vice versa) or a 7 by a 1 (or vice versa), etc. If, for example, you search for all BEIDs with a Levenshtein distance of 1 or 2 from the given BEID, and also compare the associated

business attributes with the data in the dataset or register concerned, you could potentially find the correct BEID with the associated business attributes.

4.3 *Third example*

For privacy or data protection reasons external object identifiers can be replaced by secure internal object identifiers. The advantage is that it is impossible to link other external information to the records. This does prevent direct identification of units.

If E is the set of external keys and I the set of internally used keys, this replacement can be represented by a function $k : E \rightarrow I$, which should be injective.

5. **Examples – tool specific**

6. **Glossary**

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. **References**

Date, C. J. (2000), *An Introduction to Database Systems*, 7th edition. Addison-Wesley.

Elmasri, R. and Navathe, S. B. (2004), *Fundamentals of Database Systems*. Addison-Wesley.

Willenborg, L. and Heerschap, N. (2012), *Matching*. Contribution to Methods Series, Statistics Netherlands, The Hague.

Specific section

8. Purpose of the method

Enriching records in a given microdata set with information from a second microdata set.

9. Recommended use of the method

1. The method can be applied in case object identifiers (key variables) of good quality are available in both matching data sets.

10. Possible disadvantages of the method

1. If the quality of the object identifier values (key values) is not very high, the number of mismatches or missed matches may be substantial.

11. Variants of the method

- 1.

12. Input data

1. There are two input data sets, typically a primary input data set whose records are supposed to be 'enriched' by information from records from the second input data set through (object identifier) matching.

13. Logical preconditions

1. Missing values
 1. The object identifier values used in the matching are not supposed to be missing (too often). In case they are missing in some records the corresponding objects cannot be matched using object identifier matching (but possibly with object characteristics matching).
2. Erroneous values
 1. Errors in the object identifier (key variables) are allowed for some records in the input data files.
3. Other quality related preconditions
 1. The object identifiers used for matching are not supposed to change in time. If they do this could result in matching errors (mismatches or missed matches).
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. In case the object identifier is (near) perfect (error-free) and the input data sets are small there is nothing to tune. Sorting on the object identifiers in both input files will yield an easy method to match.

2. In case the input files are big, blocking may be appropriate, that is partitioning the data sets into blocks. The choice of a good blocking variable is part of the tuning in this case.
3. In case the object identifier is not perfect (but good enough) matching on equality of object identifiers can still be carried out but may result in some mismatches or missed matches. In case more sophisticated matching criteria are used and metrics, one is in fact in the area of another type of matching, namely object characteristics matching. (An object identifier with quite some errors is more of an object characteristic.) See the method modules “Micro-Fusion – Unweighted Matching of Object Characteristics” and “Micro-Fusion – Weighted Matching of Object Characteristics” for the tuning parameters needed in those cases.

15. Recommended use of the individual variants of the method

1. In case of big input data sets the use of blocking may be applied, to split the data files in smaller blocks. This typically requires the use of one or more blocking variables.

16. Output data

1. A microdata set containing all variables of primary input data set, with variables added from the second input data set.
2. Optional data set containing all non-matching records from the primary input data set.
3. Optional data set containing all non-matching records from the second input data set.

17. Properties of the output data

1. There may be a set of matches (records from the primary input data set enriched with information from the second input data set) and a set of non-matches (from the both input data sets). In case the object key is not error-free the matches may contain false matches, and among the non-matches there may be missed matches.

18. Unit of input data suitable for the method

Objects are the units of input in this method. The objects are assumed to correspond with records in two data sets, conceptually not necessarily physically. The physical representation may be different, for instance when the objects are presented in normalised relational databases. Here the information about an object is physically scattered over various tables.

19. User interaction - not tool specific

1. Before matching the tuning parameters must be set by analysing the results for different values.
2. No user interaction during matching.
3. After matching and assessment must be made of the number of mismatches and missed matches.

20. Logging indicators

1. Number of non-matching records from the primary input data set.

2. Number of non-matching records from the second input data set.
3. Time used for the matching.

21. Quality indicators of the output data

1. The number of mismatches or missed matches and the number of missed matches can be used as quality indicators. The quality of the matching method can be assessed based on the inspection of matches of test files. It may be a labour intensive job to carry out in case the matching files are big. First impressions may be obtained from inspecting a sample of the matched records, and of the non-matched records in the input data sets.

22. Actual use of the method

1. In case good quality object identifiers are available in the two files to be matched, object identifier matching is the preferred matching method.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Object Matching (Record Linkage)

24. Related methods described in other modules

1. Micro-Fusion – Unweighted Matching of Object Characteristics
2. Micro-Fusion – Weighted Matching of Object Characteristics

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. 5.1 Integrate data

27. Tools that implement the method described in this module

- 1.

28. Process step performed by the method

Adding variables to microdata set

Administrative section

29. Module code

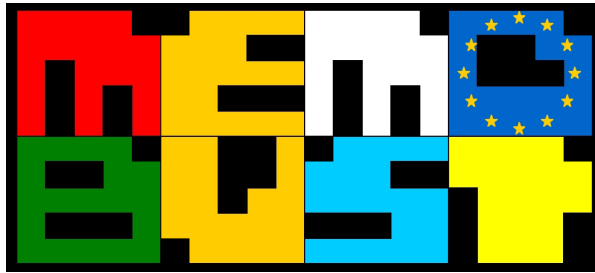
Micro-Fusion-M-Object Identifier Matching

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	23-04-2012	first version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.2	02-07-2012	second version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.3	11-07-2013	third version	Leon Willenborg	CBS (Netherlands)
0.4	09-08-2013	revised version (using review comments)	Leon Willenborg	CBS (Netherlands)
0.5	29-10-2013	revised version (using EB review comments)	Leon Willenborg	CBS (Netherlands)
0.5.1	18-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:57



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Unweighted Matching of Object Characteristics

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	5
4. Examples – not tool specific.....	6
4.1 First example	6
4.2 Second example.....	6
4.3 Third example.....	7
5. Examples – tool specific.....	8
6. Glossary.....	8
7. References	8
Specific section.....	9
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

The method discussed in the present module is intended for matching two data sets on the basis of object characteristics. It is applied in case no object identifiers (of good quality) are available from both datasets. First the potentially matching records in the two data sets are identified. This requires a suitable metric and a cut-off value so that records that are too different are not considered as candidate matches. In the next step from these potential matches, a subset is computed that maximises the number of matches, under suitable constraints. The present module is based on Willenborg and Heerschap (2012). The reader is advised to read the theme module “Micro-Fusion – Object Matching (Record Linkage)” first before reading the present one. The method module “Micro-Fusion – Weighted Matching of Object Characteristics” should also be consulted, as the method described in the present module is a special case of the method discussed there. In particular it contains relevant information on metrics and graphs that are used in the present module.

2. General description of the method

The method consists of several steps. Here we discuss only the most important ones, leaving aside the preparatory steps. Part of the preparation is finding a cut-off value for the metric. This may actually take several iterations, as the cut-off value needs to be set in such a way that not too many candidate matches are generated, and not too few.

First step: The computation of matching candidates.

Using a metric and a cut-off value the set of records that can be matched is computed. To be a candidate match the distance of the two records involved must be smaller than the cut-off value.

Second step. The computation of the final matches from the candidate matches, under constraints. The constraint is that each record from both data sets can be at most in one match. The objective is to find as many matches as possible, obeying the constraint.

Both steps are typically done by computer. The second step formally requires the solution of an optimisation model. See the next subsection.

First select all records that are in a single match, and remove these from the candidate matches.

We can apply the optimisation to the remaining candidate matches, which is ambiguous in the sense that at least one of the records involved has two or more matches. In contrast to the weighted matching method, these candidate matches are of equal value, if they lead to the same total number of matches. A (random) choice must be made from the remaining candidates, or additional object characteristics are necessary. Depending on the number of remaining candidates, there is a risk of a false match, as for this method both matching records are intended to belong to the same object. This results in a number of mismatches.

We now give some comments on the approach taken in the method in the first and in the second step. If we denote a Hamming distance by d_H and we interpret the scores on the matching key as vectors of length n , then the matching criterion used here is in fact:

First step (for Hamming metric d_H): for $\alpha \in A$, $\beta \in B$, α and β are matching candidates if and only if $d_H(\alpha, \beta) \leq k$.

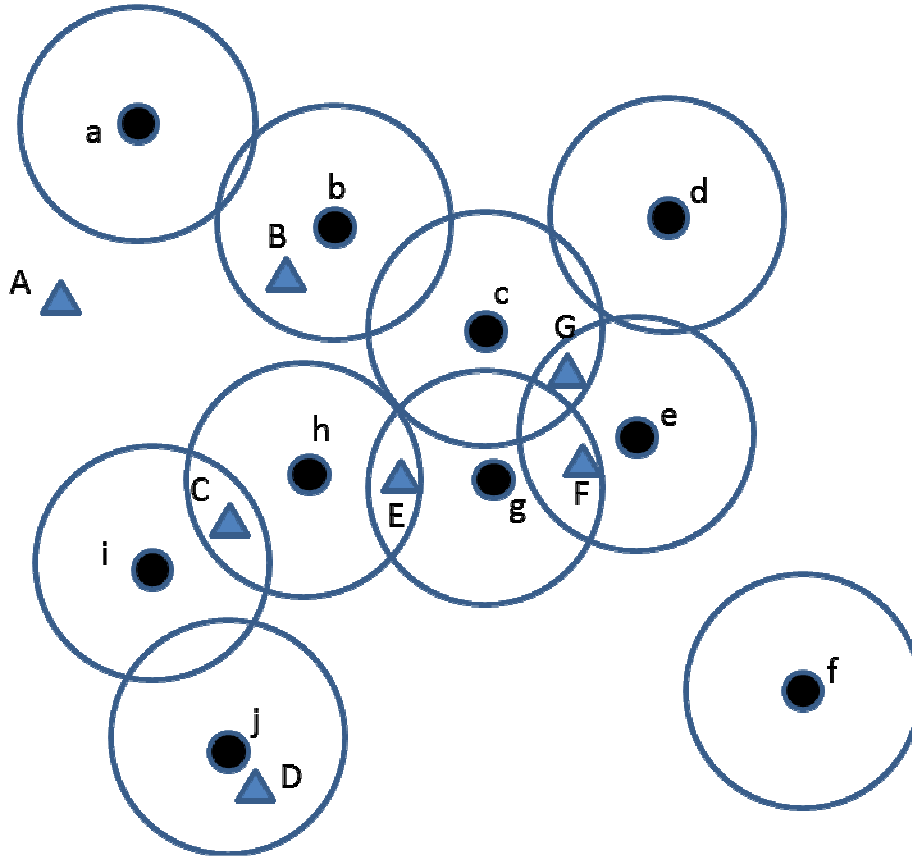


Figure 1. Records from two different files represented as points and neighbourhoods of the records from one of the files.

We can formulate this in such a way that all β from B that are inside a ball with radius k (using d_H as a metric) around α provide all matching candidates for α . See also Figure 1, illustrating the idea in case the neighbourhoods are in fact circles.

For each α in A, we can ascertain this in B. (Or vice versa, for each β in B, we can figure out which α in A are inside a circle with radius k around β . That produces the same result.) Note that, here, we only use whether a record is present in a circle around a ‘point’, not at what exact distance it is from that point. We could, in fact, use this distance as a matching weight: the smaller the distance the higher this weight. In the module “Micro-Fusion – Weighted Matching of Object Characteristics” this approach is described.

If we look at the above section critically, we can conclude that the selection of a Hamming distance is not essential for the approach taken; we could just as well have chosen another metric to arrive at a similar matching criterion. Therefore, based on a metric d , it is possible to formulate a matching criterion:

First step (for general metric d): for $\alpha \in A$, $\beta \in B$, α and β are matching candidates if and only if $d(\alpha, \beta) \leq k$.

Once again, we can formulate this in such a way that all β from B that are inside a ball with radius k around α (measured using d), provides all matching candidates for α . For each α in A , we can ascertain this in B . (Or vice versa, for each β in B , we can figure out which α in A are inside a circle with radius k around β .) All of this produces a matching candidate graph (MC graph), where records from A and B are represented, and those which are potential matches are connected by an edge. See Figure 2.

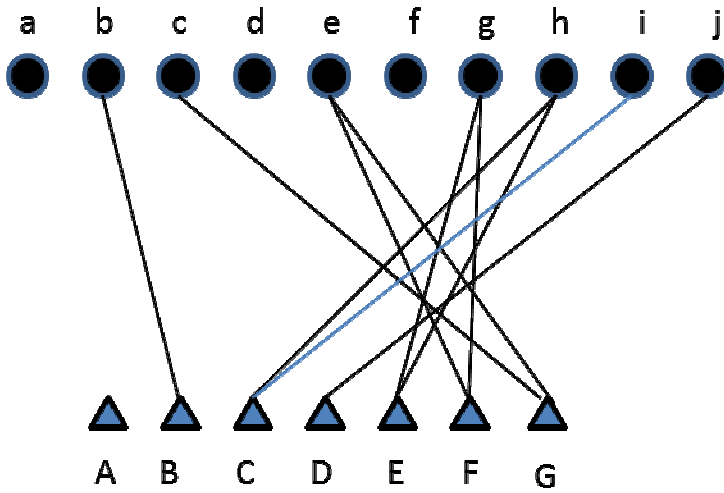


Figure 2. MC graph, representing the situation in Figure 1.

Second step: This involves the solution of the matching problem, in that matches have to be selected from candidate matches, under constraints. A typical constraint in this situation is that each record can be matched with at most one record in the other file. This is a well-known problem in combinatorial optimisation. It is discussed in books like Lawler (1976, Ch.5), Papadimitriou and Steiglitz (1998, Ch.10) or Nemhauser and Wolsey (1988, Ch. III.2), to which the interested reader is kindly referred.

3. Preparatory phase

The object characteristics common to both data sets to be matched are identified. It has to be decided if they are suitable for this type of matching; the number of potential matches should not be too big. A suitable metric for these variables should be found, as well as a suitable cut-off value. This requires some experimenting: with the cut-off value the number of potential matches can be controlled. In case the matching data sets are big special measures should be taken, such as blocking to create a manageable matching problem.

Now candidate matches can be found, from which the matches are to be calculated.

4. Examples – not tool specific

4.1 First example

An MC graph is formally a bipartite graph and is defined as follows for a matching problem where there are two data sets A and B, and a matching criterion K used on a matching key S. For an example of an MC graph without matching weights, see Figure 3.

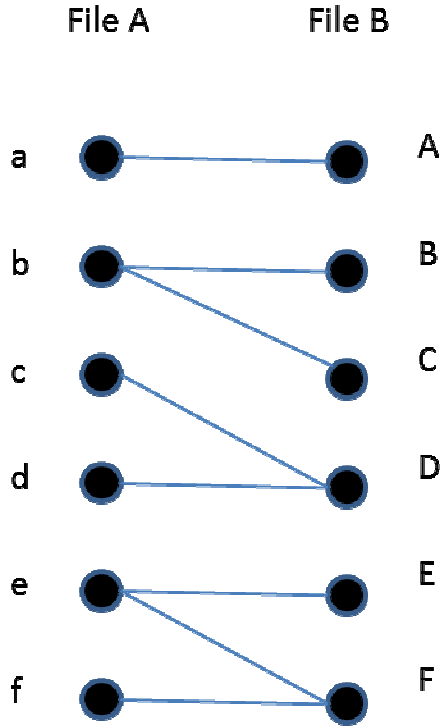


Figure 3. Example of MC-graph without matching weights.

We take data sets A and B to be sets of records. $G = (V, E)$ is the MC graph for this matching problem. The node set V is given by $V = A \cup B$ and the edge set E consist of the pairs $\{a, b\}$ where $a \in A, b \in B$ which furthermore satisfy the matching criterion K.

4.2 Second example

An MC graph is depicted in Figure 4. The edges indicate the matching candidates. The match $\{d, h\}$ is the only one that can be made unambiguously, separate from the matching criterion used. Depending on the matching criterion, more matches can be made. If this concerns a 1:1-matching, two additional matches are possible:

1. $\{a, g\}$ or $\{a, i\}$ (one of the two)
2. $\{c, f\}$ or $\{e, f\}$ (one of the two).

The choices in 1. and 2. can be made independently of each other.

In the case of an MC graph without weights, the candidate matches all count the same.

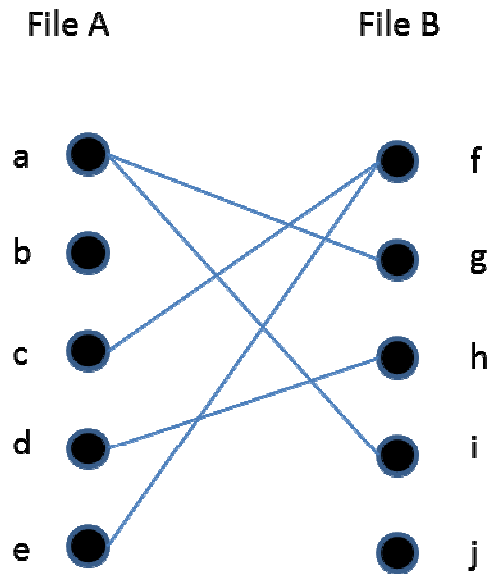


Figure 4. An MC graph

4.3 Third example

We consider two matching variables that are similar but not exactly the same. Specifically, we are talking about two age variables. One of the age variables, which specifies age in two-year classes, occurs in matching file A, and the other, which represents age in three-year classes, is found in matching file B. Depending on the reference times for each file (the time to which the data relate), we can make a connection between the age categories.

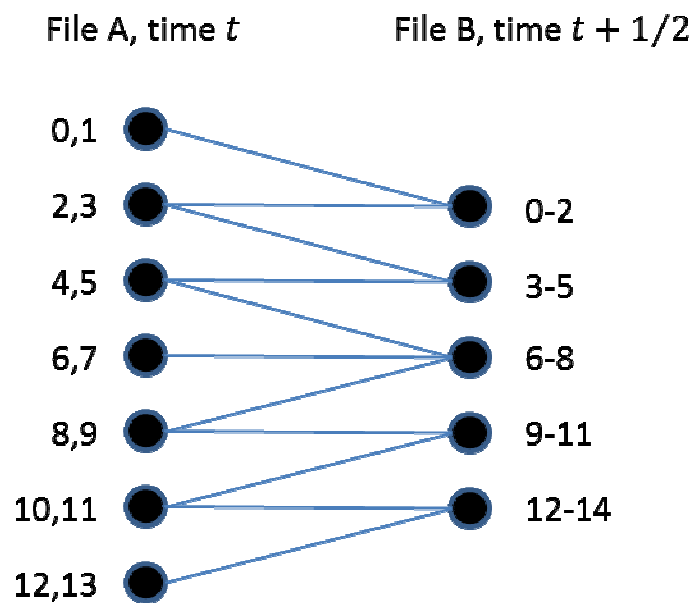


Figure 5. Two age variables and their relationship. One variable specifies age in two-year classes (file A) and the other specifies age in three (or four)-year classes (file B). The timestamps of the files differ by half a year.

Figure 5 shows a digraph that relates the age categories from the two data sets if the reference times are the same.

In practice, the reference times of two matching data sets do not have to be exactly the same. Indeed, it is more likely that they will differ. Moreover the data tend to relate to an interval rather than to a specific point in time. In Figure 5, a connection is made between the age categories if the reference times for the two data sets differ by a half a year. In this case, some people may have turned a year older in the interim period.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Lawler, E. L. (1976), *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart, Winston.

Nemhauser, G. L. and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization*. Wiley, New York.

Papadimitriou, C. H. and Steiglitz, K. (1998), *Combinatorial Optimization: Algorithms and Complexity*. Dover, Mineola (NY).

Willenborg, L. and Heerschap, N. (2012), *Matching*. Contribution to the Methods Series, Statistics Netherlands, The Hague.

Specific section

8. Purpose of the method

The purpose is adding variables to a microdata set Ds-input1 from a second microdata set Ds-input2 for the same objects in both data sets. Records from two microdata sets are combined using a set of common object characteristics. It can be viewed as a special case of the weighted matching method (described in the method module “Micro-Fusion – Weighted Matching of Object Characteristics”), with equal matching weights (e.g., all equal to 1).

9. Recommended use of the method

1. In case common object identifiers of good quality are available in both input data sets object identifiers matching should be used. If this is not the case, the method in the current module could be an option, provided the next point holds.
2. Common object characteristic values of good quality are present in both matching data sets. Also, if similar variables are present in both data sets (with a slightly different domain) this method can be considered, depending on how much the domains differ. Observation errors can occur in the scores of these variables.
3. The data in both data sets should have (approximately) the same reference period. Otherwise there may be too many differences in the object characteristics common to both files to be matched, which has a negative effect on the matching quality.

10. Possible disadvantages of the method

1. In case both matching data sets are big, the method may be too slow. A special variant of the method described in this module, i.e., the one using blocking variables, may be able to do the job and obtain satisfactory results.
2. In case the reference periods for both data sets differ, so may the scores of some objects, due to the dynamics in the population. This may increase the chances for matching errors.

11. Variants of the method

1. The degree of matchability.
 - 1.1 Two records are either matching candidates or they are not. No third option possible.
 - 1.2 Two records are matching candidates, they are not, or they could be. In the case of the distance function d , we may decide for two records a and b as follows:
 - $d(a,b) \leq p$ if a and b are considered matching candidates.
 - $p < d(a,b) \leq q$ if a and b are doubtful matching candidates, and should be inspected by a specialist on the subject who should determine whether or not a and b are matching candidates.
 - $d(a,b) > q$ if a and b are not considered matching candidates.

The parameters p and q , with $p < q$, can be chosen so as to control how many matching candidates have to be inspected. The choices may be guided by the available capacity of specialists who can assess the doubtful cases. The parameters p and q are examples of cut-off values.

1.3 Or we can use a distances. For matching based on an object identifier, it is required that records have exactly the same score on the matching key used. We can also use this matching criterion in case of object characteristics, but the matching method this implies is less attractive here. We can relax this requirement and consider two records as candidate matches if the scores for at least k (parameter to be established) of the maximum n (length of the matching key = number of matching variables in the matching key) are the same. In fact, a metric is used here, the so-called Hamming distance.

2. The number of records in the output dataset:

2.1 Each record of Ds-input1 is part of the output dataset (left outer join), or only matching records occur in the output dataset.

2.2 Each record of Ds-input1 can occur more than once in the output dataset, or can occur at most once in the output dataset.

2.3 Optionally additional output data sets are provided containing the non-matching records of Ds-input1 and Ds-input2.

3. The optional allowance of duplicate records in Ds-input1 or in Ds-input2, as an error type. Otherwise duplicate records are not allowed, as a precondition of the matching method, and a preparatory process step has to delete duplicate records, before starting the matching method.
4. Optionally one or more blocking variables can be used to partition the datasets for matching into manageable subfiles (blocks).

12. Input data

1. Ds-input1. This is the primary input data set. It is a microdata set, to which additional variables will be added.
2. Ds-input2. This input data set that contains variables that will be added to Ds-input1.

13. Logical preconditions

1. Missing values
 1. The object characteristic values used in the matching should not contain missing values.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions

1. A condition for using this method is that common object characteristics are present in both matching data sets, based on which the match can be performed. We also allow the situation that two similar variables have a different domain, for example, with another category division (for example, age in five-year classes in one data set, and in ten-year classes in the other). We also accept that observation errors can occur in the scores of these variables.
2. In practice, the decision to work with a matching method that does not use matching weights is often connected with performance. If the data sets to be matched are large, these methods generally work faster than those with matching weights. However, one should expect the quality of the matches – in terms of missed or missing matches – to be lower in general.
3. Duplicate records are not allowed, unless a variant of the matching method is used that handles the duplicate records.

14. Tuning parameters

1. A metric.
2. Cut-off values for the metric.
3. In case of big files: blocking variables to partition the files into manageable sub files (blocks).
4. In case of the variant with a zone of doubt (see item 11): the parameters p and q .

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. Ds-output1: a microdata set containing all variables of Ds-input1, with variables added from Ds-input2.
2. Optional Ds-output2 containing all non-matching records from Ds-input1.
3. Optional Ds-output3 containing all non-matching records from Ds-input2.

17. Properties of the output data

1. The output data set contains all variables from Ds-input1, but with additional variables from Ds-input2, presumably for the same objects.

18. Unit of input data suitable for the method

Processing full data sets (internally blocking variables can divide a data set into smaller parts).

19. User interaction - not tool specific

1. Before matching the tuning parameters must be set by analysing the results for different values.
2. No user interaction during matching.

3. After matching the number of mismatches must be evaluated, and quality indicators (missing and missed matches).

20. Logging indicators

1. Number of non-matching records from Ds-input1.
2. Number of non-matching records from Ds-input2.
3. Time used.

21. Quality indicators of the output data

1. The number of mismatches or missed matches and the number of missed matches can be used as quality indicators. The quality of the matching method can be assessed based on the inspection of matches of test files. It is a labour intensive job to carry out. One must examine not only the matching candidates and the matches ultimately selected, but also any missed matches under various parameter settings. The quality indicators are influenced by the use of cut-off values and the use of blocking variables to stratify large data sets.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Object Matching (Record Linkage)
2. Micro-Fusion – Probabilistic Record Linkage

24. Related methods described in other modules

1. Micro-Fusion – Object Identifier Matching
2. Micro-Fusion – Weighted Matching of Object Characteristics
3. Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. 5.1 Integrate data

27. Tools that implement the method described in this module

- 1.

28. Process step performed by the method

Adding variables to microdata set

Administrative section

29. Module code

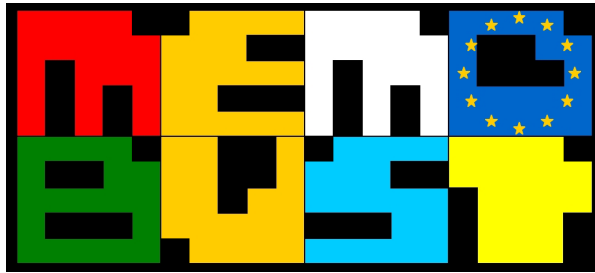
Micro-Fusion-M-Unweighted Matching

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	23-04-2012	first version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.2	02-07-2012	second version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.3	11-07-2013	third version	Leon Willenborg	CBS (Netherlands)
0.4	09-08-2013	revised version (using review comments)	Leon Willenborg	CBS (Netherlands)
0.5	18-11-2013	revised version (using EB review comments)	Leon Willenborg	CBS (Netherlands)
0.5.1	19-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:57



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Weighted Matching of Object Characteristics

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Outline	3
2.2 Preliminaries.....	3
2.3 Calculating matching weights	6
2.4 Quality of matching variables	9
2.5 MC graph with matching weights	9
2.6 Optimisation model	10
3. Preparatory phase	10
4. Examples – not tool specific.....	10
4.1 First example	10
4.2 Second example.....	10
4.3 Third example.....	11
4.4 Fourth example (Soundex algorithm).....	11
4.5 Fifth example (Trigrams)	12
5. Examples – tool specific.....	12
6. Glossary.....	12
7. References	12
Specific section.....	13
Interconnections with other modules.....	16
Administrative section.....	17

General section

1. Summary

Weighted matching is applied to match two data sets with many common units, on common object characteristics. The method is able to value the strength of possible (candidate) matches by using matching weights. Weighted matching can be formulated as an optimisation problem, in which the optimal (weighted) sum of matches is calculated, under certain constraints, such as that each record can appear in at most one match. The goal of the method is to find solutions to such problems, exact ones or good approximations. The reader is advised to consult the theme module “Micro-Fusion – Object Matching (Record Linkage)” prior to reading the present one. Also the reader should refer to the method module “Micro-Fusion – Unweighted Matching of Object Characteristics”, which can be viewed as a special case of the matching method described in the present paper. It also introduces some concepts that are not re-introduced in the current module.

2. General description of the method

2.1 Outline

Various matching methods make use of matching weights. They can be used to differentiate between the potential matches in a matching problem. There is a variety of reasons to work with matching weights: you may want to express that not all of the variables are equally reliable, that is, that they do not have reliable scores. Or you may want to indicate that different objects corresponding with records that are matching candidates demonstrate a certain degree of similarity or dissimilarity. Or you may want to demonstrate that different objects are a certain distance apart, as measured by a certain metric. Or you want to use a probability to show that two objects are probably the same. Then a probability model is needed to quantify differences in scores on the matching key, and the resulting probabilities can be used as matching weights. The method described in this module uses weights to match records on the same object from different data sets. The module draws heavily on Willenborg and Heerschap (2012) to which the interested reader is referred for additional information. It is also the reason why several examples provided are from social statistics, rather than from business statistics. They have been retained as they illustrate certain points clearly. They are also indications that matching is not only used within the business statistics area.

We start with some preliminary material on graphs and metrics in the next subsection.

2.2 Preliminaries

2.2.1 Graphs

Graphs are convenient to describe matching. We only need a few elementary concepts from this area. These are presented in the current section, along with some notation and graphical conventions.

A graph $G = (V, E)$ consists of a finite set of points V , also called nodes or vertices, of which some pairs are connected by lines (E), also called sides, edges or branches. A graph is depicted in Figure 1.

Weights can be assigned to the lines (edges) in the form of real numbers. A graph with weights associated with points or edges is called a weighted graph. In this module the weights are associated

with the edges, and they express the strength of the match between two records. There are different ways to calculate these weights. In some applications, the smaller the weights the more similar the keys of the records are.¹

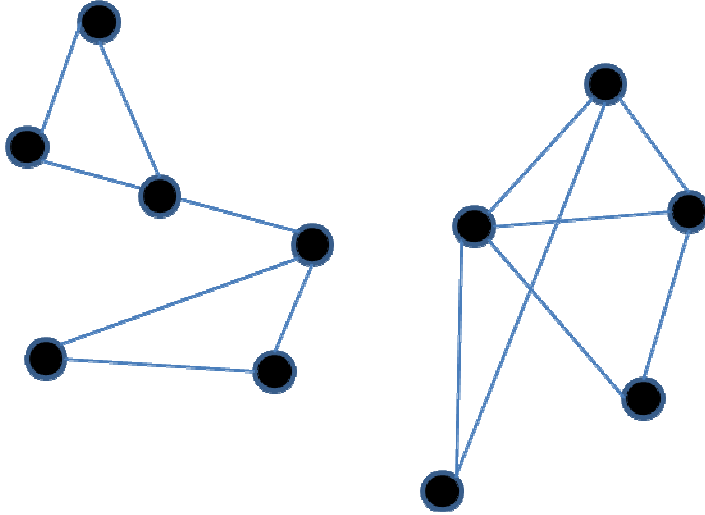


Figure 1. Example of a graph with two connectivity components

A special type of graph is the bipartite graph. See Figure 2. Here, the set of nodes V can be split into two disjoint sets A and B . The edges only connect nodes in A with nodes in B . Bipartite graphs are highly suited to illustrating matching and the theory behind it. Important is the MC graph, the matching candidate graph. This is a bipartite graph that represents the possible matches between records from two files. The edges may or may not be assigned matching weights. A matching candidate graph symbolises part of the constraints that apply for a matching problem.

A *path* in a graph is a succession of nodes arranged in such a way that an edge runs from each node to the following node in the row. Given a graph $G = (V, E)$ where v and w are two points of G , so $v, w \in V$. A path in G from v to w is a sequence v_1, \dots, v_k of points in G , such that:

1. $v_1 = v$,
2. $v_k = w$,
3. $\{v_i, v_{i+1}\} \in E$ for all $i = 1, \dots, k-1$.

If there is a path from v to w in G , then there is also one from w to v (symmetry). If there is a path in G from u to v and from v to w , then there is also one from u to w (transitivity). Here, u , v and w are points in G . For each point v in G , there is – by definition – a path from v to v (reflexivity). In other words, the relationship ‘connected by a path in a given graph’ is an *equivalence relationship* on the set of points of the graph, i.e., a binary relationship that is reflexive, symmetrical and transitive. If there is only one equivalence class for a graph G , it is said to be *connected*. In that case, all pairs of points can therefore be connected with each other via paths in G . If there are two or more equivalence classes for

¹ In other applications it may just be the other way round.

a graph, G is said to be disconnected. In that case, an equivalence class of this relationship corresponds with a *connected component* of G ; this is a connected subgraph of G .

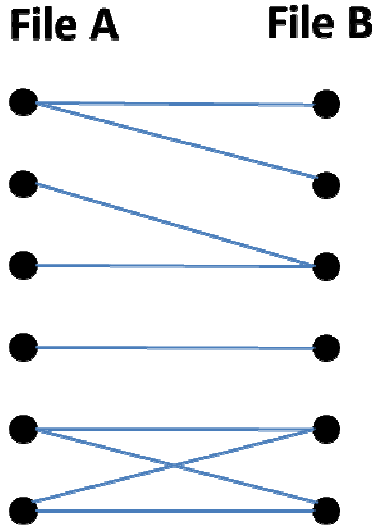


Figure 2. A bipartite graph.

2.2.2 Metrics

Metrics are an important concept for weighted matching. A metric is used in the present module to determine the matching weights. A metric is a function that defines the distance between each pair of elements of a set. Sometimes, it concerns a function that is related to that of a metric, but which deviates on several components from that of a metric. In that case, we have generalised metrics. But we will discuss metrics here first.

We assume a set X for which function $d : X \times X \rightarrow [0, \infty)$ is defined that satisfies a number of conditions:

1. $d(x, y) = 0$ if and only if $x = y$,
2. $d(x, y) = d(y, x)$ for all x, y in X (*symmetry*), and
3. $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z in X (*triangle inequality*).

A non-negative function d that satisfies conditions 1, 2, and 3 is called a metric. The conditions for a metric are not always needed. Replacing them is sometimes necessary and yields alternative distance functions, such as pseudo-metrics or hypermetrics. But in this module we stick to metrics.

In matching and specifically in the comparison of matching keys, this concerns the measurement of the distances between the scores for the matching keys, or, in other words, determining the comparability or non-comparability.

In general, we denote by d , d_H or $d(.,.)$ a metric. We denote the scores on a matching key as a vector $(\alpha_1, \dots, \alpha_n)$ for a matching key (v_1, \dots, v_n) .

A few of the metrics we use here are so special that they are specified separately. The first one is the Hamming distance. Let α and β be two strings of equal length n , viewed as vectors of symbols. The Hamming distance between α and β is defined as:

$$d_H(\alpha, \beta) = d_H((\alpha_1, \dots, \alpha_n), (\beta_1, \dots, \beta_n)) = |\{i \mid \alpha_i \neq \beta_i, 1, \dots, n\}|,$$

i.e., the number of places in which the vectors α and β have different scores. Note that the Hamming distance can be defined for all types of variables.

To illustrate the Hamming distance suppose that there are two matching keys of four alphanumeric figures, ‘1034’ and ‘1135’ respectively. In this case, the Hamming distance is 2, because the figures differ in two places, which are positions 2 and 4. In other words: the smaller the Hamming distance the greater the comparability of the matching keys. The Hamming distance is equal to the number of ‘elementary changes’ that must be made in one key value to obtain the other key value.

The next metric that we want to introduce is the Levenshtein distance. Let α and β be two strings. The Levenshtein distance $d_L(\alpha, \beta)$ counts the minimum number of elementary operations, such as deleting a character, replacing a character, adding a character, that are necessary to transform one string into the other. If another elementary operation is added, namely interchanging neighbouring characters, then we have the so-called Levenshtein-Damerau-distance. The Levenshtein distance and the Levenshtein-Damerau distance are examples of metrics that are specifically designed for strings. There are other metrics of this type, specifically tailored to certain types of variables.

Consider the words ‘apple’ and ‘pear’. Their Levenshtein distance is 4. To see this consider the following chain of elementary changes: *apple* \rightarrow *ppl*e \rightarrow *pele* \rightarrow *peae* \rightarrow *pear*. In less than 4 steps a transformation from ‘apple’ to ‘pear’ using elementary transformation associated with this distance function are not possible, as the reader is invited to check. The advantage of the Levenshtein distance, compared to the Hamming distance, is that the distance of strings of different lengths can be calculated.

More examples of metrics for strings and relevant for matching can be found in Section 4.

2.3 Calculating matching weights

There are different ways to determine matching weights that can be used in a matching problem. We will discuss several here. The list is not exhaustive, but it does provide several important examples. These matching weights are used for matching if the information about the ‘matching candidacy’ of two records is not represented in ‘either/or’ form (matching candidate? ‘yes’ or ‘no’), but with more differentiation. The extent to which two records match can be expressed in a matching weight.

In the discussion in the sections below, we look at two data sets, A and B, that contain records, for which there are common matching variables v_1, \dots, v_n that together form the matching key, based on which the records in the two data sets are matched. Weights are used for candidate matches, to indicate the ‘strength’ of a match. See Figure 3 for such a situation.

For more information and examples on MC-graphs, the interested reader should consult the method module “Micro-Fusion – Unweighted Matching of Object Characteristics”.

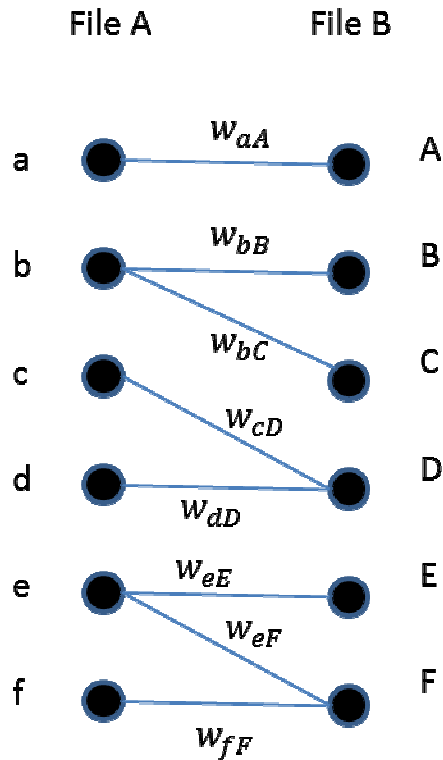


Figure 3. MC-graph with matching weights

2.3.1 Using metrics

For matching, it is important to find suitable metrics for each of the variables in a primary or secondary matching key, or rather for each type of variables. The following variables may occur in a secondary matching key: names (first names, surnames, enterprise names, street names, city names, etc.), time indications (dates of birth, ages at a certain reference time), economic activity, etc. Finding suitable metrics to be used for the secondary matching keys can be seen as a separate subfield in matching.

We take another look at strings, as they are quite important in matching. There are several aspects to strings when it comes to measuring the distance between them. This depends on how we look at them: literally, as objects built from an alphabet, or looking at other aspects such as their pronunciation (phonetics) or their meaning.

A metric can be used to calculate matching weights. These matching weights can be used to express the strength of a candidate match. We should add that, in practice, it is necessary to work with cut-off values: matches that are too weak in terms of the associated matching weight are not considered to be matching candidates. The trick is to properly establish these cut-off values: on the one hand too many irrelevant matches should be avoided but not many correct matches should be missed. In practice, this requires experimentation with various settings of the cut-off values.

All the considerations to use matching weights must be derived from the processes or mechanisms that (may) have caused differences in the data. This could be writing mistakes ('Dickson' instead of 'Dixon'), alternative designations ('Main Str.' instead of 'Main Street'), use of synonyms ('shipping' instead of 'transporting'). It is therefore important to have thorough knowledge of the way in which

the data sets to be matched have been compiled. In addition, it is possible that not exactly the same matching variables will be used in the two data sets, or that the scores do not relate to the same moment in time. As a result, the attributes of entities (e.g., businesses, enterprises, etc.) could have changed.

2.3.2 *Using probabilities*

Matching weights can also be based on probability models. Stochastic methods can enter into matching for different reasons. We offer the following reasons:

1. Errors can occur in the secondary matching keys. The errors can be present for various reasons. An answer to a question in a survey could have been understood incorrectly and therefore answered incorrectly by the respondent in question; a given answer could have been incorrectly processed, for example, keyed in wrongly; errors could have been made in the coding of answers, etc. This type of error is often referred to as a non-sampling error. The first step would be to identify and model all major sources of errors using probability models. These models can then be used to calculate the probabilities that two scores match based on corresponding object characteristics from two matching data sets.
2. The reference times of the two matching data sets differ to such an extent that the effects of the dynamics of the population are noticeable on the units contained therein: values of certain scores could have been changed for some units. An enterprise could have merged, split or gone bankrupt. Therefore, if the reference times differ significantly from one another, it is not self-evident that the units and/or their scores on object characteristic variables would have remained unchanged.
3. Some comparable matching variables are not defined exactly the same way in the two data sets. The associated question can be different, or the position in the questionnaire could have been changed, or the value range of comparable variables may differ slightly. In that case, it may sometimes be unclear which scores correspond with one another. Suppose {20,21} is an age class in one matching file and 11 - 20 and 21 - 30 are age classes in the other file. The 20 and 21-year-olds are in the same age group in the first matching file, but they are in two different age categories in the second. We can also estimate which part of the people in the category (20,21) in the first file will end up in the age category 11-20 and which part will be in the age category 21-30 in the second file: $\frac{n_{20}}{n_{20} + n_{21}}$, and $\frac{n_{21}}{n_{20} + n_{21}}$ respectively, where n_{20} is the number of 20-year-olds on the measurement date and n_{21} the number of 21-year-olds at that point in time.

In practice, combinations of these causes of differences often occur. Data sets can have different reference times, there may be processing errors in the data, and the units may not be exactly comparable. Section 4 presents examples of a situation as in point 3 above, and an example with a combination of points 2 and 3 above (variables with deviating value ranges and different reference times).

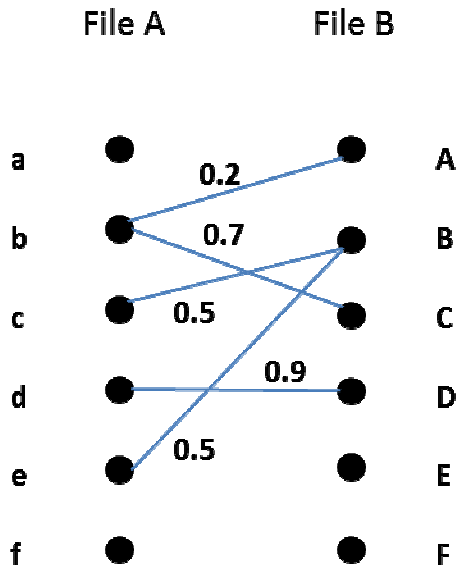


Figure 4. MC-digraph with probabilities as matching weights.

2.4 Quality of matching variables

In practice, based on the quality of the scores, we will want to differentiate between the different matching variables in the matching key. Some variables will have more reliable scores than others, and we will want to take this effect into account when determining the overall matching weight.

We consider this ‘quality weight’ as a subjective weight that the person performing the matching establishes based on his/her knowledge and experience with the different variables in the matching key. It is possible that experiments must first take place before a good choice can be made about these weights. These weights only have meaning in terms of the relationships between them, not in an absolute sense. Users can express the relative importance of a variable for the multivariate distance function. In this way, they can influence the effect of a certain variable in the total. If the variable has been reliably measured, then a relatively high weight is needed. If it is a variable with relatively more errors than the other variables in the matching key, then this variable should be given a lower weight.

For that matter, it is also possible to express the difference in the quality of matching variables in a different way, for example, when matching, by going through the scores in the order of the quality of the matching variables (from high to low), and then accepting certain deviations in the scores with increasing tolerance.

2.5 MC graph with matching weights

Once we have selected a method to determine matching weights, we can start calculating an MC graph with matching weights. We may have to use a cut-off value so that we do not have to include candidate matches of two records with a matching weight that is too low (they will not become edges in the MC graph).

2.6 Optimisation model

Once the MC graph with matching weights has been calculated we are almost ready to calculate the matching. What is needed in addition is a specification of an object function, and matching conditions. The object function could be the sum of the weights associated with the edges chosen for a particular match. If the weights are larger in case a match is stronger, the goal would be to find matches among the candidate matches that maximise the sum of the associated weights. The matching conditions yield the constraint for the matching. A common requirement is that a record can be in no more than one match.

The model that we get in this way is a well-known one in combinatorial optimisation, called bipartite matching. It is discussed in books like Lawler (1976, Ch.5), Papadimitriou and Steiglitz (1998, Ch.11) or Nemhauser and Wolsey (1988, Ch. III.2), to which the interested reader is kindly referred.

3. Preparatory phase

The object characteristics common to both data sets to be matched are identified. It has to be decided if they are suitable for this type of matching; the number of potential matches should not be too big. A suitable metric for these variables should be found, as well as a suitable cut-off value. This requires some experimenting: with the cut-off value the number of potential matches can be controlled. In case the matching data sets are big special measures should be taken, such as blocking to create a manageable matching problem.

Now candidate matches can be found, from which the matches are to be calculated.

4. Examples – not tool specific

4.1 First example

Given a matching key that consists of n variables that are all object characteristics. For the i^{th} variable we have a metric d_i . For the entire matching key, we can define a metric $d = \sum_i w_i d_i$, with weights w_i , $w_i > 0$, $i = 1, \dots, n$.

4.2 Second example

Let δ be a 0-1-indicator function, defined as follows: $\delta(a, b) = 0$ if $a = b$ and $\delta(a, b) = 1$ if $a \neq b$, for scores a, b for a matching or other variable. For score vectors α, β , we define

$$\Delta(\alpha, \beta) = (\delta(\alpha_1, \beta_1), \dots, \delta(\alpha_n, \beta_n)) \in \{0, 1\}^n.$$

This indicator vector plays a central role in the method described in Fellegi and Sunter (1969). See also the method module “Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage” in the present handbook. Note that

$$d_H(\alpha, \beta) = \sum_{i=1}^n \delta(\alpha_i, \beta_i).$$

4.3 Third example

Consider a name variable, such as first name, surname, business name, street name, place name, etc. There are several ways in which the extent to which the distance of these names, or what they stand for, can be expressed:

- **String as a sequence of symbols.** Here, you may want to express the extent to which two surnames differ from one another. The difference between ‘Jansen’ and ‘Janssen’ is smaller than the difference between ‘Jansen’ and ‘Todd’ (Jansen→Tansen→ Tonsen→ Todsen→ Todden→Todde→Todd). This concerns only the spelling of the names: the letters that are present and their order of occurrence. This can be quantified using a metric (the Levenshtein metric or Levenshtein-Damerau metric, for instance).
- **Meaning of a string.** The words ‘teacher’ and ‘instructor’ are very different from one another as strings, but in terms of meaning (concepts), they are very close, and could even be considered being synonyms.
- **Pronunciation of a string** The distance concept here relates to the meaning (semantics) associated with strings, not the way they are composed of characters from some alphabet. A similar difference is obtained if we consider pronunciation (say, in English) of strings, ‘Dixon’ and ‘Dickson’ are pronounced the same. Phonetically these strings are equal.

The last two cases are comparable, in the sense that we do not measure the distance of the literal strings, but on some associated attribute (interpretation / meaning or pronunciation).

Let d be a metric on S , and D a metric on T . Then $d(s, t)$ measures the distance between the strings s and t and $D(f(s), f(t))$ the distance between the meaning of s and t , or their pronunciation. This would be a distance between two points in a classification (a tree), which could, e.g., be the length of the shortest path (in the tree) connecting these points.

4.4 Fourth example (Soundex algorithm)

Comparing strings taking the phonetic characteristics of English into account could be done by using a so-called Soundex algorithm. This algorithm maps alphanumeric strings to Soundex strings (consisting of a letter followed by three numerical digits): the letter is the first letter of the name, and the digits encode the remaining consonants. Similar sounding consonants share the same digit.

A string (name) is mapped to a Soundex string using the following rules:

1. Retain the first letter of the name; drop all occurrences of a, e, I, o, u, y, h, w.
2. Replace consonants with digits as follows (after the first letter)
 - b, f, p, v → 1
 - c, g, j, k, q, s, x, z → 2
 - d, t → 3
 - l → 4
 - m, n → 5

- $r \rightarrow 6$
3. If two or more letters with the same number are adjacent in the original name (before step 1), only retain the first letter; also two letters with the same number separated by 'h' or 'w' are coded as a single number, whereas such letters separated by a vowel are coded twice. This rule also applies to the first letter.
 4. Iterate the previous step until you have one letter and three numbers. If you have too few letters in your word that you can't assign three numbers, append with zeros until there are three numbers. If you have more than 3 letters, just retain the first 3 numbers.

Applying these rule to 'Rupert' and 'Robert' yields the same Soundex string R163.

4.5 Fifth example (Trigrams)

The names Hendriks, Hendricks, Hendrickx, Hendriksz, Hendrikx, Hendrix are all pronounced the same (in Dutch, at least), while all are different as strings. In this example we look at two of them and see how they differ if we look at trigrams. We consider two of them, 'Hendriksz' and 'Hendrix'. For both names we consider the extended versions, which we obtain by adding a space ('_') at the start and end of each name. We then get: '_Hendriksz_' and '_Hendrix_'. The trigrams for the first string are (we write everything in lower case letters): (_he, hen, end, ndr, dri, rik, iks, ksz, sz_) and for the second string (_he, hen, end, ndr, dri, rix, ix_) . They have 5 trigrams in common, 5 out of 9 for the first string and 5 out of 7 for the second one.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

- Fellegi, I. P. and Sunter, A. B. (1969), A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1200.
- Lawler, E. L. (1976), *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart, Winston.
- Nemhauser, G. L. and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization*. Wiley.
- Papadimitriou, C. H. and Steiglitz, K. (1998), *Combinatorial Optimization: Algorithms and Complexity*. Dover, Mineola (NY).
- Willenborg, L. and Heerschap, N. (2012), *Matching*. Contribution to Methods Series. Statistics Netherlands, The Hague.

Specific section

8. Purpose of the method

The purpose is adding variables to a microdata set Ds-input1 from a second microdata set Ds-input2 for the same objects in both data sets. Records from two microdata sets are combined using a set of common object characteristics. It can be viewed as a more general case of the unweighted matching method (described in the method module “Micro-Fusion – Unweighted Matching of Object Characteristics”).

9. Recommended use of the method

1. In case object identifiers of good quality in both matching data sets are not available weighted matching may be considered as an option, under certain conditions.
2. Common object characteristic values of good quality should be present in both matching data sets. Also if similar variables are present in both data sets (with a different, but almost the same domain) this method can be considered, depending on how much the domains differ. Observation errors can occur in the scores of these variables.
3. The unweighted matching method can be characterised as being ‘black and white’: two records are either matching candidates or they are not. There is no room for any differentiation. However, there are situations where this is desirable. Some spelling mistakes or alternative designations are more likely than others.
4. In addition, it is possible that not exactly the same matching variables have been used in the two data sets, or that the scores do not relate to the same moment in time. As a result, the attributes of an entity (individual, business, etc.) could have changed. Also in this case the method aims at matching records for the same object.

10. Possible disadvantages of the method

1. It can be too slow, as compared to unweighted matching.
2. Values of tuning parameters require some experimentation or specialist knowledge.

11. Variants of the method

The text in the general section of the module places the emphasis on the basic variant for matching with matching weights, where the matches are 1:1. As stated earlier, there are also situations in which 1:n, m:1 and even n:m matches are possible. This is the case for composite units such as businesses which, over time, can split or merge into other units. Formally, this means that the conditions under which matches are possible must be adapted. Also they do not relate to the same units, but to combinations of units that produce comparable entities.

In the discussion we have so far assumed that all the scores on object characteristics are present. In practice, however, this is not always necessarily the case, and scores can also be erroneously missing. Calculating matching weights is more difficult in this situation, because the missing values cannot just be omitted: they must be replaced by stochastic variables, with a known assumed distribution. In such cases, the unknown parameter values must be estimated using, for example, the EM algorithm. For

information about the EM algorithm, see Wikipedia (http://en.wikipedia.org/wiki/EM_algorithm) and the references provided there.

We can summarise the available variants as follows:

1. Number of records in the output dataset:
 - 1.1 Each record of Ds-input1 is part of the output dataset (left outer join), or only matching records occur in the output dataset.
 - 1.2 Each record of Ds-input1 can occur more than once in the output dataset, or can occur at most once in the output dataset.
2. Duplicate records may be present in Ds-input1 or in Ds-input2.
3. One or more blocking variables can be used to divide the datasets for matching.
4. Missings in the object characteristics may be present in the input data sets.

12. Input data

1. Ds-input1. This is the primary input data set. It is a microdata set, to which additional variables will be added.
2. Ds-input2. This auxiliary input data set contains the variables that will be added to Ds-input1.

13. Logical preconditions

1. Missing values
 1. The object characteristic values used in the matching may contain missing values, but not too many, as they negatively influence the matching performance.
2. Erroneous values
 1. Errors in the object characteristic values are allowed, but it should still be possible to use them for matching. With certain assumptions on the cause of the errors, they must be usable owing to a small distance to the correct values.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. Enough object characteristic variables must be available in both input data sets to identify objects in the population. Otherwise more than one record with smallest distance remains, and an arbitrary choice should be made from them, with a high risk on Type I errors.

14. Tuning parameters

1. Optimisation function.
2. Matching weight.
3. Cut-off values.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. Ds-output1: a microdata set containing all variables of Ds-input1, with variables added from Ds-input2.
2. Optional Ds-output2 containing all non-matching records from Ds-input1.
3. Optional Ds-output3 containing all non-matching records from Ds-input2.

17. Properties of the output data

1. The output data set contains all variables from Ds-input1, but with additional variables from Ds-input2, presumably for the same objects.

18. Unit of input data suitable for the method

Processing full data sets (internally blocking variables can divide a data set in smaller parts).

19. User interaction - not tool specific

1. Before matching the tuning parameters must be set by analysing the results for different values.
2. No user interaction during matching.
3. After matching the number of mismatches must be evaluated, and quality indicators (Type1 and Type 2 errors).

20. Logging indicators

1. Number of non-matching records from Ds-input1.
2. Number of non-matching records from Ds-input2.
3. Time used.

21. Quality indicators of the output data

1. The number of mismatches or missed matches and the number of missed matches can be used as quality indicators. The quality of the matching method can be assessed based on the inspection of matches of test files. It is a labour intensive job to carry out. You must examine not only the matching candidates and the matches ultimately selected, but also any missed matches under various parameter settings. The quality indicators are influenced by the way that the weights are calculated, the use of cut-off values and the use of blocking variables to stratify large data sets.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Object Matching (Record Linkage)
2. Micro-Fusion – Probabilistic Record Linkage

24. Related methods described in other modules

1. Micro-Fusion – Object Identifier Matching
2. Micro-Fusion – Unweighted Matching of Object Characteristics
3. Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. 5.1 Integrate data

27. Tools that implement the method described in this module

- 1.

28. Process step performed by the method

Adding variables to microdata set

Administrative section

29. Module code

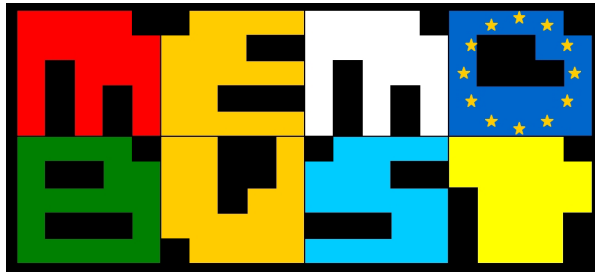
Micro-Fusion-M-Weighted Matching

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	21-04-2012	first version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.2	02-07-2012	second version	Leon Willenborg, Rob van de Laar	CBS (Netherlands)
0.3	11-07-2013	third version	Leon Willenborg	CBS (Netherlands)
0.4	09-08-2013	revised version (using review comments)	Leon Willenborg	CBS (Netherlands)
0.5	17-11-2013	revised version (using EB review comments)	Leon Willenborg	CBS (Netherlands)
0.5.1	19-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:58



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Probabilistic Record Linkage

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Search space reduction	5
2.2 The matching variables.....	5
2.3 The comparison functions	6
2.4 The decision rule and parameters estimation	6
2.5 Alternative probabilistic record linkage methods.....	7
2.6 Record linkage quality.....	8
3. Design issues	9
4. Available software tools.....	11
5. Decision tree of methods	13
6. Glossary.....	13
7. References	13
Interconnections with other modules.....	16
Administrative section.....	17

General section

1. Summary

In this section the problem of probabilistic record linkage is explored. It can be also viewed as the weighted matching in case of an explicit use of probabilities. Generally speaking record linkage (or object matching, see also module on object matching) can be defined as the set of methods and practices aiming at accurately and quickly identify if two or more records, stored in sources of various type, represent or not the same real world entity. As usually data sources are hard to integrate due to errors or lacking information in the record identifiers, record linkage can be seen as a complex process consisting of several phases involving different knowledge areas. In research literature a distinction between deterministic (matching identifiers) and probabilistic approaches (matching with matching weights) is often made, where the former is associated with the use of formal decision rules while the latter makes an explicit use of probabilities for deciding when a given pair of records is actually a match but a clear separation between the two approaches is very difficult.

Compared with the deterministic approach, the probabilistic one can solve problems caused by bad quality data and can be helpful when differently spelled, swapped or misreported variables are stored in the two data files; the attention in this section is only devoted to the probabilistic record linkage approach which allows also to evaluate the linkage errors, calculating the likelihood of the correct match.

Generally speaking, the deterministic and the probabilistic approaches can be combined in a two-step process: firstly the deterministic method can be performed on the high quality variables then the probabilistic approach can be adopted on the residuals, the units not linked in the first step; however the joint use of the two techniques depends on the aims of the whole linkage project.

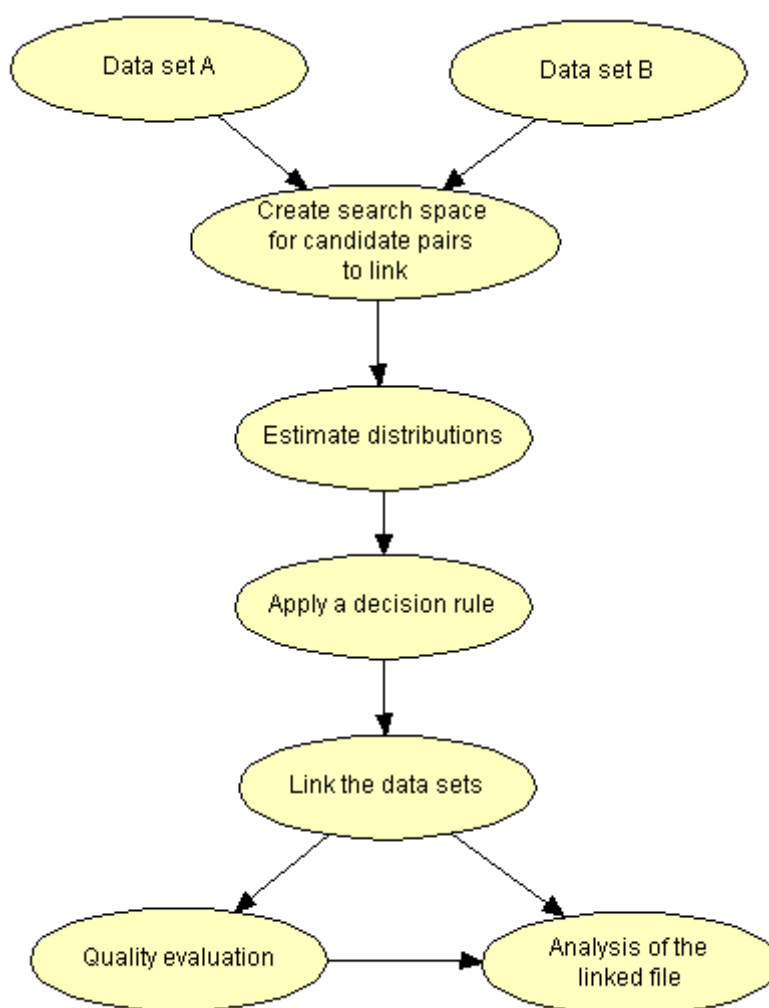
2. General description

Record linkage is widely performed in order to enrich, update or improve the information stored in different sources; to create a sampling list; to study the relationship among variables reported in different sources; to eliminate duplicates within a data frame; to assess the disclosure risk when releasing microdata files, etc. In official statistics, the advantages, in terms of quality and costs, due to the combined use of administrative data and sample surveys strongly encourage the researchers to the investigation of new methodologies and instruments to deal with record linkage projects and to identify quickly and accurately units across various sources. Since the earliest contributions to modern record linkage, dated back to Newcombe et al. (1959) and to Fellegi and Sunter (1969) where a more general and formal definition of the problem is given, there has been a proliferation of different approaches, that make use also of techniques based on data mining, machine learning, equational theory.

According to some authors (e.g., Statistic Canada) deterministic record linkage is defined just as the method that detects links if and only if there is a full agreement of unique identifiers or a set of common identifiers, the matching variables. Other authors backed up that in deterministic record linkage a pair is a link also if it satisfied some specific criteria a priori defined; actually not only the matching variables must be chosen and combined but also a threshold has to be fixed in order to establish whether a pair should be considered a link or not. Deterministic record linkage can be

adopted, instead of probabilistic method, in presence of error-free unique identifiers (such a fiscal code) or when matching variables with high quality and discriminating power are available and can be combined so as to establish the pairs link status; in this case the deterministic approach is very fast and effective and its adoption is appropriate. From the other side, the rule definition is strictly dependent on the data and on the knowledge of the practitioners. Moreover, due to the importance of the matching variable quality, in the deterministic procedure, some links can be missed due to presence of errors or missing values in the matching variables; so the choice between the deterministic and probabilistic methods must take into account “the availability, the stability and the uniqueness of the variables in the files” (Gill, 2001). It is important also to underline that, in a deterministic context, the linkage quality can be assessed only by means of re-linkage procedures or accurate and expensive clerical reviews.

Probabilistic record linkage is a complex procedure that could be decomposed in different steps. For each step we can adopt different techniques. The following workflow has been taken from the WP1 of the ESSnet on ISAD (integration of surveys and administrative data), Section 1.2 (Cibella et al., 2008a) and represents the whole record linkage process:



In a linking process of two already harmonised data sets, namely A and B of size N_A and N_B respectively, let us consider the search space $\Omega = \{(a,b), a \in A \text{ and } b \in B\}$ of size $N=N_A \times N_B$. The linkage between A and B can be defined as the problem of classifying the pairs that belong to Ω in two subsets M and U independent and mutually exclusive, such that:

M is the set of matches ($a=b$)

U is the set of non-matches ($a \neq b$)

2.1 Search space reduction

When dealing with large datasets, comparing all the pairs ($a; b$), a belonging to A and b belonging to B, in the cross product is almost impracticable and this causes computational and statistical problems. To reduce this complexity it is necessary to reduce the number of pairs ($a; b$) to be compared. There are many different techniques that can be applied to reduce the search space; blocking and sorted neighbourhood are the two main methods. Blocking consists of partitioning the two sets into blocks and of considering linkable only records within each block. The partition is made through blocking keys; two records belong to the same block if all the blocking keys are equal or if a comparison function applied to the blocking keys of the two records gives the same result. Sorted neighbourhood sorts the two input files on a blocking key and searches possible matching records only inside a window of a fixed dimension which slides on the two ordered record sets.

2.2 The matching variables

Starting from the reduced search space, we can apply different decision models that enable to classify pairs into M, the set of matches and U, the set of non-matches.

In this section the probabilistic approach is formalised according to the Fellegi and Sunter theory which is described in details in the module “Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage”. The method requires an estimation of the model parameters that can be performed via the EM algorithm, Bayesian methods, etc.

In order to classify the pairs, some k common identifiers, either quantitative or qualitative, called matching variables,

$$\mathbf{X}_1^A \quad \mathbf{X}_2^A \quad \dots \quad \mathbf{X}_K^A ; \quad \mathbf{X}_1^B \quad \mathbf{X}_2^B \quad \dots \quad \mathbf{X}_K^B$$

have to be chosen so that, for each pairs, a comparison vector $\gamma = \{\gamma_1, \gamma_2, \dots, \gamma_K\}$ can be defined, where

$$_{(a,b)} \gamma_k = \begin{cases} 1 & \text{if } X_k^A = X_k^B \\ 0 & \text{otherwise} \end{cases}$$

It is important to choose matching variables that are as suitable as possible for the considered linking process. The matching attributes are generally chosen by a domain expert. If unique identifiers are available in the linkable data sources, the easiest and most efficient way is to use these ones as link variables; but very strict controls need to be made in case of using numeric identifiers alone. Variables like name, surname, address, date of birth, can be used jointly instead of using each of them separately; in such a way, one can overcome problems like the wide variations of the name spelling or the changes in surname depending on the variability of the marital status. It is evident that the more

heterogeneous are the items of a variable, the higher is its identification power; moreover, if missing cases are relevant in a field it is not useful to choose it as a matching variable.

2.3 The comparison functions

The comparison functions are used to compute the distance between records compared on the values of the chosen matching variables. Some of the most common comparison functions are (for a review, see Koudas and Srivastava, 2005):

- a) equality that returns 1 if two strings fully agree, 0 otherwise;
- b) edit distance that returns the minimum cost in terms of insertion, deletions and substitutions needed to transform a string of one record into the corresponding string of the compared record;
- c) Jaro counts the number of common characters and the number of transpositions of characters (same character with a different position in the string) between two strings;
- d) Hamming Distance that computes the number of different digits between two numbers;
- e) Smith-Waterman that uses dynamic programming to find the minimum cost to convert one string into the corresponding string of the compared record; the parameters of this algorithm are the insertions cost, deletions cost and transposition cost;
- f) TF-IDF that is used to match strings in a document. It assigns high weights to frequent tokens in the document and low weights to tokens that are also frequent in other documents.

2.4 The decision rule and parameters estimation

Following Fellegi and Sunter (1969), the ratio

$$r = \frac{P(\gamma | (a,b) \in M)}{P(\gamma | (a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)}$$

between the probabilities of γ given the pair (a,b) membership either to the subset M or U is used so as classifying the pair. Fellegi and Sunter proposed an equation system to achieve the explicit formulas for the estimates of $m(g)$ and $u(g)$ when the matching variables are at most three (see the method module “Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage” for details).

Once the probabilities m and u are estimated, all the pairs can be ranked according to their ratio $r=m/u$ in order to detect which pairs are to be matched by means of a decision rule based on two thresholds T_m and T_u ($T_m > T_u$)

$$\begin{aligned} r_{(a,b)} > T_m &\Rightarrow (a,b) \in M^* \\ T_m \geq r_{(a,b)} \geq T_u &\Rightarrow (a,b) \in Q \\ r_{(a,b)} < T_u &\Rightarrow (a,b) \in U^* \end{aligned}$$

- those pairs for which r is greater than the upper threshold value can be considered as linked
- those pairs for which r is smaller than the lower threshold value can be considered as not-linked

The thresholds are chosen so as to minimise two types of possible errors: false matches (FMR, or mismatch, false positive match, Type I error, see module on object matching) and false non-matches

(FNMR, missed match, false negative match, Type II error) that refers respectively, as stated above, to the matched records which do not represent the same entity and to the unmatched records not correctly classified, that imply truly matched entities were not linked.

The Fellegi and Sunter approach is heavily dependent on the accuracy of $m(\gamma)$ and $u(\gamma)$ estimates. Misspecifications in the model assumptions, lack of information and other problems can cause a loss of accuracy in the estimates and, as a consequence, an increase of both false matches and non-matches.

Armstrong and Mayda (1993) assume that the frequency distribution of the observed patterns γ is a mixture of the matches $m(\gamma)$ and non-matches $u(\gamma)$ distributions

$$\begin{aligned} P(\gamma) &= P(\gamma|(a,b) \in M)P((a,b) \in M) + P(\gamma|(a,b) \in U)P((a,b) \in U) \\ &= m(\gamma) \cdot p + u(\gamma) \cdot (1-p) \end{aligned}$$

where $p=P(M)$. The latent variable C denotes the unknown linkage status and is equal to 1 in case of a match, with the probability p , so the joint distribution of the observations γ and the latent variable $C=c$ ($c=(0,1)$) is given by:

$$P(C = c, \gamma) = [pm(\gamma)]^c [(1-p)u(\gamma)]^{1-c}. \quad (1)$$

Since vector C is not directly measurable, the maximum likelihood estimates of parameters $m_k(\gamma)$, $u_k(\gamma)$ and p can be obtained through EM algorithm (Dempster et al., 1977) as proposed in Jaro (1989). A simplification of the estimates, which is often made in order to keep easier the parameters estimation, is the so called local independence assumption, where r is written as

$$r = \frac{P(\gamma|M)}{P(\gamma|U)} = \prod_{k=1}^K \frac{P(\gamma_k|M)}{P(\gamma_k|U)} = \prod_{k=1}^K \frac{m_k}{u_k}.$$

Even local independency assumption works well in most of the practical application, it cannot be sure that this hypothesis is automatically satisfied. Some authors (Winkler, 1989, and Thibaudeau, 1989) extend the standard approach by means of log-linear models with latent variable by introducing appropriate constraints on parameters so to overcome to some extent local independence assumption. In these cases, however, it is not sure if the best model in terms of fitting could be also considered as the most accurate in terms of linkage results and errors.

2.5 *Alternative probabilistic record linkage methods*

Also other approaches could be considered in the estimation of parameters (the following description has been taken from the WP1 of the ESSnet on ISAD, Section 1.5 (Cibella et al., 2008a)):

The Bayesian approaches – Fortini et al. (2001, 2002) look at the status of each pair (match and non-match) as the parameter of interest. For this parameter and for the parameters of the latent variables that generate matches and non-matches they define natural prior distributions. The Bayesian approach consists in marginalising the posterior distribution of all these parameters with respect to the parameters of the comparison variables (nuisance parameters). The result is a function of the status of the different pairs that can be analysed for finding the most probable configuration of matched and unmatched pairs.

Iterative approaches – Larsen and Rubin (2001) define an iterative approach which alternates a model based approach and clerical review for lowering as much as possible the number of records whose status is uncertain. Usually, models are estimated among the set of fixed loglinear models, through parameter estimation computed with the EM algorithm and comparisons with “semi-empirical” probabilities by means of the Kullback-Leibler distance.

Other approaches – Different papers do not estimate the distributions of the comparison variables on the data sets to link. In fact, they use ad hoc data sets or training sets. In this last case, it is possible to use comparison variables more informative than the traditional dichotomous ones. For instance, a remarkable approach is considered in Copas and Hilton (1990), where comparison variables are defined as the pair of categories of each key variable observed in two files to match for matched pairs (i.e., comparison variables report possible classification errors in one of the two files to match). Unmatched pairs are such that each component of the pair is independent of the other. In order to estimate the distribution of comparison variables for matched pairs, Copas and Hilton need a training set. They estimate model parameters for different models, corresponding to different classification error models.

2.6 *Record linkage quality*

As not every record matched in the linkage process refers to the same identity, at the end of the record linkage process is really important to assess the “quality” of the procedure establishing whether a match is a “true one” or not. In other words, during a linkage project is necessary to classify records as true link or true non link, minimising, according to the Fellegi and Sunter theory, the two types of possible errors: false matches and false non-matches that refers respectively, as stated above, to the matched records which do not represent the same entity and to the unmatched records not correctly classified, that imply truly matched entities were not linked. False non-matches of matching cases are the most critical ones because of the difficulty of checking and detecting them. In general, it’s not easy to find automatic procedures to estimate these types of errors so as to evaluate the quality of record linkage procedures. The same accuracy indicators are also used in the research field of information retrieval, although they are usually named precision and recall and can be evaluated even if the linkage procedure is performed through techniques different from the probabilistic one, as for instance supervised or unsupervised machine learning (Elfeky et al., 2003).

Errors can also be introduced by the choices that are made in the matching process itself. For instance, an incorrect or overly limited matching key may be used, the way in which the weights are calculated may be incorrect, or the cut-off values against which the weights are set off may lead to matching errors.

Also the time consumed by software programmes and by the number of records that require manual review could be considered additional performance criteria for the process (see the WP1 of the ESSnet on ISAD, Section 1.7 for details (Cibella et al., 2008a)) or also, as stated in the module “Micro-Fusion – Object Matching (Record Linkage)”, all the choices that are made in the matching process itself could have an impact on the record linkage quality (e.g., an incorrect or overly limited matching key).

The final step of the whole record linkage process is devoted to the subsequent studies of the linked data set, taking in mind that this file can contain matching errors and all the derived analysis could be affected by the two types of errors: the percentage of incorrect acceptance of false matches and, on the

other hand, the incorrect rejection of true matches. Record linkage procedures must deal with the existing trade-off between these two errors and/or measure the effects on the parameter estimates of the models that are associated to the obtained files.

The following description has been selected from Section 1.8 of the WP1 of the ESSnet on ISAD (Cibella et al., 2008a): different approaches have tackled the problem, the first due to Neter et al. (1965) that has studied bias in the estimates of response errors when the results of a survey are partially improved through record checks, and raises awareness of substantial effects in the results with relatively small errors in the matching process.

Scheuren and Oh (1975) focus on different problems noticed in a large-scale matching task as a Census - Social Security match through Social Security Number (SSN)¹. They focus attention to the impact of different decision rules on mismatching and erroneous no matching. Furthermore they point out the constraints to develop an appropriate comparison vector when statistical purposes differ from administrative aims that generated the file and that regulate its maintenance. Nevertheless their approach does not offer general criteria to estimate the parameters of the distributions, as $m(\gamma_{ab})$ and $u(\gamma_{ab})$. Their approach is to select a sample of records, manually check their status of matched and unmatched pair, and estimate those parameters from the observed proportions.

Some more complete methodologies have been developed by Scheuren and Winkler (1993, 1997) through recursive processes of data editing and imputation.

Larsen (2004) and Lahiri and Larsen (2000 and 2005) have widely discussed the use of the former methodology for mixture models, trying to improve the estimates of the probability that a pair of records is actually a match. Those estimates can be found through maximum likelihood or Bayesian analysis, and then adjust the regression models by an alternative to the bias correction method used in Scheuren and Winkler.

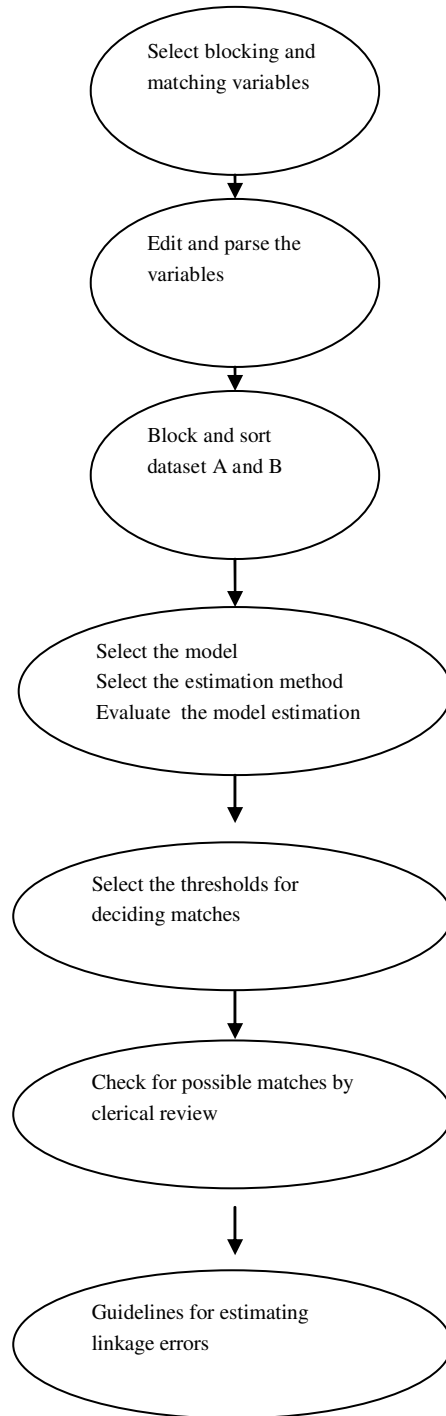
Additionally, Liseo and Tancredi (2004) develop a brief regression analysis based on a Bayesian approach to record linkage while Winkler (2006) suggests that the use of a regression adjustment to improve matching can be done by means of identifying variables that are not strictly the same, but actually include the same information from different points of view.

3. Design issues

This present section has been taken from the WP2 of the ESSnet on ISAD (integration of surveys and administrative data), Section 2.1 (Cibella et al., 2008b).

Record linkage is a complex procedure that can be decomposed in many different phases. Each phase implies a decision by a practitioner, which cannot always be justified by theoretical methods. In the following figure, a workflow of the decisions that a practitioner should assume is given. The figure is adapted from a workflow in Gill et al. (2001), p. 33.

¹ Although a unique common identifier is used to fuse data from two files, some different problems can arise even when linkage is achieved through some automated process. Scheuren and Oh (1975) report problems related to misprints, absence of SSN in one of the two records that are candidate to be matches, unexplainable changes of SSN in records known to be from the same person, etc.



The workflow describing the practical actions of a practitioner for applying record linkage procedures shows that the actual record linkage problem (as described in WP1 in Section 1, Cibella et al., 2008a) is tackled only in a few steps (the selection of model with the estimation method and the evaluation of the model estimation; the selection of the thresholds for deciding matches).

The steps to be performed are summarised in the following list.

- 1) At first a practitioner, should decide which are the variables of interest available distinctly in the two files. To the purpose of linking the files, the practitioner should understand which

variables are able to identify the correct matched pairs among all the common variables. These variables will be used as either matching or blocking variables.

- 2) The blocking and matching variables should be appropriately harmonised before applying any record linkage procedure.
- 3) When the files A and B are too large (as usually happens) it is appropriate to reduce the search space from the Cartesian product of the files A and B to a smaller set of pairs, as described above in par.2.
- 4) After the selection of a comparison function a suitable model should be chosen. This should be complemented by the selection of an estimation method, and possibly an evaluation of the obtained results. After this step, the application of a decision procedure needs the definition of cut-off thresholds.
- 5) There is the possibility of different outputs, logically dependent on the aims of the match. The output can take the form of a one-to-one, one-to-many or many-to-many links.
- 6) The output of a record linkage procedure is composed of three sets of pairs: the links, the non-links, and the possible links. This last set of pairs should be analysed by trained clerks.
- 7) The final decision that a practitioner should consider consists in deciding how to estimate the linkage errors and how to include this evaluation in the analyses of linkage files.

4. Available software tools

The main use of the record linkage techniques in official statistics produced many software and tools both in the academic and private sectors, like BigMatch (Yancey, 2007), GRLS (Fair, 2001), Febrl (<http://www.sourceforge.net/projects/febrl>), Link Plus (<http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm>), Tailor (Elfeky et al., 2002), etc.

In the ESSnet on Integration of Surveys and Administrative data (ISAD) the characteristics of some available software tools explicitly developed for record linkage and based on a probabilistic paradigm were analysed (see WP3, Chapter 1, section 1.1 and 1.3, Cibella et al., 2008c).

The probabilistic record linkage tools that have been selected among the most well-known and adopted ones are:

1. AutoMatch, developed at the US Bureau of Census, now under the purview of IBM [Herzog et al., 2007, chap.19].
2. Febrl - Freely Extensible Biomedical Record Linkage, developed at the Australian National University [FEBRL].
3. Generalized Record Linkage System (GRLS), developed at Statistics Canada [Herzog et al., 2007, chap.19].
4. RELAIS, developed at ISTAT [RELAIS].
5. The Link King, commercial software [LINKKING].
6. Link Plus, developed at the U.S. Centre for Disease Control and Prevention (CDC), Cancer Division [LINKPLUS].

An interesting feature of some tools is related to the fact that some record linkage activities are performed “within” other tools. For instance, there are several data cleaning tools that include record linkage but they are mainly dedicated to standardisation, consistency checks etc. A second example is provided by the recent efforts by major database management systems’ vendors (like Microsoft and Oracle) that include record linkage functionalities for data stored in relational databases (Koudas et al., 2006).

In the following, two comparison tables are presented and described with the aim of summarising and pointing out the principal features of each tool so far described. In Table 1, the selected values for the characteristics specified above for each of the analysed tools are reported.

Table 1: Main features

	Free/Commercial	Domain Specificity	Level of Adoption
AUTOMATCH	commercial	functionalities for English words	high
FEBRL	free/source code available	no specific domain	medium
GRLS	commercial (government)	functionalities for English words	medium
RELAIS	free/source code available	no specific domain	low
THE LINK KING	free/source code available (SAS licence is needed)	mixed/requires first and last names, date of birth	high
LINK PLUS	free/source code not available	mixed- general features	high

In Table 2 the details on the specific method used for the estimation of the Fellegi and Sunter model parameters are reported.

Table 2: Estimation methods implemented in the record linkage tools

	Fellegi Sunter Estimation Techniques
AUTOMATCH	Parameter estimation via frequency based matching
FEBRL	Parameter estimation via EM algorithm
GRLS	Parameter estimation under agreement/disagreement patterns
RELAIS	EM method Conditional independence assumption of matching variables
THE LINK KING	Ad hoc weight estimation method Not very clear theoretical hypotheses
LINK PLUS	Default M-probabilities + user-defined M-probabilities EM algorithm

The WP3 of the Essnet DI (Data Integration) was focused on the development of common software tools. In particular, as far as record linkage method is concerned, the goal was to improve Relais, the software for record linkage developed by a team of the Italian National Statistical Institute (ISTAT), with pre-processing facilities and a new manual (<http://www.essnetportal.eu/sites/default/files/131/Relais2.3Preprocessing.pdf>).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Armstrong, J. and Mayda, J. E. (1993), Model-based estimation of record linkage error rates. *Survey Methodology* **19**, 137–147.
- Cibella, N. et al. (2008a), Chapter 1. Literature review on probabilistic record linkage. Section 1.2, 1.5 and 1.7 of *WP1 Report of the ESSnet on Integration of Surveys and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>).
- Cibella, N. et al. (2008b), The practical aspects to be considered for record linkage. Section 2.1 of the *Report on WP2 of the ESSnet on Integration of Surveys and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>).
- Cibella, N. et al. (2008c), Software tools for record linkage. Chapter 1 of the *Report on WP3 of the ESSnet on Integration of Survey and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>).
- Copas, J. R., and Hilton, F. J. (1990), Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A* **153**, 287–320.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- Elfeky, M., Verykios, V., Elmagarmid, A. K. (2002), A Record Linkage Toolbox. *Proceedings of the 18th International Conference on Data Engineering IEEE Computer Society*, San Jose, CA, USA.
- Elfeky, M. G., Verykios, V. S., Elmagarmid, A., Ghanem, M., and Huwait, H. (2003) Record Linkage: A Machine Learning Approach, a Toolbox, and a Digital Government Web Service. Department of Computer Sciences, Purdue University, Technical Report CSD-TR 03-024.
- Fair, M. (2001), Recent developments at Statistics Canada in the linking of complex health files. Federal Committee on Statistical Methodology, Washington D.C., USA.
- Fellegi, I. P. and Sunter, A. B. (1969), A Theory for Record Linkage. *Journal of the American Statistical Association* **64**, 1183–1210.

- Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001), On Bayesian record linkage. *Research in Official Statistics* **4**, 185–198. Published also in: E. George (ed.), *Bayesian Methods*, Monographs of Official Statistics, Eurostat, 155–164.
- Fortini, M., Nuccitelli, A., Liseo, B., and Scanu, M. (2002), Modelling issues in record linkage: a Bayesian perspective. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1008–1013.
- Herzog T. N., Scheuren F. J., and Winkler, W. E. (2007), *Data Quality and Record Linkage Techniques*. Springer Science+Business Media, New York.
- Jaro, M. A. (1989), Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Association* **84**, 414–420.
- Koudas, N. and Srivastava, D. (2005), Approximate joins: Concepts and techniques. *Proceedings of VLDB 2005*.
- Koudas, N., Sarawagi, S., and Srivastava, D. (2006), Record linkage: similarity measures and algorithms. *SIGMOD Conference 2006*, 802–803.
- Gill, L. (2001), Methods for automatic record matching and linkage and their use in national statistics. National Statistics Methodological Series No. 25, London (HMSO).
- Lahiri, P. and Larsen, M. D. (2000), Model-based analysis of records linked using mixture models. *Proceedings of the Section on Survey Research Methods Section*, American Statistical Association, 11–19.
- Lahiri, P. and Larsen, M. D. (2005), Regression Analysis With Linked Data. *Journal of the American Statistical Association* **100**, 222–230.
- Larsen, M. D. (2004), Record Linkage of Administrative Files and Analysis of Linked Files. In: *IMS-ASA's SRMS Joint Mini-Meeting on Current Trends in Survey Sampling and Official Statistics*, The Ffort Radisson, Raichak, West Bengal, India.
- Larsen, M. D. and Rubin, D. B. (2001), Iterative automated record linkage using mixture models. *Journal of the American Statistical Association* **96**, 32–41.
- Liseo, B. and Tancredi, A. (2004), Statistical inference for data files that are computer linked. *Proceedings of the International Workshop on Statistical Modelling*, Firenze Univ. Press, 224–228.
- Neter, J., Maynes, E. S., and Ramanathan, R. (1965), The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association* **60**, 1005–1027.
- Newcombe, H., Kennedy, J., Axford, S., and James, A. (1959), Automatic Linkage of Vital Records. *Science* **130**, 954–959.
- Scheuren, F. and Oh, H. L. (1975), Fiddling around with nonmatches and mismatches. *Proceedings of the Social Statistics Section*, American Statistical Association, 627–633.
- Scheuren, F. and Winkler, W. E. (1996), Recursive analysis of linked data files. U.S. Bureau of the Census, Statistical Research Division Report Series, n.1996/08.

- Scheuren F. and Winkler W. E. (1997), Regression analysis of data files that are computer matched – part II. *Survey Methodology* **23**, 157–165.
- Thibaudeau, Y. (1989), Fitting log-linear models when some dichotomous variables are unobservable. *Proceedings of the Section on statistical computing*, American Statistical Association, 283–288.
- Winkler, W. E. (1989), Frequency-based matching in Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778–783 (longer version report rr00/06 at <http://www.census.gov/srd/www/byyear.html>).
- Winkler, W. E. (1993), Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 274–279.
- Winkler, W. E. (2006), Overview of Record Linkage and Current Research Directions. U.S. Bureau of the Census, Statistical Research Division Report Series, n.2006/2.
- Yancey, W. (2007), BigMatch: A Program for Extracting Probable Matches from a Large File. Research Report Series Computing, 2007-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Object Matching (Record Linkage)

9. Methods explicitly referred to in this module

1. Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage
2. Micro-Fusion – Weighted Matching of Object Characteristics

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 – Process

12. Tools explicitly referred to in this module

1. AUTOMATCH
2. Febrl
3. GRLS
4. RELAIS (REcord Linkage At IStat)
5. THE LINK KING
6. LINK PLUS

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.1: Integrate data

Administrative section

14. Module code

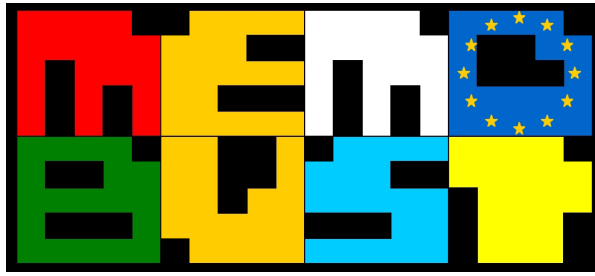
Micro-Fusion-T-Probabilistic Record Linkage

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	11-05-2012	first version	Nicoletta Cibella	Istat
0.2	02-10-2012	second version	Nicoletta Cibella	Istat
0.2.1	03-10-2013	preliminary release		
0.3	09-10-2013	EB comments	Nicoletta Cibella	Istat
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:58



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Fellegi-Sunter and Jaro Approach to Record Linkage

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Estimation of matching probabilities.....	5
3. Preparatory phase	6
4. Examples – not tool specific.....	7
4.1 Linkage between survey business data and administrative data.....	7
4.2 Estimating number of units in a population amount by capture-recapture method.....	8
4.3 Estimating number of under coverage farms in Agricultural Census.....	9
4.4 Enriching and updating the information stored in different sources	10
5. Examples – tool specific.....	10
6. Glossary.....	10
7. References	11
Specific section.....	13
Interconnections with other modules.....	17
Administrative section.....	19

General section

1. Summary

The Fellegi and Sunter method is a probabilistic approach to solve record linkage problem based on decision model. Records in data sources are assumed to represent observations of entities taken from a particular population (individuals, companies, enterprises, farms, geographic region, families, households...). The records are assumed to contain some attributes identifying an individual entity. Examples of identifying attributes are name, address, age and gender when dealing with people; style (or name) of a firm, legal form, address, number of local units, number of employees, turnover value when dealing with businesses. According to the method, given two (or more) sources of data, all pairs coming from the Cartesian product of the two sources has to be classified in three independent and mutually exclusive subsets: the set of matches, the set of non-matches and the set of pairs requiring manual review. In order to classify the pairs, the comparisons on common attributes are used to estimate for each pair the probabilities to belong to both the set of matches and the set of non-matches. The pair classification criteria is based on the ratio between such conditional probabilities. The decision model aims to minimise both the misclassification errors and the probability of classifying a pair as belonging to the subset of pairs requiring manual review.

2. General description of the method

Record linkage consists in matching the records belonging to different data sets when they correspond to the same unit. Records in data sources are assumed to represent observations of entities taken from a particular population (individuals, companies, enterprises, farms, geographic region, families, households...). The records are assumed to contain some attributes (variables) identifying an individual entity. Examples of identifying attributes are name, address, age and gender. Let A and B be two data sets, partially overlapping and containing the same type of units, of size N_A and N_B respectively. Suppose also that the two files consist of vectors of variables (X_A, Z_A) and (X_B, U_B) , either quantitative or qualitative, assuming that X_A and X_B are sub-vectors of common attributes, called key variables or matching variables in what follows, so that any single unit is univocally identified by an observation x . The goal of record linkage is to find all the pairs of units $(a,b) \in \Omega = \{(a,b): a \in A, b \in B\}$, such that a and b refer actually to the same unit ($a=b$). Hence, a record linkage procedure can be considered as a decision model based on the comparison of the key variables; for each single pair of records either one of the following decisions can be taken: link, possible link and non-link. Since the key variables can be prone both to measurement errors and misreporting, the record linkage problem is far from being a trivial one and Fellegi and Sunter (1969) propose an approach to the probabilistic record linkage based on decision model to minimise the incidence of both the non-decision area and false and missed links.

Let us consider $\Omega = \{(a,b), a \in A \text{ and } b \in B\}$ of size $N=N_A \times N_B$. This method considers all the pairs as a sample of $N_A \times N_B$ records independently generated by a mixture of two distributions: one for the matched pairs and the other for the unmatched ones. The linkage between A and B can be defined as the problem of classifying the pairs that belong to Ω in two subsets M and U independent and mutually exclusive, such that:

M is the set of matches ($a=b$)

U is the set of non-matches ($a \neq b$)

Actually, the model assumption fails to be true for the sample defined by the set of $N_A \times N_B$ records for the two data sets to link. In that case, it is not possible to state that comparison variables are independently generated by appropriate distributions. For more details about this weakness, see Kelley (1984). It is not yet clear how the failure of this independence hypothesis affects the record linkage results.

In order to classify the pairs, K common identifiers (called matching variables)

$$\mathbf{X}_1^A \quad \mathbf{X}_2^A \quad \dots \quad \mathbf{X}_K^A ; \quad \mathbf{X}_1^B \quad \mathbf{X}_2^B \quad \dots \quad \mathbf{X}_K^B$$

have to be chosen (the variables with the same subindex are comparable). So, for each pair, a comparison vector $\gamma = \{\gamma_1, \gamma_2, \dots, \gamma_K\}$ can be defined, by means of distance functions applied to matching variables for each pair. For instance, Fellegi and Sunter consider the binary comparison vector

$$_{(a,b)}\gamma_k = \begin{cases} 1 & \text{if } X_k^A = X_k^B \\ 0 & \text{otherwise} \end{cases}$$

For an observed comparison vector γ in Γ , the space of all comparison vectors, $m(\gamma)$ is defined to be the conditional probability of observing γ given that the record pair is a true match: in formula $m(\gamma) = P(\gamma | (a,b) \in M)$. Similarly, $u(\gamma) = P(\gamma | (a,b) \in U)$ denotes the conditional probability of observing γ given that the record pair is a true non-match.

There are two kinds of possible misclassification errors: false matches and false non-matches. The probability of false matches is:

$$\mu = P(M^* | U) = \sum_{\gamma \in \Gamma} u(\gamma) P(M^* | \gamma)$$

and the probability of a false non-matches is:

$$\lambda = P(U^* | M) = \sum_{\gamma \in \Gamma} m(\gamma) P(U^* | \gamma)$$

where M^* and U^* are the sets of estimated matches and estimated non-matches, respectively. For fixed values of μ and λ , Fellegi and Sunter define the optimal linkage rule as the rule that minimises the probability of assigning a pair in the set of no-decision Q , that is the set of pairs requiring clerical review so to be solved. The optimal rule is a function of the probability ratio

$$r = \frac{P(\gamma | (a,b) \in M)}{P(\gamma | (a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)}.$$

In practice, once the probabilities m and u are estimated, all the pairs can be ranked according to their ratio $r = m/u$ in order to detect which pairs are to be matched by means of this classification criterion based on the two thresholds T_m and T_u ($T_m > T_u$)

$$\begin{aligned}
r_{(a,b)} > T_m &\Rightarrow (a,b) \in M^* \\
T_m \geq r_{(a,b)} \geq T_u &\Rightarrow (a,b) \in Q \\
r_{(a,b)} < T_u &\Rightarrow (a,b) \in U^*
\end{aligned}$$

- those pairs for which r is greater than the upper threshold value can be considered as linked
- those pairs for which r is smaller than the lower threshold value can be considered as not-linked

The thresholds are assigned solving equations that minimise both the size of the set Q and the false match rate (FMR) and false non-match rate (FNMR).

$$\begin{aligned}
FMR &= \sum_{\gamma \in \Gamma} u(\gamma) P(M^* | \gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma) \quad \text{where} \quad \Gamma_{M^*} = \{\gamma : T_m \leq m(\gamma)/u(\gamma)\} \\
FNMR &= \sum_{\gamma \in \Gamma} m(\gamma) P(U^* | \gamma) = \sum_{\gamma \in \Gamma_{U^*}} m(\gamma) \quad \text{where} \quad \Gamma_{U^*} = \{\gamma : T_u \geq m(\gamma)/u(\gamma)\}
\end{aligned}$$

2.1 Estimation of matching probabilities

In order to apply the model for record linkage described in the previous section, a method for estimating the likelihood ratio $r=m/u$ is required. In their seminal paper, Fellegi and Sunter define a system of equations for estimating the parameters of the distributions for matched and unmatched pairs, based on the method of moments; it gives estimates in closed form when the comparison variables are at least three. Currently, the most widespread method for estimating the conditional probabilities m and u is the expectation-maximisation (EM) algorithm (Dempster et al., 1977), in the record linkage field first used by Jaro (1989). This is why the presented method is called the Fellegi-Sunter and Jaro one. According to this approach, the frequency distribution of the observed patterns γ is viewed as a mixture of the matches $m(\gamma)$ and non-matches $u(\gamma)$ distributions

$$\begin{aligned}
P(\gamma) &= P(\gamma | (a,b) \in M) P((a,b) \in M) + P(\gamma | (a,b) \in U) P((a,b) \in U) \\
&= m(\gamma) \cdot p + u(\gamma) \cdot (1 - p)
\end{aligned}$$

where $p=P(M)$. This means to consider a latent variable C , indicating the actual unknown matching status of the record pair, that takes value 1 corresponding to a match with probability p and value 0 corresponding to non-match with probability $1-p$.

The joint distribution of the observations γ and the latent variable C is given by:

$$P(C = c, \gamma) = [pm(\gamma)]^c [(1-p)u(\gamma)]^{1-c}.$$

Jaro restricts to 0/1 values the possible outcomes for the comparison vector γ , as in the previous Fellegi and Sunter model, and assumes conditional independence of the γ_k . These assumptions are currently often made in order to simplify the parameter estimation; in this case the likelihood function for $m_k(\gamma)$, $u_k(\gamma)$ ($k=1, \dots, K$) and p can be written as:

$$L = \prod_{(a,b)} [pm(\gamma^{(a,b)})]^{c^{(a,b)}} [(1-p)u(\gamma^{(a,b)})]^{1-c^{(a,b)}}.$$

The EM algorithm uses maximum likelihood estimates of $m_k(\gamma)$, $u_k(\gamma)$ and p to estimate the unobserved c . The EM algorithm needs initial estimates of $m_k(\gamma)$, $u_k(\gamma)$ and p and then iterates. Generally, the EM algorithm solutions don't depend on the initial values.

Under the conditional independence assumption the likelihood ratio r is given by:

$$r = \frac{P(\gamma|M)}{P(\gamma|U)} = \prod_{k=1}^K \frac{P(\gamma_k|M)}{P(\gamma_k|U)} = \prod_{k=1}^K \frac{m_k}{u_k}.$$

Even conditional independence assumption works well in most of the practical applications, it cannot be sure that this hypothesis is automatically satisfied. Some authors (Winkler 1989, and Thibaudeau 1989) extend the standard approach by means of log-linear models with latent variable by introducing appropriate constraints on parameters so to overcome to some extent conditional independence assumption. In these cases, however, it is not sure if the best model in term of fitting could be also considered as the most accurate in terms of linkage results and errors. See item 2 of the following section 11 (Variants of the method) for more details.

The Fellegi–Sunter and Jaro approach is heavily dependent on the accuracy of $m(\gamma)$ and $u(\gamma)$ estimates. Misspecifications in the model assumptions, lack of information and other problems can cause a loss of accuracy in the estimates and, as a consequence, an increase of both false matches and non-matches.

For this reason the appropriate thresholds are often identified mainly through empirical methods which need of scrutiny by experts, such as a diagram of the weights distribution as the one showed in the figure below.

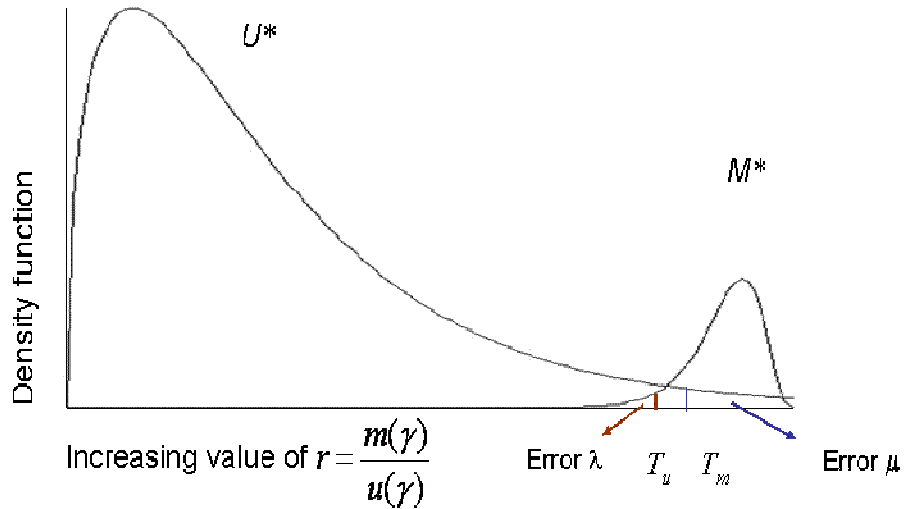


Figure 1. The mixture model for m - and u -distributions

3. Preparatory phase

Probabilistic record linkage, as proposed by Fellegi–Sunter and Jaro, is a complex procedure that can be decomposed in many different phases. The actual probabilistic record linkage model, as described in the previous section, is tackled only in few steps, but, as a matter of fact, all the previous steps are necessary when considering the use of the Fellegi–Sunter and Jaro method in practical situations.

The main points faced in this section are treated in depth in the WP2 of the ESSnet on ISAD (integration of surveys and administrative data), Section 2.1 (Cibella et al., 2008a) and in the theme

module “Micro-Fusion – Probabilistic Record Linkage”. In these papers, recommendations and suggestions are proposed as well, on the basis of the requirements of specific application.

The steps to be performed to apply the method can be summarised in the following list.

- 1) At first, a practitioner should decide which are the variables of interest available distinctly in the two files. To the purpose of linking the files, the practitioner should understand which variables are able to identify the correct matched pairs among all the common variables. These variables will be used as either matching or blocking variables.
- 2) The matching variables should be appropriately harmonised before applying any record linkage procedure. Harmonisation is in terms of variable definition, classification, codification, categorisation and so on.
- 3) When the files A and B are too large (as usually happens) it is appropriate to reduce the search space from the Cartesian product of the files A and B to a smaller set of pairs.
- 4) For probabilistic record linkage, after the selection of a comparison function, the suitable model should be chosen. This should be complemented by the selection of an estimation method, and possibly an evaluation of the obtained results. After this step, the application of a decision procedure needs the definition of the cut-off thresholds.
- 5) There is the possibility of different outputs, logically dependent on the aims of the match. The output can take the form of a one-to-one, one-to-many or many-to-many links.
- 6) The output of a record linkage procedure is composed of three sets of pairs: the links, the non-links, and the possible links. This last set of pairs should be analysed by trained clerks.
- 7) The final decision that a practitioner should consider consists in deciding how to estimate the linkage errors and how to include this evaluation in the analyses of linkage files.

4. Examples – not tool specific

4.1 Linkage between survey business data and administrative data

The following example is summarised from the paper Ichim et al. (2009). It is regarding the exploitation of administrative data in the NSIs. The administrative data may be useful in the sample design stage or in the estimation phase. Auxiliary information may also be useful in data validation and editing. The original paper reports several experiments undertaken to identify an optimal matching strategy for business data. The problem of linking survey and administrative data is addressed. The business survey data is the Small and Medium Enterprises (SME) survey while the administrative data source is the Balance Sheets (BIL). Small and Medium Enterprises sample survey (SME) is carried out annually by sending a postal questionnaire with the purpose of investigating profit-and-loss account of enterprises with less than 100 persons employed, as requested by SBS EU Council Regulation n. 58/97. The main variables of interest are Turnover, Value added at factor cost, Employment, Total purchases of goods and services, Personnel costs, Wages and salaries, Production value, etc. The frame for the SME survey is the Italian Statistical Business Register (ASIA). ASIA results from the logical and physical combination of data from both statistical and administrative sources. The Business Fiscal Turnover is provided from the Fiscal Register, this variable being a good proxy of the Turnover collected in SME. SME sample survey population of interest is about 4 millions of active

enterprises. Both the selection and estimation phases are based on the information available in ASIA, but a time lag exists between the reference years of SME and BR. The sample size is about 120.000 units. On the other side, the Italian limited enterprises are obliged to fill their financial statements according to the standards specified in the EEC fourth Directive and to transmit them to the Chambers of Commerce. The resulting database is called Balance sheets (BIL). This data source is actually the most used in the production of SBS estimates. In industry and services sectors there are about 500,000 limited enterprises which account for one half of the total employment. BIL data's coverage is 11.3% among 1-19 persons employed size class, it reaches 80.7% in the size class 20-99 and it is 96.2% among larger enterprises. The main aims of the linkage between SME and BIL are related to

- check and validation of survey results;
- obtain auxiliary information to deal with survey non-responses.
- update the frame of the survey reference population
- supply information to plan future survey wages.

The linkage procedures, described in the paper, mainly stress the efforts devoted to the pre-processing phase, the blocking step and the choice of the matching variables and the corresponding distances. Standardisation were applied to the streets typology and the types of the enterprises in both datasets. The unusual characters were deleted (@, -, =, \$, #, &, double spaces). The most frequent strings in the name of the enterprise were standardised, too. In order to select the matching variables, some descriptive statistics and correlations between the numerical variables were calculated. For the identified matching variables, different combinations of the several distance functions were tested to identify the best setting of the linkage experiment. Two reduction methods were applied in these experiments: blocking and sorted neighbourhood. The applied probabilistic model follows the Fellegi-Sunter and Jaro approach. In this work, the thresholds were derived from the probabilities of false nonmatch (0.90) and false match (0.95). Finally, the reduction one-to-one was solved as a linear programming problem. In both BIL and SME datasets, there is a unique identifier, namely the fiscal code. Even if the fiscal code may be subject to some errors, it was used for evaluating the quality of the record linkage through *precision* and *recall* (see section 21 below for these quality measures). Details on the several tests and results can be achieved in the full paper.

4.2 *Estimating number of units in a population amount by capture-recapture method*

The following example is summarised from the paper Cibella et al. (2008c). It involves data from the 2001 Italian Population Census and its Post Enumeration Survey (PES). The main goal of the Census was to enumerate the resident population at the Census date, 21/10/2001. The PES instead had the objective of estimating the coverage rate of the Census; it was carried out on a sample of enumeration areas (called EA in the following), which are the smallest territorial level considered by the Census. The size of the PES's sample was about 70000 households and 180000 individuals while the variables stored in the files are name, surname, gender, date and place of birth, marital status, etc. Correspondingly, comparable amounts of households and people were selected from the Census database with respect to the same EAs. The PES was based on the replication of the Census process inside the sampled EAs and on the use of a capture-recapture model (Wolter, 1986) for estimating the hidden amount of the population. In order to apply the capture-recapture model, after the PES enumeration of the statistical units (households and people), a record linkage between the two lists of

people built up by the Census and the PES was performed. In this way the rate of coverage, consisting of the ratio between the people enumerated at the Census day and the hidden amount of the population, was obtained.

The Fellegi–Sunter and Jaro linkage procedure, as described in previous sections, is applied on two sub-sets of size 8000 records, corresponding to the EAs of Rome. As matching variables all the strongest identifiers were used: name and surname, gender, day, month, and year of birth. Even if, generally, string variables as name and surname can complicate the linkage process due to diminutives or synonyms, in this example they didn't need further work due to their high quality level. So, the equality were applied as comparison function. The parameters of the Fellegi-Sunter probabilistic model were estimated via the EM algorithm. Two thresholds were fixed in order to individuate the three sets of Matches, of Unmatches and of Possible Links. The upper threshold was fixed assigning to the set of Matches all the pairs with the likelihood ratio corresponding to estimated matching probabilities higher than 0.99; the set of the possible links was created fixing the lower threshold level with the likelihood ratio corresponding to the estimated matching probability lower than 0.50. The pairs falling into the set of the Possible Links were assigned to the set of Matches without clerical supervision of the results.

A blocking phase was used considering as blocking variable the month of birth of the household head. In this way 12 blocks were created, plus a residual block formed by the units with missing information about the month of birth of the household header. The resulting blocking size are quite similar and homogeneous. The overall match rate is equal to 88%, the false match rate is 0.5% and the false non-match rate is 12%. Those results are comfortable and quite optimistic if compared with those coming from the scientific community, when a record linkage is performed in analogous conditions in terms of identification variables, number of matched records, kind of matched units. The results have to be regarded also more optimistic considering the unsupervised possible link data processing. Anyway, when the linkage is finalised to evaluate coverage rate, as in Census Post Enumeration Survey, the value of the false non-match rate has to be as small as possible and the resulting 12% false non-match rate is too high. In this situation, a further linkage procedure should be applied to the records non-linked at the first time, if it is possible without using blocking phase, so to minimise the risk of losing matches. The estimates of the Census coverage rate through capture-recapture model has required to match Census and PES records, assuming no errors in matching operations. Therefore the linkage between the two sources was both deterministic and probabilistic and the results was checked manually; all the linkage operations lasted several working days. Due to the accuracy of the matching procedures adopted, we know the true linkage status of all candidate pairs. In this way we can evaluate the effectiveness of the Fellegi-Sunter and Jaro linkage method in terms of match rate, false match rate and false non-match rate.

4.3 Estimating number of under coverage farms in Agricultural Census

The capture-recapture model introduced in the previous example has been also applied for the estimation of the unknown true number of farms, or, equivalently, for the estimate of the under-coverage rate of the Agricultural census. With respect to the previous example, the general workflow of the linkage procedure is the same, but different problem arise in comparing farms rather than people. In this case, the matching variables are the name (of the company name) of the farm or the name of its owner, the legal form, the utilised agricultural area, the address of the farm or the address

of its owner. Dealing with farms, the pre-processing procedures for name and address (in rural area) standardisation are very important and time-consuming.

4.4 Enriching and updating the information stored in different sources

The following example is summarised from the paper Cibella and Tuoto 2012. A record linkage is applied in order to study the fecundity of married foreign-women with residence in Italy. The Fellegi-Sunter and Jaro linkage method regards data referred to marriages with almost one of the married couple foreign and resident and data referred to babies born in the same Region in 2005-2006, from the registers of births. The size of each file is about 30000 records. The common variables are: fiscal code of the bride/mother, the 3-digit-standardised name and surname of both spouses/parents, the day/month/year of birth of the bridegroom/father and of the bride/mother, the municipality of the event (marriage/birth). Due to the data size, a data reduction method is needed, avoiding to deal with 900 millions of candidate pairs; analyses on the accuracy and of the frequency distribution of the available variables has limited the choice to the 3-digit-standardised name and surname of the bride/mother as blocking keys. The adopted blocking strategy is based on sorted neighbourhood method using as order variable the 6-digit-string of name and surname (composed from joining the 3-digit-standardised name and surname) over a window of size 15. The Fellegi-Sunter and Jaro method has been applied on the about 400000 candidate pairs produced by the sorted neighbourhood reduction, considering as matching variables: the 3-digit-standardised name of the mother and her day/month/year of birth. Equality function was used to compare the variables. The two thresholds to identify the tree sets of Matches, of Unmatches and of Possible Links were fixed in the following way: the upper threshold assigns to the set of Matches all the pairs with the likelihood ratio correspondent to estimated matching probability higher than 0.95; the lower threshold assigns to the set of the possible links all the pairs with the likelihood ratio correspondent to the estimated matching probability lower than 0.80. The procedure identified 567 matches and 457 possible matches. Among the matches, even 499 pairs have the same fiscal code or agree in all the bridegroom/father variables, while, among the possible matches, the concordance in the pairs is 25; so, totally, 592 true matches are identified by this procedure. This result can be compared with the total amount of pairs with common fiscal code in the files (they are 517 records).

5. Examples – tool specific

The examples reported in the previous section were carried out by using the RELAIS tool. It implements the Fellegi and Sunter method for record linkage, using the EM algorithm for the estimation of the conditional probabilities. For the EM algorithm, the initial values of the parameters are $m(g)=0.8$, $u(g)=0.2$ and $p=0.1$; the maximum number of iteration is 5.000 and the stop criterion is achieved when the difference between the estimates of two iterations is 0.000001.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Armstrong, J. and Mayda, J. E. (1993), Model-based estimation of record linkage error rates. *Survey Methodology* **19**, 137–147.
- Cibella, N., Scanu, M., and Tuoto, T. (2008a), The practical aspects to be considered for record linkage. Section 2.1 of the *Report on WP2 of the ESSnet on Integration of Survey and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>).
- Cibella, N. and Tuoto, T. (2008b), Quality assessments. Section 1.7 of the *Report on WP1 of the ESSnet on Integration of Survey and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>).
- Cibella, N., Fortini, M., Scannapieco, M., Tosco, L., and Tuoto, T. (2008c) Theory and practice of developing a record linkage software. In: *Proceedings of the International Workshop “Combination of surveys and administrative data”, 29-30 May, Vienna, Austria.*
- Cibella, N. and Tuoto, T. (2012), Statistical perspectives on blocking methods when linking large data-sets. In A. Di Ciaccio et al. (eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets*, ISBN 978-3-642-21036-5, Springer-Verlag, Berlin Heidelberg.
- Copas, J. R. and Hilton, F. J. (1990), Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A* **153**, 287–320.
- Da Silva, A. D., Martins Romeo, O. S., Soares, T. S., and Xavier, V. L. (2011), Study of Record Linkage Software for the 2010 Brazilian Census Post Enumeration Survey. *Proceedings of the 58th ISI congress, 21-26 August, Dublin.*
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- Fellegi, I. P. and Sunter, A. B. (1969), A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1210.
- Gill, L. (2001), Methods for automatic record matching and linkage and their use in national statistics. National Statistics Methodological Series No. 25, London (HMSO).
- Heasman, D., Bailliel, M., Danielis, J., McLeod, P., and Elkin, M. (2011), Applications of record linkage to population statistics in the UK. Workshop of the ESSnet Data Integration, 24-25 November, Madrid, Spain.
- Ichim, D., Casciano, C., and Seri, G. (2009), A linkage experiment between survey business data and administrative data. Workshop “2009 European Establishment Statistics”, September 7-9, Stockholm, Sweden.
- Jaro, M. A. (1989), Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* **84**, 414–420.
- Kelley, R. B. (1984), Blocking considerations for record linkage under conditions of uncertainty. Statistical Research Division Report Series, SRD Research Report No. RR-84/19, Bureau of the Census, Washington, D.C.

- Larsen, M. D. and Rubin, D. B. (2001), Iterative automated record linkage using mixture models. *Journal of the American Statistical Association* **96**, 32–41.
- Scanu, M. (2008), Estimation of the distributions of matches and nonmatches. Section 1.5 of the *Report on WP1 of the ESSnet on Integration of Survey and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>).
- Thibaudeau, Y. (1989), Fitting log-linear models when some dichotomous variables are unobservable. *Proceedings of the Section on statistical computing*, American Statistical Association, 283–288.
- Thibaudeau, Y. (1993), The discrimination power of dependency structures in record linkage. *Survey Methodology* **19**, 31–38.
- Thompson, G. (2011), Linking Information to the Australian Bureau of Statistics Census of Population and Housing in 2011. Workshop of the ESSnet Data Integration, 24-25 November, Madrid, Spain.
- Winkler, W. E. (1989a), Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington D.C., 145–155.
- Winkler, W.E. (1989b), Frequency-based matching in Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778–783 (longer version report rr00/06 at <http://www.census.gov/srd/www/byyear.html>).
- Winkler, W. E. (1993), Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 274–279.
- Wolter, K. (1986), Some coverage error models for Census data. *Journal of the American Statistical Association* **81**, 338–346.

Specific section

8. Purpose of the method

The purpose of Fellegi-Sunter and Jaro record linkage procedure is to identify the same real world entity that can be differently represented in data sources, even when unique identifiers are not available or are affected by errors. This operation is suitable when two or more partially or completely overlapping sets of data have to be integrated at micro level so as the information available in one frame for a unit can be linked to the information related to exactly the same unit stored in the other frame. The different frame can be statistical or coming from administrative data.

9. Recommended use of the method

1. The Fellegi-Sunter and Jaro method is recommended when unique identifiers are not available for all the units or when they are affected by errors. Regardless of the record linkage purposes, the following logic is adopted in extreme cases: when a pair of records is in complete disagreement on some *key* issues it will be almost certainly composed of different entities; conversely, a perfect agreement will indicate an almost certain match. All the intermediate cases, whether a partial agreement between two different units is achieved by chance or a partial disagreement between a couple of records relating to the same entity is caused by errors in the comparison variables, have to be properly resolved. This method, under the suitable conditions, solve these ambiguous situations.

10. Possible disadvantages of the method

1. The Fellegi-Sunter and Jaro approach is heavily dependent on the accuracy of $m(\gamma)$ and $u(\gamma)$ estimates. Misspecifications in the model assumptions, lack of information, inappropriate choices in the previous steps of the whole record linkage process and so on can cause a loss of accuracy in the estimates. Generally speaking, estimation cannot be reliable when one of the categories of the latent variable (the matches) is too rare. In general, the set of the matched pairs M should be large enough (generally, more than 5% of the overall set of $N_A \times N_B$ pairs). For instance, this is one of the motivations for the application of blocking procedures. However, in most practical cases, even when the parameter estimates are not very reliable, the linkage procedure is robust with respect to the identification of the matches, while it does not allow a reliable estimation of the matching errors.

11. Variants of the method

This section has been taken from the WP1 of the ESSnet on ISAD (integration of surveys and administrative data), Section 1.5 (Scanu 2008).

1. Independence between the comparison variables – This assumption is usually called the Conditional Independence Assumption (CIA), i.e., the assumption of independence between the comparison variables γ_j^{ab} , $j=1, \dots, k$, given the match status of each pair (matched or unmatched pair). Fellegi and Sunter define a system of equations for estimating the parameters of the distributions for matched and unmatched pairs, based on the method of moments which gives estimates in closed form when the comparison variables are at least three. Jaro (1989)

solves this problem for a general number of comparison variables with the use of the EM algorithm (Dempster et al., 1977).

2. Dependence of comparison and latent variable defined by means of loglinear models – Thibaudeau (1989, 1993) and Armstrong and Mayda (1993) have estimated the distributions of the comparison variables under appropriate loglinear models of the comparison variables. They found out that these models are more suitable than the CIA. The problem is estimating the appropriate loglinear model. Winkler (1989, 1993) underlines that it is better to avoid estimating the appropriate model, because tests are usually unreliable when there is a latent variable. He suggests using a sufficiently general model, as the loglinear model with interactions larger than three set to zero, and incorporating appropriate constraints during the estimation process. For instance, an always valid constraint states that the probability of having a matched pair is always smaller than the probability of having a nonmatch. A more refined constraint is obviously the following:

$$p \leq \frac{n_A}{n_B \cdot n_A} = \frac{1}{n_B}.$$

Estimation of model parameters under these constraints may be performed by means of appropriate modifications of the EM algorithm, see Winkler (1993).

3. Iterative approaches – Larsen and Rubin (2001) define an iterative approach which alternates a model based approach and clerical review for lowering as much as possible the number of records whose status is uncertain. Usually, models are estimated among the set of fixed loglinear models, through parameter estimation computed with the EM algorithm and comparisons with “semi-empirical” probabilities by means of the Kullback-Leibler distance.
4. Other approaches – Different papers do not estimate the distributions of the comparison variables on the data sets to link. In fact, they use ad hoc data sets or training sets. These variants simplify the estimation procedure and can be applied in particular when the linkage is done for files that become available regularly and don’t change too much in time. In this last case, it is possible to use comparison variables more informative than the traditional dichotomous ones. For instance, a remarkable approach is considered in Copas and Hilton (1990), where comparison variables are defined as the pair of categories of each key variable observed in two files to match for matched pairs (i.e., comparison variables report possible classification errors in one of the two files to match). Unmatched pairs are such that each component of the pair is independent of the other. In order to estimate the distribution of comparison variables for matched pairs, Copas and Hilton need a training set. They estimate model parameters for different models, corresponding to different classification error models.

12. Input data

1. Input data for the Fellegi-Sunter and Jaro method for record linkage are two or more microdata files referred, partially or completely, to the same units.
2. The input datasets have to contain three or more matching variables, with high level of identification power and quality (few errors, few missing data). Note that the number of

matching variables and some of their characteristics (as the number of categories and their rarity) influence the identification of links.

3. Another type of input of the method is the distance function used to compare each pair of records. This function must be appropriate for reporting the characteristics of the selected matching variables. The equality function is the most widespread. Distance functions based on string comparators (as Levenstein, Jaro, Jaro-Winkler, Soundex, 3grams) can be useful applied when the matching variables are names and are affected by typos or other kind of errors.
4. A further input of the method is the level of acceptable error rates.

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. The acceptable levels of error rates are user-defined. These levels serve to assign the threshold values of the decision rule. Sometimes, due to the poor accuracy of the $m(\gamma)$ and $u(\gamma)$ estimates, the appropriate thresholds are often identified mainly through empirical methods which need scrutiny by experts.

15. Recommended use of the individual variants of the method

1. The model under the conditional independence assumption (CIA) has to be preferred if there is no evidence of marginal dependency among the matching variables and the linkage status, as usual.
2. When training set of data with the true matching status is available, for instance because an error-free identification code is available for a sub-set of records, the Copas and Hilton variant can be applied in order to improve the accuracy of the estimates.

16. Output data

1. The Fellegi and Sunter method produces a single set of data collecting the pairs in common in the two input datasets, i.e., the set of matches. In this dataset, for all matched pairs, all the original variables are available and more an output variable reporting the matching probability.

2. The method generally produces a file of possible links, i.e., pairs that need a manual review or further analyses in order to be assigned to the match set or to be discarded as non-matches.
3. The method also allows to create residual files, i.e., from the original datasets can be created reduced dataset composed of the records that haven't been linked.
4. Finally, the method allows to create the set of non-matched pairs, i.e., the file composed of the pairs that, according to the decision rules, are declared as non-matches. This file can be useful in order to investigate the false non-matches.

17. Properties of the output data

1. The main advantage in using Fellegi-Sunter and Jaro method to solve record linkage problem is the availability of the linkage probability for each pair assigned to the set of matches. This probability allows to evaluate the quality of the linkage and it has to be taken into account in the following phase of the whole process.

18. Unit of input data suitable for the method

Processing full data sets

19. User interaction - not tool specific

- 1.

20. Logging indicators

1. Number of records in Dataset1
2. Number of records in Dataset2
3. Number of matching variables considered in the model
4. Comparison function used for each variable
5. Error levels considered acceptable

21. Quality indicators of the output data

This section has been taken from the WP1 of the ESSnet on ISAD (integration of surveys and administrative data), Section 1.7 (Cibella and Tuoto, 2008).

1. The first indicator of the output data is the match rate, i.e., the total number of linked record pairs divided by the total number of true match record pairs. In order to compute the match rate, the total number of true matches has to be known. In alternative, when the total number of true matches is unknown and it is not possible to achieve it in different way, a maximum value of the indicator can be calculated as the ratio between the total number of linked record pairs and the number of records of the smallest of the two input datasets.
2. Another indicator is the false match rate is defined the number of incorrectly linked record pairs divided by the total number of linked record pairs. The false match rate corresponds to the well-known $1-\alpha$ error in a one-tail hypothesis test. The estimate of such indicator is an output of the estimation step of the Fellegi-Sunter and Jaro method. In the epidemiological

field, instead of the false match rate, it is largely used the positive predictive value, defined as one minus the false match rate and corresponding to the number of correctly linked record pairs divided by the total number of linked record pairs.

3. One more indicator is the false non-match rate is defined as the number of incorrectly unlinked record pairs divided by the total number of true match record pairs. The false non-match rate corresponds to the β error in a one-tail hypothesis test. The estimate of such indicator is an output of the estimation step of the Fellegi-Sunter and Jaro method. In the epidemiological field, the sensitivity indicator is defined as the number of correctly linked record pairs divided by the total number of true match record pairs. It can be easily obtained from the false non-match rate.
4. A different performance measure is specificity, defined as the number of correctly unlinked record pairs divided by the total number of true non-match record pairs. The difference between sensitivity and specificity is that sensitivity measures the percentage of correctly classified record matches, while specificity measures the percentage of correctly classified non-matches.
5. In information retrieval the previous accuracy measures take the name of precision and recall. Precision measures the purity of search results, or how well a search avoids returning results that are not relevant. Recall refers to completeness of retrieval of relevant items. Hence, precision can be defined as the number of correctly linked record pairs divided by the total number of linked record pairs, i.e., it coincides with the positive predicted value. Similarly, recall is defined as the number of correctly linked record pairs divided by the total number of true match record pairs, i.e., recall is equivalent to sensitivity. As a matter of fact, precision and recall can also be defined in terms of non-matches.

22. Actual use of the method

1. The method is used in the linkage steps of the Post Enumeration Surveys of Agricultural Census in several countries (for instance in USA since 1985, in Italy since 2011).
2. The method is used in the linkage steps of the Post Enumeration Surveys of Population Census in several countries (in Italy, since 2011).
3. The method is used in linking information to the ABS Census of Population and Housing in 2011 (see Thompson, 2011).
4. The method is used in applications of record linkage to population statistics in the UK (see Heasman et al., 2011).
5. The method is used in the linkage steps of the 2010 Brazilian Census Post Enumeration Survey (see da Silva et al., 2011).
6. The method is used for building preparatory lists for the 2011 Population Census in Italy.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Data Fusion at Micro Level

2. Micro-Fusion – Object Matching (Record Linkage)
3. Micro-Fusion – Probabilistic Record Linkage

24. Related methods described in other modules

- 1.

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

1. RELAIS (Record linkage at Istat) is a toolkit providing a set of techniques for dealing with record linkage projects. It allows to dynamically select the most appropriate solution for each phase of record linkage and to combine different techniques for building a record linkage workflow of a given application. It is developed as an open source project. It is released under the EUPL license (European Union Public License) and it can be downloaded for free at <http://www.istat.it/it/strumenti/metodi-e-software/software/relais> with its User Guide, as well. It has been implemented by using two languages based on different paradigms: Java, an object oriented language, and R, a functional language. It is based on relational database architecture, mySql environment. The RELAIS project aims to provide record linkage techniques easily accessible to non-expert users. Indeed, the developed system has a GUI (Graphical User Interface) that on the one hand permits to build record linkage work-flows with a good flexibility. On the other hand it checks the execution order among the different provided techniques whereas precedence rules must be controlled. The current version of RELAIS provides several techniques to execute record linkage applications, in particular it allows to perform the Fellegi–Sunter and Jaro method for probabilistic record linkage, estimating the conditional matching probabilities via the EM algorithm. Moreover it provides different methods for search space reduction, several comparison functions, some metadata on the common variables in order to select them as matching or blocking variables. It runs under Windows and Linux environments.

28. Process step performed by the method

GSBPM Sub-process 5.1: Integrate data

Administrative section

29. Module code

Micro-Fusion-M-Fellegi-Sunter and Jaro Approach

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	11-05-2012	first version	Tiziana Tuoto	Istat
0.2	01-10-2012	second version	Tiziana Tuoto	Istat
0.2.1	04-10-2013	preliminary release		
0.3	09-10-2013	EB comments	Tiziana Tuoto	Istat
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:59



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Matching

Contents

General section.....	3
1. Summary	3
2. General description.....	3
3. Design issues	5
4. Available software tools.....	7
5. Decision tree of methods	7
6. Glossary.....	8
7. References	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

This section explores the problem of data integration in the following context: there are two non-overlapping surveys (in the sense that the two sets of units collected in the two surveys are distinct) that refer to the same target population, the variables of interest for the statistical analyses are available distinctly in the two surveys, due to the nature of the data sets it is not possible to create joint information on these variables by means of their common identifiers. This problem is usually referred to as statistical matching. As a matter of fact, this is a non-standard problem in statistics, for which naïve methods based on data imputation were defined at the beginning. Nowadays the complex nature of statistical matching is dealt differently, by the exploration of all the possible models that could give as a result the two sample surveys at hand, giving rise to “sets” of estimates instead of the more usual “point estimates”. These sets of estimates should not be confused with confidence intervals: they just reflect the fact that joint information on the target variables is missing.

2. General description

Statistical matching (sometimes called data fusion, synthetical matching) aims at combining information available in distinct sample surveys referred to the same target population. Formally, let Y and Z be two random variables (r.v.). Statistical matching is defined as the estimation of the joint (Y, Z) distribution function (e.g., a contingency table or a regression coefficient) or of some of its parameters when:

- Y and Z are not jointly observed in a survey, but
- Y is observed in a sample A , of size n_A ,
- Z is observed in a sample B , of size n_B ,
- A and B are independent, and the set of observed units in the two samples do not overlap (it is not possible to use record linkage),
- A and B both observe a set of additional variables X .

A figure representing this situation is the following.

	Y	X	Z
Data source A			missing
	Y	X	Z
Data source B	missing		

A detailed list of statistical matching applications is in D’Orazio et al. (2006) and Ridder and Moffit (2007). Generally speaking, this problem has been considered as an imputation problem. One of the

files, e.g., A, was considered the recipient, the other the donor file, and the statistical matching procedure consists in imputing Z in A by means of the available common information X. Among the procedures applied in this context, it is possible to distinguish

1. Use of imputation techniques that reproduce the assumption of independence of Y and Z given X (conditional independence assumption, henceforth CIA). One of the first statistical matching attempts is in Okner (1972). In this case, statistical matching consisted of the application of imputation techniques of taxable income observed on 1966 Internal Revenue Service Tax File on the 1967 Survey of Economic Opportunity. Denoting the common variables in the two files as X, the variables observed only in the Survey of Economic Opportunity as Y and those only in the Tax File as Z, these imputation techniques were able to reproduce the model of conditional independence between Y and Z given X. Appropriateness of CIA is discussed in several papers. We quote, among the others, Sims (1972) and Rodgers (1984).
2. Use of external auxiliary information for avoiding the CIA. This second group of techniques uses external auxiliary information on the statistical relationships between Y and Z, e.g., an additional file C where (X, Y, Z) are jointly observed is available (as in Singh et al., 1993).

The imputation procedures used in the two previous contexts can be clustered in:

1. parametric: i.e., explicit use of a parametric model (e.g., a regression) between X, Y and Z
2. nonparametric: use of hot-deck methods
3. mixed: two step procedures that partially make use of parametric models and then apply hot-deck methods for imputation of “live” values

These approaches are actually theoretically justified when the joint probability distribution of the variables of interest in the population coincides with the probability distribution of the same variables in the synthetic (imputed) data file, or at least when these two distributions are “very close”. The discrepancy between the joint distribution of the variables of interest (a) in the population, and (b) in the synthetic data file is usually referred to as matching noise Paass (1986). Attempts at evaluating the “closeness” of the empirical distribution of imputed data to the empirical distribution of “real” data have been performed in the literature, see D’Orazio et al. (2006). In a nonparametric setting an important role is played by hot-deck methods, as well as k-nearest neighbor (kNN) methods. Their properties are studied in Marella et al. (2008), where both theoretical and simulation results are obtained.

As a matter of fact, the CIA is usually a misspecified assumption, and external auxiliary information is most of the times not available. The lack of joint information on the variables of interest is the cause of uncertainty on the model of (X, Y, Z). The problem is that sample information provided by A and B is actually unable to discriminate among a set of plausible models for (X, Y, Z). In other terms, the adopted statistical model is not identifiable on the basis of sample data. Hence, a third group of techniques that does not directly aim at reconstructing a complete data set is introduced. This group of techniques addresses the so-called identification problem. The main consequence of the lack of identifiability is that some parameters of the model cannot be estimated on the basis of the available sample information. Instead of point estimates, one can only reasonably construct sets of “possible

point estimates”, compatible with what can be estimated (i.e., each point estimate is obtained by imposing a model which is compatible with the estimable distributions $Y|X$ and $Z|X$).

These sets (usually intervals) formally provide a representation of uncertainty about the model parameters (note that these intervals are not confidence intervals, the problem is not sampling variability, but the lack of joint information on Y and Z).

In this setting, the main task consists in constructing a coherent measure that can reasonably quantify the uncertainty about the (estimated) model. From an operational point of view, a measure of uncertainty essentially quantifies how “large” is the class of models estimated on the basis of the available sample information. The smaller the measure of uncertainty, the smaller the class of estimated models. Preliminary studies on this have been considered in Kadane (1979), Rubin (1986), Raessler (2002), D’Orazio et al (2006, Chapter 4). A thorough discussion on uncertainty measures is in Conti et al (2012).

When dealing with samples drawn according to complex survey designs, there is the problem of how to use the possibly different survey weights in a statistical matching context. Up to now there are essentially two distinct approaches.

1. File concatenation. This approach was suggested by Rubin (1986) and consists in defining the probabilities of inclusion that the units in the A sample would have had if the survey design of sample B was adopted (say π_a^B , $a=1,\dots,n_A$), and the probabilities of inclusion that the units in the B samples would have had if the survey design of sample A was adopted (say π_b^A , $b=1,\dots,n_B$). Then, the file obtained concatenating the two samples will have n_A+n_B units with probability of inclusion: $\pi_h^{A\cup B} = \pi_h^A + \pi_h^B - \pi_h^{A\cap B}$, $h=1,\dots, n_A+n_B$, where the last term indicates the probability of inclusion of a unit in the intersection between the two samples. Most of the times this last probability is negligible, and as suggested by Rubin it can be eliminated in the formula. This is not the case when, for instance, there are “take-all” strata in the two samples with a non-empty intersection (as it is typical for enterprise surveys, where take-all strata usually consist of large enterprises). Rubin suggests to use multiple imputation in order to fill in the missing data in the concatenated file.
2. Calibration. This approach was suggested by Renssen (1998), and consists in estimating all the distributions of X , $Y|X$ and $Z|X$ from A and B after a calibration step that makes the two surveys coherent on the common information (X). These distributions allow to apply statistical matching procedures under the CIA (Renssen suggests to use imputation by regression functions). Renssen studies also the case a complete third sample C is available and suggests two different procedures for making information on A, B and C coherent by means of calibration procedures. This use of an external auxiliary file C allows to avoid the assumption of conditional independence for Y and Z given X . Again, a complete file can be obtained by using imputation by regression.

3. Design issues

This section has been taken from the WP2 of the ESSnet on ISAD (integration of surveys and administrative data), Section 3.1 (Scanu, 2008a).

Figure 1 represents the steps that need to be performed for solving a statistical matching problem.

- 1) A key role is represented by the choice of the target variables, i.e., of the variables observed distinctly in two sample surveys. The objective of the study will be to obtain joint information on these variables. This task is important because it influences all the subsequent steps. In particular, the matching variables (i.e., those variables used for linking the two sample surveys) will be chosen according to their capacity to preserve the direct relationship between the target variables.
- 2) The second step is the identification of all the common variables in the two sources (potentially all these variables can be used as matching variables). Not all these variables can actually be used. The reasons can be different, as lack of harmonisation between the variables. To this purpose, some steps need to be performed as the harmonisation of their definition and classification, the need to take only accurate variables whose statistical content is homogeneous.
- 3) Once the common variables have been cleaned of those variables that cannot be harmonised, it is necessary to choose only those that are able to predict the target variables. To this purpose, it is possible to apply some statistical methods whose aim is to discover the relationship between variables, as statistical tests or appropriate models.

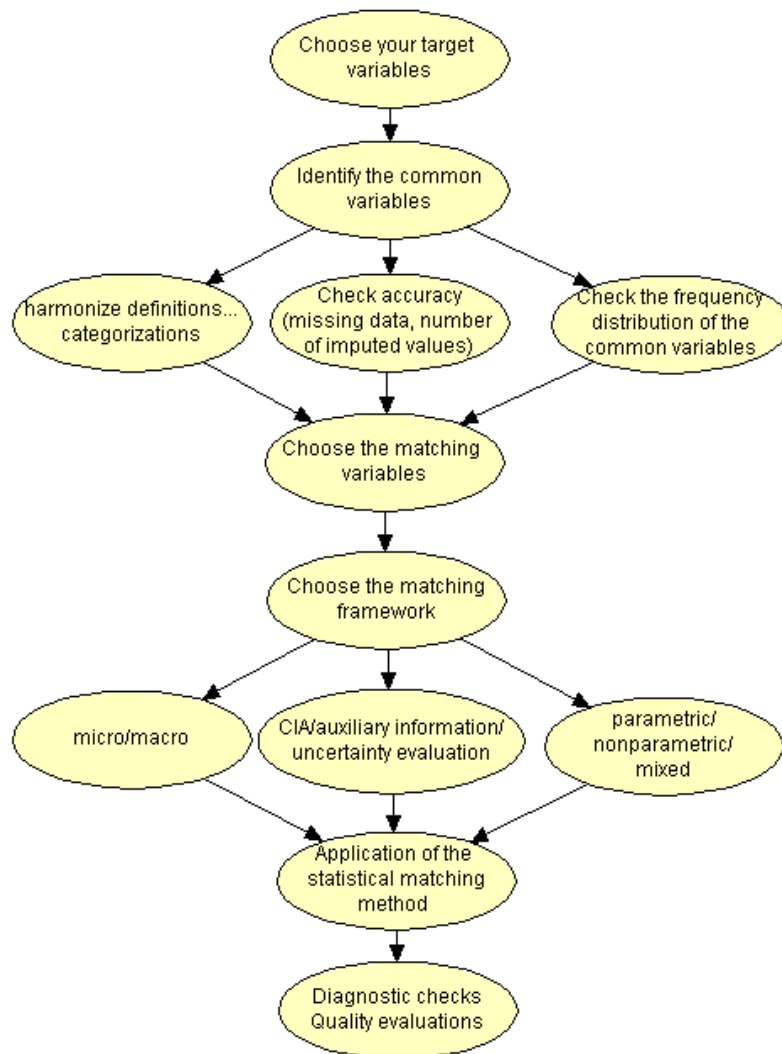


Figure 1: workflow of the actions to perform in statistical matching

- 4) As already introduced in the beginning, the statistical matching aim can be solved in different ways:
 - a. By a micro objective (i.e., construction of a complete data file with joint information on X, Y, and Z) or a macro objective (i.e., estimation of a parameter on the joint distribution of (Y,Z), (Y,Z|X), (X,Y,Z))
 - b. By the use of specific models (as the conditional independence assumption), the use of auxiliary information, or the study of uncertainty
 - c. By parametric, nonparametric or mixed procedures (this will be specified in Section “Statistical matching methods”).
- 5) Once a decision has been taken, the procedure is applied on the available data sets.
- 6) Quality evaluations of the results are the final step to perform.

Chapter 3 of the Report on WP2 of the ESSnet on ISAD describes in detail all the previous steps. The previous steps correspond to choices taken by the researcher that is performing a statistical matching application. What happens if some of the steps cannot be performed? This problem is especially connected with step 3, i.e., on the choice of the matching variables. If the common variables are unable to predict the target variables (e.g., they are independent of the target variables), statistical matching cannot be performed, because the common variables do not add any information on the relationship between the target variables.

4. Available software tools

The ESSnet on Integration of Surveys and Administrative data (ISAD) dealt with the problem of software tools in data integration. Workpackage 3 includes a thorough discussion on the available software tools (see Chapter 2, Scanu 2008b).

SAMWIN (Sacco, 2008): The software package SAMWIN was built for the production of an integrated archive for the social accounting matrix. This integrated archive was built by means of statistical matching techniques based on nonparametric imputation methods (hot-deck). For this reason, SAMWIN includes only matching algorithms based on the donors, more precisely distance hot-deck algorithms. The platform for SAMWIN is Visual Studio 6 (Visual C++). The developer is Giuseppe Sacco. Any question on SAMWIN should be sent to the email address sacco@istat.it.

StatMatch (D’Orazio, 2011). This is an R package consisting of functions for the implementation of statistical matching methods based on imputation procedures, under both the conditional independence assumption and the use of auxiliary information. It also includes functions for the evaluation of uncertainty.

SPlus codes (Raessler, 2002). These codes were written by Raessler for the implementation of proper multiple imputation methods for statistical matching in a Bayesian context.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Conti, P. L., Marella, D., and Scanu, M. (2012), Uncertainty analysis in statistical matching. *Journal of Official Statistics* **28**, 1–21.
- D’Orazio, M. (2011), *Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment*. Vignette for the application of the R package StatMatch, available on CRAN and at <http://www.cros-portal.eu/content/wp3-development-common-software-tools>.
- D’Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical matching: theory and practice*. Wiley, Chichester.
- Kadane, J. B. (1978), Some Statistical Problems in Merging Data Files. Compendium of Tax Research, Department of Treasury, U.S. Government Printing Office, Washington D.C., 159–179. (Reprinted in 2001, *Journal of Official Statistics* **17**, 423–433).
- Marella, D., Scanu, M., and Conti, P. L. (2008), On the Matching Noise of Some Nonparametric Imputation Procedures. *Statistics and Probability Letters* **78**, 1593–1600.
- Okner, B. A. (1972), Constructing a New Microdata Base from Existing Microdata Sets: The 1966 Merge File. *Annals of Economic and Social Measurement* **1**, 325–362.
- Paass, G. (1986), Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information. In: G.H. Orcutt and H. Quinke (eds.), *Microanalytic Simulation Models to Support Social and Financial Policy*, Elsevier, Amsterdam.
- Raessler, S. (2002), *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Lecture Notes in Statistics, Springer Verlag, New York.
- Renssen, R. H. (1998), Use of Statistical Matching Techniques in Calibration Estimation. *Survey Methodology* **24**, 171–183.
- Ridder, G. and Moffitt, R. (2007), The Econometrics of Data Combination. In: J. J. Heckmann and E. E. Leamer (eds.), *Handbook of Econometrics*, vol. 6A, Elsevier, Amsterdam.
- Rodgers, W. L. (1984), An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics* **2**, 91–102.
- Rubin, D. B. (1986), Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics* **4**, 87–94.
- Sacco, G. (2008), SAMWIN: a software for statistical matching. Document of WP3 of the *ESSnet on Integration of Surveys and Administrative Data*, available at http://cenex-isad.istat.it/archivio/Technical_reports_and_documentation/software_on_statistical_matching/SAMWIN_manual.pdf.
- Scanu, M. (2008a), The practical aspects to be considered for statistical matching. Section 3.1 of the *Report on WP2 of the ESSnet on Integration of Surveys and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>)

- Scanu, M. (2008b), Software tools for statistical matching. Chapter 2 of the *Report on WP3 of the ESSnet on Integration of Survey and Administrative Data* (<http://www.cros-portal.eu/content/isad-finished>).
- Sims, C. A. (1972), Comments and Rejoinder (On Okner (1972)). *Annals of Economic and Social Measurement* **1**, 343–345 and 355–357.
- Singh, A. C., Mantel, H., Kinack, M., and Rowe, G. (1993), Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. *Survey Methodology* **19**, 59–79.

Interconnections with other modules

8. Related themes described in other modules

1. Imputation – Main Module
2. Imputation – Donor Imputation
3. Weighting and Estimation – Main Module
4. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

1. StatMatch (R package)
2. SamWin

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.1: Integrate data

Administrative section

14. Module code

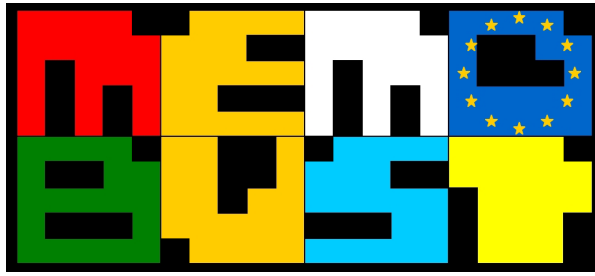
Micro-Fusion-T-Statistical Matching

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-03-2012	first version	Mauro Scanu	Istat (Italy)
0.2	02-05-2012	second version	Mauro Scanu	Istat (Italy)
0.3	25-09-2013	EB comments	Mauro Scanu	Istat (Italy)
0.3.1	03-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:59



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Statistical Matching Methods

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Parametric approach	4
2.2 Nonparametric approach	4
2.3 Mixed methods	4
3. Preparatory phase	4
4. Examples – not tool specific.....	5
5. Examples – tool specific.....	5
6. Glossary.....	6
7. References	6
Specific section.....	8
Interconnections with other modules.....	11
Administrative section.....	13

General section

1. Summary

Statistical matching (SM) methods for microdata aim at integrating two or more data sources related to the same target population in order to derive a unique synthetic data set in which all the variables (coming from the different sources) are jointly available. The synthetic data set is the basis of further statistical analysis, e.g., microsimulations. The word synthetic refers to the fact that the records are obtained by integrating the available data sets rather than direct observation of all the variables. Usually the matching is based on the information (variables) common to the available data sources and, when available, on some auxiliary information (a data source containing all the interesting variables or an estimate of a correlation matrix, contingency table, etc.). When the additional information is not available and the matching is performed on the variables shared by the starting data sources, then the results will rely on the assumption of independence among variables not jointly observed given the shared ones.

The synthetic data set can be derived by applying a parametric or a nonparametric approach. They can be mixed too.

2. General description of the method

Statistical matching at micro level attempts to derive a synthetic data source by integrating the available data sources. In the traditional framework, there are two data sets $A = \{X, Y\}$ and $B = \{X, Z\}$, sharing a number variables X (common variables) while the variable Y is observed just in A and Z is available just in B . In practice the synthetic data source $S = \{X, Y, Z\}$ is derived by exploiting the shared information, i.e., the common variables X (usually a subset of them) and, when available, eventual auxiliary information concerning the relationship among X , Y and Z or just Y and Z which can be in terms of an additional data source in which all the variables are jointly observed or an estimate of a parameter of interest (correlation matrix, contingency table, etc.). It is worth noting that when the matching is solely based on the available common variables (X), then the results of the matching will rely on a strong assumption of conditional independence of Y and Z given X . Hence the entire analysis carried out on the synthetic data set will reflect such assumption (Chapter 2, D'Orazio et al., 2006)

From the practical viewpoint, the synthetic data set can be simply one of the origin data sources (A or B) in which the values of the missing variable are imputed using techniques developed for imputing missing values in a survey. Usually it is preferred to refer to the smaller data source (in terms of observations) which becomes the recipient; the other one, the larger data sets, plays the role of the donor. In some cases it may happen that the synthetic data set is the result of concatenating the original data sources ($S = A \cup B$), then two imputation steps are required, Z is imputed in A while Y is imputed in B . The file concatenation procedure is proposed by Rubin (1986) in order to deal with data arising from complex sample surveys carried out from the same target population. A similar procedure is suggested by Renssen (1998) whose approach, based on weights calibration, is essentially developed for macro purposes (estimation of two-way contingency table $Y \times Z$). A discussion about the methods for statistical matching data from complex sample surveys can be found in the Report of WP1 of the ESSnet on Data Integration (2011, pp. 43-49).

The methods that can be used to impute the values for the missing variable in the recipient data set (or the concatenated file) can be based on a parametric, nonparametric or mixed approach. For the sake of simplicity, it will be considered the case of two i.i.d. samples *A* and *B* and the conditional independence (CI) is assumed to hold.

2.1 Parametric approach

A model characterised by a finite number of parameters is explicitly considered; once its parameters are estimated it is possible to impute the values of the missing variables via conditional expectation (conditional mean matching) or by drawing values from the predicted distribution.

2.2 Nonparametric approach

Many applications of statistical matching are based on the usage on nonparametric methods which do not require specifying in advance a model. The most used nonparametric techniques in statistical matching derive from hot deck methods applied in sample surveys to fill in missing values. Usually the objective is that of creating the synthetic data set by imputing the missing variables in the recipient data set. Imputed values are those observed in a similar statistical unit observed in the donor data set. Random hot deck and nearest-neighbour hot deck are the most used techniques in statistical matching (cf. Section 2.4, D'Orazio et al., 2006). A discussion about the use of hot deck techniques is in D'Orazio et al. (2006) and Singh et al. (1993). Paass (1985) and Conti et al. (2006) enlighten that such methods may introduce a matching noise, i.e., a discrepancy among the joint probability density function of the variables of interest in the synthetic data set the ones in the target population.

2.3 Mixed methods

This class of techniques mixes parametric and nonparametric approach. More precisely, in a first step a parametric model is adopted and its parameters are estimated, then, in the second step, a completed synthetic data set is obtained by means of some hot deck procedures. This approach exploits the advantages of models, being more parsimonious as far as estimation is concerned, and, on the other hand, provides imputed values that are not artificial (i.e., predicted by the model with possibly a random term) but are really observed (taken from the donor records). Interesting papers in this context are those of Rubin (1986), Singh et al. (1993), Moriarity and Scheuren (2001, 2003).

3. Preparatory phase

Before integrating two data sources through statistical matching some practical steps are necessary (Chapter 3, ESSnet-ISAD, 2009):

- i. identification of the common variables and harmonisation issues;
- ii. choice of the matching variables
- iii. Definition of a model (when using a parametric or mixed approach)

The harmonisation issue can be quite time consuming, because it may be necessary to harmonise the definition of units, the reference periods, the variables, the classifications etc. Sometimes harmonisation cannot be reached and if two variables available in both the data sources cannot be harmonised then they cannot be used as matching variables.

Once completed the harmonisation step, most of the matching methods listed in Section 3. require a crucial step for the choice of the matching variables X_M , i.e., the subset of the common variables ($X_M \subseteq X$) that should be used in the models or in computing distances among units. The commonly used approach to identify the set of matching variables consists in disregarding all those variables which are not statistically connected with Y or Z (Singh et al., 1988; Cohen, 1991). In this context it is possible to use methods commonly used to select the best subset of predictor when fitting regression models or nonparametric procedure based on the fitting of classification or regression trees. In the case of all categorical variables, D’Orazio (2011a) suggested a procedure which is based on the exploration of the uncertainty due to the statistical matching framework.

When it is necessary to specify a model it is worth noting that in the basic statistical matching framework since the variables (X,Y,Z) are not jointly observed on data, it is not possible to test the fit of the model to data. In this case, experts in the phenomena under investigation can provide guidance on the choice of the model. Another possibility consists in considering different alternative models where different results are evaluated (a kind of sensitivity analysis). It is worth noting that a mixed approach offers a certain level of protection against model misspecification if compared to a fully parametric one.

4. Examples – not tool specific

5. Examples – tool specific

Some statistical matching methods are implemented in a specific library, called “StatMatch” (D’Orazio, 2011b), made freely available for the R environment (R Development core team, 2014). As far as statistical matching at micro level is concerned the following functions are available:

- (a) functions to perform nonparametric statistical matching at micro level by means of hot deck imputation (NND.hotdeck, RANDwNND.hotdeck, rankNND.hotdeck). The following examples taken from D’Orazio (2011a) show how to use the functions:

```
# example of usage of nearest-neighbour hotdeck
# with the R function NND.hotdeck

> group.v <- c("rb090", "db040")
> X.mtc <- c("hsize", "age")
> out.nnd <- NND.hotdeck(data.rec=rec.A, data.don=don.B,
+                       match.vars=X.mtc, don.class=group.v,
+                       dist.fun="Manhattan")

# to derive the sythetic data set
> fA.nnd.m <- create.fused(data.rec=rec.A, data.don=don.B,
+                         mtc.ids=out.nnd$mtc.ids,
+                         z.vars=c("netIncome", "c.netI"))

# example of random hotdeck
# with the R function RANDwNND.hotdeck

> group.v <- c("db040", "rb090")
```



```

> rnd.1 <- RANDwNND.hotdeck(data.rec=rec.A, data.don=don.B,
+                           match.vars=NULL, don.class=group.v)
> fA.rnd <- create.fused(data.rec=rec.A, data.don=don.B,
+                       mtc.ids=rnd.1$mtc.ids,
+                       z.vars=c("netIncome", "c.netI"))

> rnk.1 <- rankNND.hotdeck(data.rec=rec.A, data.don=don.B,
+                          var.rec="age", var.don="age")
> fA.rnk <- create.fused(data.rec=rec.A, data.don=don.B,
+                       mtc.ids=rnk.1$mtc.ids,
+                       z.vars=c("netIncome", "c.netI"),
+                       dup.x=TRUE, match.vars="age")

```

- (b) a function to perform mixed SM at micro level for continuous variables (mixed.mtc); the following example is taken from D’Orazio (2011a):

```

> X.mtc <- c("Sepal.Length", "Sepal.Width") # matching variables
# parameters estimated using ML
> mix.1 <- mixed.mtc(data.rec=iris.A, data.don=iris.B,
+                   match.vars=X.mtc, y.rec="Petal.Length",
+                   z.don="Petal.Width", method="ML", rho.yz=0,
+                   micro=TRUE, constr.alg="lpSolve")

> mix.1$filled.rec # provides A filled in with Z

```

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Cohen, M. L. (1991), Statistical matching and microsimulation models. In: Citro and Hanushek (eds.), *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. Vol II Technical papers*, Washington D.C.
- Conti, P. L., Marella, D., and Scanu, M. (2006), Nonparametric evaluation of matching noise. *Proceedings of the IASC conference “Compstat 2006”, Roma, 28 August – 1 September 2006*, Physica-Verlag/Springer, 453–460.
- ESSnet on Data Integration (2011), *Report on WPI State of the art on statistical methodologies for data integration*. <http://www.cros-portal.eu/content/wp1-state-art>
- D’Orazio, M. (2011a), Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment. *R Package Vignette*. http://rm.mirror.garr.it/mirrors/CRAN/web/packages/StatMatch/vignettes/Statistical_Matching_with_StatMatch.pdf
- D’Orazio, M. (2011b), StatMatch: Statistical Matching. R package version 1.0.3. <http://CRAN.R-project.org/package=StatMatch>
- D’Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical Matching, Theory and Practice*. Wiley, Chichester.

- ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data (2009), *Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data*. <http://cenex-isad.istat.it/>
- Little, R. J. A and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd Edition. Wiley, New York.
- Moriarity, C. and Scheuren, F. (2001), Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* **17**, 407–422.
- Moriarity, C. and Scheuren, F. (2003), A note on Rubin’s statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics* **21**, 65–73.
- Paass, G. (1985), Statistical record linkage methodology: state of the art and future prospects. *Bullettin of the International Statistical institute, Proceedings of the 45th Session*, vol. LI, Book 2, Voorburg, The Netherlands.
- R Development Core Team (2014), *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing*, Vienna, Austria. <http://www.R-project.org/>
- Renssen, R. H. (1998), Use of Statistical Matching Techniques in Calibration Estimation. *Survey Methodology* **24**, 171–183.
- Rubin, D. B. (1986), Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business and Economic Statistics* **4**, 87–94.
- Singh, A. C., Mantel, H., Kinack, M., and Rowe, G. (1993), Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology* **19**, 59–79.

Specific section

8. Purpose of the method

Statistical matching (SM) techniques when applied at micro level aim at integrating the available data sources, related to the same target population, in order to derive a unique synthetic data set in which all the variables (coming from the different sources) are jointly available. The synthetic data set is the basis of further statistical analysis, e.g., microsimulations.

9. Recommended use of the method

1. Statistical matching techniques usually are applied to investigate the relationship between two variables, Y and Z , never jointly observed in the available data sources, by considering the available common information, usually X variables. When no auxiliary information is available the statistical matching is based on the conditional independence of Y and Z given X ; unfortunately, this assumption cannot be tested on the available data. If the analyst does not consider it to be valid then the SM cannot be performed.

10. Possible disadvantages of the method

- 1.

11. Variants of the method

1. Parametric approach: conditional mean matching

The conditional mean matching in the simple case of three continuous variables X , Y , and Z reduces to a regression imputation; the recipient data set A is filled in with the predicted values:

$$\hat{z}_k^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_k, \quad k = 1, 2, \dots, n_A$$

The parameters of the model should be estimated in order to exploit all the available information in both the data sets. For instance, the simple estimation of the parameters obtained through their observed counterpart may lead to unacceptable results, like a non-positive semi-definite covariance matrix. A solution to this problem is to use the maximum likelihood estimation (cf. D'Orazio et al., 2006, pp. 16-19). A discussion of the problems concerning the combination of estimates obtained from the different data sets is in Moriarity and Scheuren (2001, 2003) and D'Orazio et al. (2006). Extension to this method to the multivariate case are provided in D'Orazio et al. (2006).

2. Parametric approach: stochastic regression imputation

Regression imputation provides values lying on the regression and there is no variability around it. For this reason in the case of the previous example it would be better to refer to stochastic regression imputation, such that the imputed value is obtained as (cf. Little and Rubin, 2002):

$$\tilde{z}_k^{(A)} = \hat{z}_k^{(A)} + e_k = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_k + e_k, \quad k = 1, 2, \dots, n_A$$

being e_k a residual generated randomly from a normal distribution with zero mean and variance equal to the estimated residual variance $\hat{\sigma}_{z|x}$. This is an example of the drawings based on the conditional predictive distributions. Extension to this method to the multivariate case are provided in D’Orazio et al. (2006).

3. Nonparametric approach: Random hot deck

Random hot deck consists in randomly choosing a donor record in the donor file for each record in the recipient file. The random choice is often done within groups obtained by considering subsets of homogeneous units characterised by presenting the same values for one or more common variables X (usually categorical).

4. Nonparametric approach: nearest-neighbour hot deck

Nearest-neighbour hot deck is widely used in the case of continuous variables. The donor unit is the closest to the given recipient units in terms of a distance measured by considering all or a subset of the common variables X . The distance can be measured in different ways (cf. Appendix C in D’Orazio et al., 2006). Sometimes the search of the donors is restricted to suitable subsets of the donor units, sharing the same characteristics of the recipient unit (as for random hot deck).

The constrained nearest-neighbour hot deck represents an interesting variation of the nearest-neighbour hot deck. In this approach, each donor record can be chosen as donor only once: the subset of the donors to choose is the one obtained as a solution of the transportation problem whose objective is the minimisation of the overall matching distance (sum of the recipient-donor distances). This constraint helps in better preserving of the marginal distribution of the imputed variable in the synthetic data set.

In general the methods based on distances pose the problem of deciding the subset of the common variables X to be used for computing it. Using all or too many common variables may affect negatively the matching results because variables with low predictive power on the target variable may influence negatively the distances.

5. Nonparametric approach: rank hot deck

Singh et al. (1993) proposed the usage of the rank hot deck distance method; it searches for the closest donor for the given recipient record with distance computed on the percentage points of the empirical cumulative distribution function of the (continuous) common variable X being considered. Considering the percentage points of the empirical cumulative distribution provides values uniformly distributed in the interval $[0,1]$; moreover, this permits to compare observations when the values values of X cannot be directly compared because of measurement errors which however do not affect the “position” of a unit in the whole distribution.

6. Mixed approach: stochastic regression imputation followed by nearest-neighbour hot deck

In case of continuous variables, the procedure resembles the *predictive mean matching* imputation methods; let A play the role of recipient then procedure follows these steps:

- (step 1) Estimate (on B) the regression parameters of Z on X ; then use the model to impute the predicted values of Z in A (it is preferable to add a residual error term to the predicted values);
- (step 2) For each record in A impute the value of Z observed on the closest value in B according to a distance computed on the values of Z (predicted values of Z in A and truly observed values of Z in B).

Such a two steps procedure presents various advantages: it offers protection against model misspecification and also reduces the risk of bias in the marginal distribution of the imputed variable because the distances are computed on intermediate and truly observed values of the target variable, instead of a suitable subset of the common variables X . In fact when computing the distances by considering all the matching variables, variables with low predictive power on the target variable may influence negatively the distances. Various alternative similar mixed procedures are listed in D'Orazio et al. (2006, Section 2.5).

12. Input data

1. Ds-input1: is the data set that contains data referred to the variables X and variable(s) Y , usually denoted as A in the statistical matching framework. This data set contains n_A records (observations) usually representing a sample of i.i.d. observations or the results of a complex sample survey carried out on a given finite population U .
2. Ds-input2: is the data set that contains data referred to the variables X and variable(s) Z , usually denoted as B in the statistical matching framework. This data set contains n_B records (observations) usually representing a sample of i.i.d. observations or the results of a complex sample survey carried out on a given finite population U .
3. Ds-input3: this is an optional data set that may be available in statistical matching as a source of auxiliary information. In such case it may contain all the necessary variables X , Y and Z or just Y and Z , the variables that are never jointly observed in the two basic input data sets (DS-input1 and ds-input2); usually this data set is denoted as C in the statistical matching framework and it contains n_C records (observations) usually representing a sample of i.i.d. observations or the results of a complex sample survey carried out on the same finite population U but in the past or on a smaller scale.

13. Logical preconditions

1. Missing values
 1. Usually the common variables are expected to be free of missing values and the same happens as far as the target variables are concerned. In some applications of nearest-neighbour hot deck it is possible to refer to distance functions that account for the missing values of the matching variables.
2. Erroneous values
 - 1.
3. Other quality related preconditions

- 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

- 1.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. Ds-output1: The output of the statistical matching at micro level is a synthetic data set in which all the interest variables X , Y and Z are available. The synthetic data set can be simply one of the origin data sources (ds-input1 or ds-input2) in which the values of the missing variables are imputed using methods listed before. Usually it is preferred to refer to the smaller data source (in terms of observations) which becomes the recipient; the other one, the larger data sets, plays the role of the donor. In some cases it may happen that the synthetic data set is the result of concatenating the origin data sources ($S = A \cup B$).

17. Properties of the output data

- 1.

18. Unit of input data suitable for the method

19. User interaction - not tool specific

- 1.

20. Logging indicators

- 1.

21. Quality indicators of the output data

- 1.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

- 1.

24. Related methods described in other modules

- 1.

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

1. R library *StatMatch* (D’Orazio, 2011b), made freely available for the R environment

28. Process step performed by the method

GSBPM Sub-process 5.1: Integrate data

Administrative section

29. Module code

Micro-Fusion-M-Statistical Matching Methods

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-03-2012	first version	Marcello D'Orazio	Istat (Italy)
0.2	02-05-2012	second version	Marcello D'Orazio	Istat (Italy)
0.3	25-09-2013	EB comments	Marcello D'Orazio	Istat (Italy)
0.3.1	03-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:59



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Reconciling Conflicting Microdata

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Composite records arising in micro-fusion and imputation	3
2.2 Introduction to the micro-level consistency problem	4
2.3 Overview of adjustment methods to achieve consistency	6
3. Preparatory phase	7
4. Examples – not tool specific.....	7
5. Examples – tool specific.....	7
6. Glossary.....	7
7. References	7
Specific section.....	8
Interconnections with other modules.....	9
Administrative section.....	11

General section

1. Summary

In data fusion we consider microdata consisting of records that are composed of information from different sources. Such composite records may consist of several combinations of sources (see the module “Micro-Fusion – Data Fusion at Micro Level”). Records may be a combination of values obtained from a register with values obtained from a survey for the same units (obtained by record linkage). Records may also combine information from several surveys with non-overlapping units, in which case a unit from one source is matched with a similar (but not identical) unit from another source. In addition, records with values obtained from different sources can also arise as a consequence of item non-response and subsequent imputation in which case the two sources are the directly observed values versus the values generated by the imputation method.

In all these cases the composition of a record by combining information obtained from different sources may give rise to consistency problems because the information is conflicting in the sense that edit rules that involve variables obtained from the different sources will often be violated.

The purpose of reconciling conflicting microdata is to solve the consistency problems by making slight changes or adjustments to some of the variables involved. Apart from the choice of variables to be adjusted, an adjustment method should also be specified since there are a number of methods to handle the adjustment problem. In this module three different approaches to the reconciliation problem will be described and the properties of the solutions will be discussed.

2. General description of the method

2.1 *Composite records arising in micro-fusion and imputation*

In this module we are concerned with the task of reconciling conflicting information in statistical microdata that may arise if (some of) the individual records are composed of data obtained from different sources. In the module “Micro-Fusion – Data Fusion at Micro Level” two general cases have been described that give rise to such composite units: record linkage (see also the theme module “Micro-Fusion – Object Matching (Record Linkage)”) and statistical matching (see also the theme module “Micro-Fusion – Statistical Matching”). In addition, imputation for non-response (see the topic “Imputation”) also creates a composite record. Thus we have the following three situations in which composite records can arise:

Record linkage

This type of data fusion, which is a common and increasing practice in the production of business statistics, concerns the linkage of (usually) a sample survey to a register. In this case the linked records consist of register information combined and enriched with survey information on the same units. Both sources will usually also have a few variables in common, apart from the variables used to identify the unit that are necessary for the linking process. In business statistics the main administrative source today is the tax register, providing information on at least the total turnover, which will be a common variable since it will also be measured in the survey. It should be noted that such common variables may have different values in the register and the survey.

Statistical matching

The second case concerns the integration of two (or more) sample surveys which have some variables in common while others are specific for each of the sources. Let the set of common variables be denoted by X and the sets of specific variables by Y and Z . Usually the samples will be (almost) non overlapping and therefore there will be no units with all sets of variables observed. In this case synthetic records can be constructed from one of the sources, say with Y observed, by filling in or imputing the variables Z . These imputations can be obtained by a regression model relating Z to X , which can be estimated using the other source where both Z and X are observed. Alternatively a hot-deck imputation method can be used where values for Z are obtained from a similar record from the other source, found by matching on the common variables X (see Figure 2 and the accompanying text in the module “Micro-Fusion – Data Fusion at Micro Level” or D’Orazio et al., 2006). In the case of hot-deck imputation the composite record consists of values obtained from different but similar units.

Imputation

Records with values obtained from different sources can also arise as a consequence of item non-response and subsequent imputation. In this case one of the sources of the composite record consists of observed values and the other of imputed values derived from a parametric or nonparametric imputation model. This situation is similar to the one arising from statistical matching since in both cases the composite record consists of observed and imputed values. The difference is, however, that the synthetic records in statistical matching all have the same variables imputed, while in the item non-response case the non-response pattern and hence the variables requiring imputation, can be different for each record.

2.2 Introduction to the micro-level consistency problem

To illustrate the consistency problem at micro level, we consider the following situation that arises in business statistics (cf. Pannekoek, 2011). There is information on some key variables available from reliable administrative data. Let these variables be the total turnover (*Turnover*), the number of employees (*Employees*) and total amount of wages paid (*Wages*). These variables are used to compile the short term economic statistics (STS) and are published quarterly as well as yearly. The yearly structural business statistics (SBS), requires much more detail and this more detailed information is not available from registers. Therefore, a sample survey is conducted to obtain the additional details. After linking the sample data to the register, the situation arises that for the key variables, two sources are available for each responding unit in the sample: the register value and the survey value and for the other variables only survey values are obtained. To be consistent with already published STS figures on *Turnover* and possibly other key variables, the register values are used for the key variables and the survey values for the other variables. Thus we create composite records based on two sources: register and survey. This is illustrated in table 1 below. The column *Survey values* displays the survey values of the eight variables for a responding unit. In the column *Composite (I)* the values of the composite record are shown; the survey values for the key variables are replaced by the register values (in bold). As an alternative we also consider, for illustrative purposes, the situation that we only have *Turnover* available from administrative sources resulting in the values in the column *Composite (II)*.

Business records generally have to adhere to a number of accounting rules and logical constraints. These constraints are widely employed for checking the validity of a record and are, in this context,

referred to as edit rules (see “Statistical Data Editing – Main Module”). For the example record above, the following three edit rules are formulated:

$$e_1: x_1 - x_5 + x_8 = 0 \text{ (Profit = Turnover - Total Costs)}$$

$$e_2: -x_3 + x_5 - x_4 = 0 \text{ (Turnover = Turnover main + Turnover other)}$$

$$e_3: -x_6 - x_7 + x_8 = 0 \text{ (Total Costs = Wages + Other costs)}$$

Notice that these edits are connected by the variables *Turnover* and *Total Costs*, which is true for many of the edits used in business statistics and has consequences for adjustment for consistency.

Table 1. Example Business record with data from two sources

Variable	Name	Survey values	Composite (I)	Composite (II)
x_1	Profit	330	330	330
x_2	Employees (Number of employees)	20	25	20
x_3	Turnover main (Turnover main activity)	1000	1000	1000
x_4	Turnover other (Turnover other activities)	30	30	30
x_5	Turnover (Total turnover)	1030	950	950
x_6	Wages (Costs of wages and salaries)	500	550	500
x_7	Other costs	200	200	200
x_8	Total costs	700	700	700

Both composite records lead to violation of the edit rules, which we refer to as the micro-level consistency problem. In particular, composite record (I) violates all three edit rules and composite record (II) violates the two edit rules involving *Turnover*. To obtain a consistent record some of the values have to be changed or “adjusted”. Since the register values are considered reliable and already used in publications, the survey values are an obvious choice in this case.

When the data are obtained from a single source, e.g., a single survey questionnaire, the violation of hard edit rules that describe relations between variables, such as the balance edits, indicate that a response error has occurred. When data are from different sources, edit rules that describe relations between variables can also be violated by (slight) differences in definitions of variables or time differences between the two sources. In such cases the cause of the violation need not be a response error and is therefore termed an inconsistency between the sources.

The example above is just a simple illustration, in practice the number of variables as well as the number of edit rules can be much larger. The structural business statistics (SBS) are an example with a large number of variables and edit rules. An SBS questionnaire can be divided in sections. It contains, for instance, sections on employees, revenues, costs and results. In each of these sections a total is broken down in a number of components that can again be broken down in sub-components. Components of the total number of employees can be part-time and full-time employees and components of total revenues may be subdivided in turnover and other operating revenues. The total costs can have as components: purchasing costs, depreciations, personnel costs and other costs. Each of these breakdowns of a (sub)total corresponds to what is called a balance edit. SBS questionnaires also contain a profit and loss section where revenues are balanced against the costs to obtain the

results (profit or loss), which leads to edits of the form e_1 . This last type of edit connects the edits from the costs section with the edits from the revenues section. Therefore, almost all variables are connected by edit rules and changing one variable will lead to necessary changes in most other values if the structure as laid down in the edit rules is to be preserved. In some cases there is no explicit connection between variables specifying employment in terms of the numbers of employees in different categories and the other, financial, variables. Since relations between, e.g., number of employees and wages should be preserved, adjustment methods should take care of relations not specified by edit rules and methods to accomplish this are the method described in the module “Micro-Fusion – Generalised Ratio Adjustments” and an approach discussed in the module “Micro-Fusion – Minimum Adjustment Methods” (section 2.5.2).

2.3 *Overview of adjustment methods to achieve consistency*

Adjustment methods change (or adjust) some of the values of some variables (the adjustable variables) in a record such that the resulting adjusted record satisfies all the specified edit constraints. Three different adjustment methods are treated in three separate modules: “Micro-Fusion – Prorating”, “Micro-Fusion – Minimum Adjustment Methods”, and “Micro-Fusion – Generalised Ratio Adjustments”. Below we give a short overview of these methods.

Prorating is a simple ratio adjustment for balance edits (see Banff Support Team, 2008). It solves the possible inconsistencies for each constraint separately. It is an intuitively appealing method that is easy to interpret and to apply. For composite record (II) in table 1, a prorating adjustment to resolve the violation of edit-rule e_2 would entail multiplying the components of *Turnover*, x_3 and x_4 , by the ratio of the register and survey values for *Turnover* (1030/950). This ratio adjustment has the effect that the ratios of the components of turnover to their total become equal to the values of these ratios obtained from the survey, but the levels of the components are consistent with the register value of the total. This reflects the availability of information in the two sources and the priority of the total from the register. A drawback of this method is that for interrelated balance edits the result is dependent on the order in which the edits are treated, which introduces arbitrariness in the solution. In practice different orders can indeed lead to substantially different solutions. Especially for the extensive systems of balance edits encountered in the SBS this can be a problem. This method is treated in more detail in the module “Micro-Fusion – Prorating”.

The minimum adjustment approach is to make adjustments to the adjustable variables that are minimal in some sense, such that the adjusted record satisfies all constraints (see Pannekoek, 2011). The minimal adjustments are thus obtained by minimising a chosen distance metric subjected to the edit constraints. Since this optimisation approach treats all edits simultaneously there is no problem with the order in which the edits are handled and it leads to a single optimal solution. This solution does, however, depend on the chosen optimisation criterion. In the module “Micro-Fusion – Minimum Adjustment Methods” the optimisation approach is described and properties of the solutions for three different optimisation criteria are discussed. Some solutions are characterised by additive adjustments that preserve the differences between variables that are part of the same (set of) constraint(s) and other solutions are characterised by multiplicative constraints that preserve the ratios between variables that are part of the same (set of) constraint(s).

The third adjustment method is generalised ratio adjustment (see Pannekoek and Zhang, 2011). The method uses multiplicative adjustments, just as the methods Prorating and one of the minimum

adjustment methods (the KL-adjustments, see the module “Micro-Fusion – Minimum Adjustment Methods”). The generalised ratio adjustments method aims to make the adjustments as uniform as possible. Furthermore, and in contrary to the other methods, the method can result in adjustments to variables that are not involved in the constraints. In this sense it can solve the problem, mentioned at the end of the previous section, of preserving relations between variables that are not connected by edit rules.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Banff Support Team (2008), Functional Description of the Banff System for Edit and Imputation. Technical Report, Statistics Canada.

D’Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical matching: theory and practice*. John Wiley, Chichester.

Pannekoek, J. (2011), Models and algorithms for micro-integration. In: *Report on WP2: Methodological developments*, ESSNET on Data Integration, available at <http://www.cros-portal.eu/content/wp2-development-methods>.

Pannekoek, J. and Zhang, L.-C. (2011), Partial (donor) imputation with adjustments. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing.

van der Loo, M. (2012), *rspa: Adapt numerical records to fit (in)equality restrictions with the Successive Projection Algorithm*. R package version 0.1-1.

Specific section

8. Purpose of the method

The purpose of the method is to adjust the values of some variables in a data record to remove edit violations to ensure consistency of the data values obtained from different sources.

9. Recommended use of the method

1. The method should be used after detection and treatment of errors and missing values.

10. Possible disadvantages of the method

1. When inconsistencies arise due to large errors in some values, these errors may propagate to other values due to adjustment. Influential errors should therefore be treated before the method is applied.

11. Variants of the method

1. Prorating
2. Minimum adjustment methods
3. Generalised ratio adjustments

12. Input data

1. Data records with possibly inconsistent values and edit rules.

13. Logical preconditions

1. Missing values
 1. Missing values are allowed but edit rules involving variables with missing values cannot be checked and no adjustment with respect to these edit rules will take place.
2. Erroneous values
 1. Influential erroneous values should be treated before the method is applied.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. The amount of change applied to individual variables can be controlled by specifying weights for the variables.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. The output consists of the same individual records as the input, with values adapted when needed to ensure consistency with the edit rules.

17. Properties of the output data

1. The output data are ensured to be consistent with all specified edit rules that do not involve variables with missing values.

18. Unit of input data suitable for the method

The input consists of individual records that are treated one-by-one, independently.

19. User interaction - not tool specific

- 1.

20. Logging indicators

- 1.

21. Quality indicators of the output data

- 1.

22. Actual use of the method

1. Adjustments of imputed values to ensure that edit rules are satisfied is used in the production process for Structural Business Statistics at Statistics Netherlands.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Data Fusion at Micro Level
2. Micro-Fusion – Object Matching (Record Linkage)
3. Micro-Fusion – Statistical Matching
4. Statistical Data Editing – Main Module
5. Statistical Data Editing – Editing Administrative Data
6. Imputation – Main Module

24. Related methods described in other modules

1. Micro-Fusion – Prorating
2. Micro-Fusion – Minimum Adjustment Methods
3. Micro-Fusion – Generalised Ratio Adjustments

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

Available software options vary for the three (classes) of methods discussed in this module: prorating, the optimisation approach and generalised ratio adjustment.

1. Statistics Canada's generalised edit and imputation software Banff, contains a routine PRORATE that provides an off-the-shelf, generalised prorating application. However, for specific applications the prorating calculations are not difficult to implement. So, without the availability of generalised prorating software, the application of prorating could be performed by an ad hoc implementation using general statistical packages with programming facilities such as R or SAS.
2. The optimisation methods can be implemented, in general, by using standard (commercially) available solvers for convex optimisations problems and the same holds for the generalised ratio approach. For the optimisation methods based on Least Squares and Weighted Least Squares a specific R-package is freely available (van der Loo, 2012).

28. Process step performed by the method

GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

29. Module code

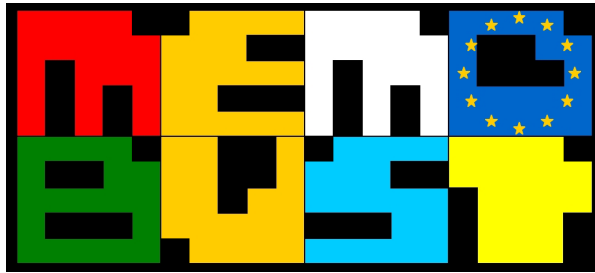
Micro-Fusion-M-Reconciling Conflicting Microdata

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-03-2013	first version	Jeroen Pannekoek	CBS (Netherlands)
0.2	17-04-2013	second version	Jeroen Pannekoek	CBS (Netherlands)
0.2.1	09-09-2013	preliminary release		
0.3	20-12-2013	improvements based on the EB-review	Jeroen Pannekoek	CBS (Netherlands)
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:00



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Prorating

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 The prorating method	3
2.2 Weighted prorating.....	4
3. Preparatory phase	4
4. Examples – not tool specific.....	4
4.1 Prorating applied in two different orders.....	4
5. Examples – tool specific.....	5
6. Glossary.....	5
7. References	5
Specific section.....	7
Interconnections with other modules.....	8
Administrative section.....	10

General section

1. Summary

Prorating is a simple method to reconcile conflicting information as described in the module “Micro-Fusion – Reconciling Conflicting Microdata”. The method is designed for equality edits, especially with business statistics in mind, where often a total (turnover, costs etc.) is broken down into a number of specifications (turnover from different activities, different kinds of costs). Inconsistencies arising when the specifications do not add up to the total are often handled by prorating. The method handles a single edit rule at a time and is therefore in practice applied to each of the edit rules one by one. This has the drawback that the order in which the edits are treated does matter and quite different results can be obtained by different orders. This drawback has led to the more principled approaches described in the modules “Micro-Fusion – Minimum Adjustment Methods” and “Micro-Fusion – Generalised Ratio Adjustments”.

2. General description of the method

2.1 The prorating method

Consider the following situation described in the module “Micro-Fusion – Reconciling Conflicting Microdata”. In a business data set obtained by linking a survey to an administrative source, we observe for some unit the following values for three variables (x_3, x_4, x_5) describing turnover:

x_3 : <i>Turnover main</i>	x_4 : <i>Turnover other</i>	x_5 : <i>Turnover total</i>
1000	30	950

The variable *Turnover total* is obtained from an administrative source while the component variables are observed in a survey. An inconsistency arises because the sum of x_3 and x_4 is 1030 instead of 950 and the edit rule $x_5 = x_3 + x_4$ is violated. Suppose that the administrative value of *Turnover total* is not to be changed but the other values may be changed in order to make the record consistent. The prorating method (Banff Support Team, 2008; Pannekoek, 2011; Pannekoek and Zhang, 2011) changes the adjustable values by a uniform multiplicative adjustment. Thus, in this case, the adjusted values for x_3 and x_4 become $(950/1030) \times 1000$ and $(950/1030) \times 30$.

For a general description it is convenient to express the equality edit in the form $\sum_i x_i = 0$, which involves changing the sign of some of the original variables. In the example above this could be accomplished by defining $x_5 + (-x_3) + (-x_4) = 0$. Furthermore, let δ denote the prorating factor and let I_{fre} and I_{fix} be the index sets of, respectively, the adjustable (free) variables and the un-adjustable (fixed) variables. Then, the adjusted values are given by

$$\tilde{x}_i = \delta x_i \text{ for } i \in I_{fre}. \quad (1)$$

Now, since we must have $\sum_{i \in I_{fre}} \tilde{x}_i + \sum_{i \in I_{fix}} x_i = 0$ for the adjusted values to satisfy the equality edit, we can write:

$$\delta \sum_{i \in I_{fre}} x_i = - \sum_{i \in I_{fix}} x_i, \text{ and so,}$$

$$\delta = - \sum_{i \in I_{fix}} x_i / \sum_{i \in I_{fre}} x_i. \quad (2)$$

From (1) we can see that for a solution to this adjustment problem it is necessary that there are free variables with non-zero values which is understandable because a multiplicative adjustment would otherwise be ineffective.

2.2 Weighted prorating

A weighted version of the prorating method makes it possible to control the relative amount of change in the free variables. A weight is assigned to each free variable and the amount of change is inversely proportional to the weight.

In this case we can write, for the adjusted values,

$$\tilde{x}_i = \frac{\delta}{w_i} x_i \text{ for } i \in I_{fre}. \quad (3)$$

Furthermore, since we must have $\sum_{i \in I_{fre}} \tilde{x}_i = - \sum_{i \in I_{fix}} x_i$, we obtain the following expression for δ :

$$\delta = - \sum_{i \in I_{fix}} x_i / \sum_{i \in I_{fre}} \frac{x_i}{w_i}. \quad (4)$$

3. Preparatory phase

4. Examples – not tool specific

4.1 Prorating applied in two different orders

Prorating is defined as a treatment for a single edit inconsistency. It also applies to several edit inconsistencies without complications as long as the edits have no variables in common. However, it does not, in itself, provide a unique solution for systems of connected edits. For such cases, a strategy is followed that involves treating the edits in a predefined order and fixing each variable that has been treated (see Banff Support Team, 2008). This is illustrated in the example below.

In this example we show the results of applying prorating with two different orders to resolve the violation of the edit rules for the values of the business record shown in Table 1 of the module “Micro-Fusion – Reconciling Conflicting Microdata”, column *Composite (I)*. The data in this column consist of administrative values for the variables in bold in Table 1 below, *Employees*, *Turnover* and *Wages*, and values observed in a survey for the other variables. This composite record violates three edit rules:

$$e_1: x_1 - x_5 + x_8 = 0 \text{ (Profit = Turnover - Total Costs);}$$

$$e_2: -x_3 + x_5 - x_4 = 0 \text{ (Turnover = Turnover main + Turnover other);}$$

$$e_3: -x_6 - x_7 + x_8 = 0 \text{ (Total Costs = Wages + Other costs).}$$

Now, we assume that the administrative values are fixed and adjust the other values by prorating so that the three edit rules are satisfied. The result for the edit e_2 is independent of the order in which prorating is applied because the free variables in this edit do not appear in other edits and are only adjusted to sum up to the total *Turnover*. The order in which the edit rules e_1 and e_3 are treated does make a difference for the result because these variables have a free variable (*Total costs*) in common. If a top-down strategy is followed in which first the edit e_1 is treated (which entails adjustment of *Profit* and *Total costs*) and then the edit e_3 is treated (which amounts to adjustment of *Other costs*), we obtain the results in the column “ e_1 adjusted first”. If the prorating adjustments are applied the other way around, that is first treating e_3 (which in this case entails adjusting *Other costs* and *Total costs*) and then e_1 , we obtain the results in the column “ e_3 adjusted first”.

The differences in the results for the two different orders are quite large as are the adjustments themselves. If e_1 is treated first, this results in a moderate proportional downwards adjustment of *Profit* and *Total costs* to make them sum up to 950. When *Other costs* is adjusted next, the adjustment is very large because before adjustment *Other costs* was already larger than *Total costs* – *Wages* and since in the first step *Total costs* was reduced this discrepancy has become larger so that *Other costs* has to be reduced by more than 50%. For the other order in which e_3 is treated first, the adjustment to *Other costs* is only 10% but in this case we end up with a very large downwards adjustment of *Profit*.

Table 1. Example business record: prorating using two orders of application.

Variable	Name	Unadjusted	e_1 adjusted first	e_3 adjusted first
x_1	Profit	330	304	180
x_2	Employees	25	25	25
x_3	Turnover main	1000	922	922
x_4	Turnover other	30	28	28
x_5	Turnover	950	950	950
x_6	Wages	550	550	550
x_7	Other costs	200	96	220
x_8	Total costs	700	646	770

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Banff Support Team (2008), Functional Description of the Banff System for Edit and Imputation. Technical Report, Statistics Canada.

Pannekoek, J. (2011), Models and algorithms for micro-integration. In: *Report on WP2: Methodological developments*, ESSNET on Data Integration, available at <http://www.cros-portal.eu/content/wp2-development-methods>.

Pannekoek, J. and Zhang, L.-C. (2011), Partial (donor) imputation with adjustments. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing.

Specific section

8. Purpose of the method

The purpose of the method is to adjust the values of some variables in a data record to remove violations of balance edits by a uniform multiplicative adjustment to some variables involved in the edit.

9. Recommended use of the method

1. The method should be used after detection and treatment of errors and missing values.

10. Possible disadvantages of the method

1. The order in which the edit rules are treated can influence the result.

11. Variants of the method

1. Unweighted prorating
2. Weighted prorating

12. Input data

1. Data records with possibly inconsistent values and edit rules.

13. Logical preconditions

1. Missing values
 1. Edits with missing values cannot be handled by this method.
2. Erroneous values
 1. Influential erroneous values should be treated before the method is applied.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. The amount of change applied to individual variables can be controlled by specifying weights for the variables.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. The output consists of the same individual records as the input, with values adapted when needed to ensure consistency with the edit rules.

17. Properties of the output data

1. In the output data inconsistencies with respect to equality edits that existed in the input data are resolved.

18. Unit of input data suitable for the method

19. User interaction - not tool specific

- 1.

20. Logging indicators

- 1.

21. Quality indicators of the output data

- 1.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Data Fusion at Micro Level
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Editing Administrative Data
4. Imputation – Main Module

24. Related methods described in other modules

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Micro-Fusion – Minimum Adjustment Methods
3. Micro-Fusion – Generalised Ratio Adjustments

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

1. Statistics Canada's generalised edit and imputation software Banff contains a routine PRORATE that provides an off-the-shelf, generalised prorating application. However, for specific applications the prorating calculations are not difficult to implement. So, without the availability of generalised prorating software, the application of prorating could be performed by an ad hoc implementation using general statistical packages with programming facilities such as R or SAS.

28. Process step performed by the method

GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

29. Module code

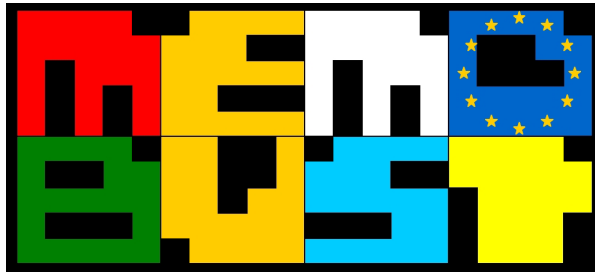
Micro-Fusion-M-Prorating

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-03-2013	first version	Jeroen Pannekoek	CBS (Netherlands)
0.2	17-04-2013	second version	Jeroen Pannekoek	CBS (Netherlands)
0.3	10-12-2013	third version	Jeroen Pannekoek	CBS (Netherlands)
0.3.1	12-12-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:00



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Minimum Adjustment Methods

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Formal description of the optimisation problem	3
2.2 Least squares adjustments	5
2.3 Weighted least squares adjustments	6
2.4 Kullback-Leibler adjustments	7
2.5 Generalisations: Adjusting to multiple sources and soft constraints	7
3. Preparatory phase	9
4. Examples – not tool specific.....	9
4.1 Comparison of distance functions using the example record	9
5. Examples – tool specific.....	10
6. Glossary.....	10
7. References	10
Specific section.....	11
Interconnections with other modules.....	12
Administrative section.....	14

General section

1. Summary

The problem of reconciling possibly conflicting information as described in the module “Micro-Fusion – Reconciling Conflicting Microdata” can be treated by an optimisation approach. In this approach, the values in the record with inconsistent microdata are changed, as little as possible, such that the modified record with microdata is consistent in the sense that it satisfies all edit rules. Formally then, the minimum adjustment method can be described as minimising a chosen distance between the original (inconsistent) record and the adjusted record, subject to the constraint that all edit rules are satisfied by the adjusted record. By specifying different distance functions, the minimum adjustment approach leads to different methods. For three common and, for the adjustment problem plausible, distance functions the corresponding adjustment methods will be described in this module and their differences will be illustrated by a numerical example.

2. General description of the method

2.1 Formal description of the optimisation problem

The optimisation approach resolves inconsistencies in data records with numerical variables that are required to adhere to a set of specified linear edit rules. The numerical variables in a record are denoted by x_i with $i = (1, \dots, n)$ and can be represented as a vector of variables: $\mathbf{x} = (x_1, x_2, \dots, x_n)$. The general form of a linear edit rule is as follows (see the module “Statistical Data Editing – Automatic Editing”):

$$e_{j1}x_1 + \dots + e_{jn}x_n - c_j = 0, \quad (1)$$

for equalities and

$$e_{j1}x_1 + \dots + e_{jn}x_n - c_j \geq 0 \quad (2)$$

for inequalities. Where $j = (1, \dots, J)$ numbers the edit rules, e_{ji} are numerical coefficients and c_j are numerical constants.

To describe the minimum adjustment methods it is convenient to express the edit rules in matrix notation. The equalities (1) can be expressed as $\mathbf{E}\mathbf{x} = \mathbf{c}$, with \mathbf{E} the $J \times n$ “edit matrix” with elements e_{ji} and \mathbf{c} the J -vector with elements c_j .

For the example record in table 1 of the module “Micro-Fusion – Reconciling Conflicting Microdata” we have

$\mathbf{x} = (1.Profit, 2.Employees, 3.Turnover\ main, 4.Turnover\ other, 5.Turnover, 6.Wages, 7.Other\ costs, 8.Total\ costs)$.

The three equality edits:

$$e_1: x_1 - x_5 + x_8 = 0 \text{ (Profit = Turnover - Total Costs)}$$

$$e_2: -x_3 + x_5 - x_4 = 0 \text{ (Turnover = Turnover main + Turnover other)}$$

$$e_3: -x_6 - x_7 + x_8 = 0 \text{ (Total Costs = Wages + Other costs)}$$

can be expressed in the form $\mathbf{E}\mathbf{x} = \mathbf{c}$ with

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{pmatrix} \text{ and } \mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Notice that the second column of \mathbf{E} contains all zeroes because the second variable is not involved in any of the edit rules.

In this example a composite record was considered where three variables, *Turnover*, *Employees* and *Total costs* were obtained from reliable administrative sources and the other variables from a survey. As a consequence of obtaining the data from different sources, the edit rules are violated. The adjustment problem was to adjust the survey values such that the edit rules are satisfied while leaving the administrative values unchanged. For the optimisation approach it is necessary to take the distinction between free variables that are allowed to be adjusted and fixed variables that are not, into account. The complete data vector can be partitioned into \mathbf{x}_{fre} for the free variables and \mathbf{x}_{fix} for the fixed ones. A corresponding partitioning of the edit matrix yields, say, \mathbf{E}_{fre} and \mathbf{E}_{fix} . Now we can write

$$\mathbf{E}\mathbf{x} = \mathbf{E}_{fre}\mathbf{x}_{fre} + \mathbf{E}_{fix}\mathbf{x}_{fix} = \mathbf{c},$$

$$\text{and so } \mathbf{E}_{fre}\mathbf{x}_{fre} = \mathbf{c} - \mathbf{E}_{fix}\mathbf{x}_{fix},$$

which can be expressed as

$$\mathbf{A}\mathbf{x}_{fre} = \mathbf{b}, \text{ say.}$$

The r.h.s. of this last expression contains all constants including the values of fixed variables and the l.h.s. contains the free variables that may be changed. They are the actual variables for the optimisation problem. For ease of notation we will, in the context of the optimisation problem, simply write \mathbf{x} for the relevant, not fixed, variables and suppress the suffix *fre*. Thus we will write $\mathbf{A}\mathbf{x} = \mathbf{b}$ for the constraints on the relevant variables.

In addition to the equality constraints we also often have linear inequality constraints. The simplest case is the non-negativity of most economic variables. The optimisation approach can also handle linear inequality constraints. The constraints can then be formulated as $\mathbf{A}_{eq}\mathbf{x} = \mathbf{b}_{eq}$ and $\mathbf{A}_{ineq}\mathbf{x} \geq \mathbf{b}_{ineq}$, where \mathbf{A}_{eq} contains the rows of \mathbf{A} corresponding to the equality constraints and \mathbf{A}_{ineq} the ones corresponding to the inequality constraints. For ease of exposition we shall, without noting otherwise, write these equality/inequality constraints more compactly as $\mathbf{A}\mathbf{x} \geq \mathbf{b}$

With the notation and conventions introduced above we can write the optimisation approach to the problem of finding the smallest possible adjustments compactly as

$$\begin{aligned} \tilde{\mathbf{x}} &= \arg \min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0), \\ \text{s.t. } \mathbf{A}\tilde{\mathbf{x}} &\geq \mathbf{b} \end{aligned} \tag{3}$$

with \mathbf{x}_0 the adjustable part of the record *before* adjustment and $\tilde{\mathbf{x}}$ the corresponding sub-record *after* the adjustment and $D(\mathbf{x}, \mathbf{x}_0)$ a function measuring the distance or deviance between \mathbf{x} and \mathbf{x}_0 . In the next section we will consider different functions D for the adjustment problem.

The conditions for a solution of the minimisation problem formulated in (3) can be found by inspection of the Lagrangian for this problem, which can be written as

$$L(\mathbf{x}, \boldsymbol{\alpha}) = D(\mathbf{x}, \mathbf{x}_0) + \boldsymbol{\alpha}^T (\mathbf{A}\mathbf{x} - \mathbf{b}), \quad (4)$$

with $\boldsymbol{\alpha}$ a vector of Lagrange multipliers, one for each of the constraints j .

From optimisation theory it is well known that for a convex function $D(\mathbf{x}, \mathbf{x}_0)$ and linear (in)equality constraints, the solution vector $\tilde{\mathbf{x}}$ must satisfy the so-called Karush-Kuhn-Tucker (KKT) conditions (see, e.g., Luenberger, 1984). One of these conditions is that the gradient of the Lagrangian w.r.t. \mathbf{x} is zero when evaluated at the optimal point, i.e.,

$$L'_{x_i}(\tilde{x}_i, \boldsymbol{\alpha}) = D'_{x_i}(\tilde{x}_i, \mathbf{x}_0) + \sum_j \alpha_j a_{ji} = 0, \quad (5)$$

with L'_{x_i} the gradient of L w.r.t. x_i and D'_{x_i} the gradient of D w.r.t. x_i . From this condition alone, we can already see how different choices for D lead to different solutions to the adjustment problem. Below we shall consider three familiar choices for D , Least Squares, Weighted Least Squares and Kullback-Leibler divergence, and show how these different choices result in different structures of the adjustments, which we will refer to as the adjustment models. The form of these adjustment models gives some guidance to the choice of metric and the following properties may also be helpful in this respect. Weights in the WLS-criterion can be used to adjust some variables more than others, for instance because they are considered less reliable. Weights can also be used to make the amount of adjustment dependent on the size of the original value. Without knowledge about the preferred relative size of the adjustments for the different variables, the ordinary LS special case arises. The KL-criterion is only defined for positive variables: the original values need to be positive and the adjusted values are also guaranteed to be positive. The KL-adjustments can be expressed as positive multiplicative factors, larger original values will be adjusted more than smaller ones. More details of these adjustment models and their interpretation is given below.

2.2 Least squares adjustments

First, we consider the least squares criterion to find an adjusted \mathbf{x} -vector that is closest to the original unadjusted data, that is: $D(\mathbf{x}, \mathbf{x}_0) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0)$, and so $D'_{x_i}(\tilde{x}_i, \mathbf{x}_0) = \tilde{x}_i - x_{0,i}$, and we obtain from (5)

$$\tilde{x}_i = x_{0,i} + \sum_j a_{ji} \alpha_j. \quad (6)$$

This shows that the least squares criterion results in an additive structure for the adjustments: the total adjustment to variable $x_{0,i}$ decomposes as a sum of adjustments to each of the constraints j . Each of these adjustments consists of an adjustment parameter α_j that describes the amount of adjustment due

to constraint j and the entry a_{ji} of the constraint matrix \mathbf{A} pertaining to variable i and constraint j . Values of 1, -1 or 0 for a_{ji} imply that $x_{0,i}$ is adjusted by α_j , $-\alpha_j$ or not at all.

For variables that are part of the same constraints and have the same value a_{ji} , the adjustments are equal and the differences between adjusted variables are the same as in the unadjusted data. In particular, this is the case for variables that add up to a fixed total, given by a register value, and are not part of other constraints.

2.3 Weighted least squares adjustments

For the weighted least squares criterion, $D(\mathbf{x}, \mathbf{x}_0) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \text{Diag}(\mathbf{w})(\mathbf{x} - \mathbf{x}_0)$, with $\text{Diag}(\mathbf{w})$ a diagonal matrix with a vector with weights along the diagonal. The derivative of this loss function in the optimum is $w_i(\tilde{x}_i - x_{0,i})$ and we obtain from (5)

$$\tilde{x}_i = x_{0,i} + \frac{1}{w_i} \sum_j a_{ji} \alpha_j. \quad (7)$$

Contrary to the least squares case where the amount of adjustment to a constraint is equal in absolute value (if it is not zero) for all variables in that constraint, the amount of adjustment now varies between variables according to the weights: variables with large weights are adjusted less than variables with small weights.

For variables that are part of the same constraints and have the same value a_{ji} , the adjustments are equal up to a factor $1/w_i$ and the differences of the weighted adjusted variables are the same as in the unadjusted data, that is, for variables i and i' we have $w_i \tilde{x}_i - w_{i'} \tilde{x}_{i'} = w_i x_{0,i} - w_{i'} x_{0,i'}$.

The weighted least squares approach to the adjustment problem has been applied by Thomson et al. (2005) in the context of adjusting records with inconsistencies caused by imputation. Some of the variables were missing and the missings were filled in by imputed values without taking care of edit constraints. This caused inconsistencies that were resolved by minimal adjustments, in principle to all variables, observed or imputed, according to the WLS-criterion. They used weights of 10,000 for observed values and weights of 1 for imputed values. Effectively, this means that if a consistent solution can be obtained by changing only imputed variables, this solution will be found. Otherwise (some of the) observed variables will also be adjusted.

One specific form of weights that is worth mentioning is obtained by setting the weight w_i equal to $1/x_{0,i}$ resulting, after dividing by $x_{0,i}$ in the adjustment model

$$\frac{\tilde{x}_i}{x_{0,i}} = 1 + \sum_j a_{ji} \alpha_j, \quad (8)$$

which is an additive model for the *ratio* between the adjusted and original values. It may be noticed that the expression on the right-hand side of (8) is the first-order Taylor expansion (i.e., around 0 for all the α_j 's) to a multiplicative adjustment given by

$$\frac{\tilde{x}_i}{x_{0,i}} = \prod_j (1 + a_{ji} \alpha_j) \quad (9)$$

From (8) we see that the α_j 's determine the difference from 1 of the *ratio* between the adjusted and original values, which is usually much smaller than unity in absolute value (e.g., an effect of 0.2 implies a 20% increase due to adjustment which is large in practice). The products of the α_j 's are therefore often much smaller than the α_j 's themselves, in which cases (9) becomes a good approximation to (8), i.e., the corresponding WLS adjustment is roughly given as the product of the constraint-specific multiplicative adjustments.

2.4 Kullback-Leibler adjustments

The Kullback-Leibler divergence measures the difference between \mathbf{x} and \mathbf{x}_0 by the function $D_{KL} = \sum_i x_i (\ln x_i - \ln x_{0,i} - 1)$. The derivative of this loss function is $\ln \tilde{x}_i - \ln x_{0,i}$ and we obtain from (5)

$$\tilde{x}_i = x_i \times \prod_j \exp(-a_{ji} \alpha_j). \quad (10)$$

In this case the adjustments have a multiplicative form and the adjustment for each variable is the product of adjustments to each of the constraints. The adjustment factor $\gamma_j = \exp(-a_{ji} \alpha_j)$ in this product represents the adjustment to constraint j and equals 1 if a_{ji} is 0 (no adjustment), $1/\gamma_j$ if a_{ji} is 1 and γ_k , if a_{ji} is -1.

For variables that are part of the same constraints and have the same value a_{ji} , the adjustments factors are equal and the ratios between adjusted variables are the same as between the unadjusted variables, $\tilde{x}_i / \tilde{x}_j = x_{0,i} / x_{0,j}$.

2.5 Generalisations: Adjusting to multiple sources and soft constraints

In this section we consider the possibilities for further modelling of the adjustment problem by using, simultaneously, information from multiple sources. First, we consider the situation that both register and survey values are considered to provide information for the final adjusted record rather than discarding survey values for which register values are available. Then we show that the approach used to combine information from multiple sources can be viewed as using, in addition to the “hard” constraints that are to be satisfied exactly, also “soft” constraints that only need to be fulfilled approximately.

2.5.1 Adjusting to both survey and register values

So far we considered the case where one of the sources (the administrative one) provides the reference values that are considered to be the correct ones and these values replace the values of the corresponding survey variables. Another situation arises when both data sources are considered to be fallible. In this situation we do not want to discard the data from one of the sources but we consider both sources to provide useful information on the variables of interest. This means that in the final consistent estimated vector we should not simply copy the values from the register values but obtain adjusted values that depend on both the survey values and the available register values. The data from the survey will be denoted by $\mathbf{x}_{0,S}$ and the data from the register by $\mathbf{x}_{0,R}$. In particular, for the example in table 1 of the module “Micro-Fusion – Reconciling Conflicting Microdata” we have the following:

$\mathbf{x}_{0,S}=(Profit, Employees, Turnover\ main, Turnover\ other, Turnover, Wages, Other\ costs, Total\ costs),$
 $\mathbf{x}_{0,R}=(Employees_reg, Turnover_reg, Total\ costs_reg).$

where the suffix *_reg* is used to distinguish the register variables from their survey counterparts.

A consistent minimal adjustment procedure based on the information from both the survey values, the register values and the edit rules can be set up by considering the following constrained optimisation problem

$$\begin{aligned} \tilde{\mathbf{x}} &= \arg \min_{\mathbf{x}} \{D(\mathbf{x}, \mathbf{x}_{0,S}) + D(\mathbf{x}_R, \mathbf{x}_{0,R})\} \\ \text{s.t. } \mathbf{Ax} &\geq \mathbf{0} \end{aligned} \quad (11)$$

where the vector \mathbf{x}_R denotes the subvector of \mathbf{x} that contains the variables that are observed in the register. The vectors \mathbf{x} and $\mathbf{x}_{0,S}$ both contain all variables and can be partitioned as $\mathbf{x} = (\mathbf{x}_{\bar{R}}^T, \mathbf{x}_R^T)^T$ and $\mathbf{x}_{0,S} = (\mathbf{x}_{0,S\bar{R}}^T, \mathbf{x}_{0,SR}^T)^T$, with \bar{R} denoting the set of variables not in the register. Using this partitioning and the property that the distance functions considered in this paper are all decomposable in the sense that they can be written as a sum over variables, (11) can be re-expressed as

$$\begin{aligned} \tilde{\mathbf{x}} &= \arg \min_{\mathbf{x}} \{D(\mathbf{x}_{\bar{R}}, \mathbf{x}_{0,S\bar{R}}) + D(\mathbf{x}_R, \mathbf{x}_{0,SR}) + D(\mathbf{x}_R, \mathbf{x}_{0,R})\} \\ \text{s.t. } \mathbf{Ax} &\geq \mathbf{0} \end{aligned} \quad (12)$$

This clearly shows that the values of the variables *R* that are in both the register and the survey are adjusted to satisfy the edit constraints and remain as close as possible to both the register value and the survey value. Note that variables that are in both the register and the survey will be adjusted, if the two values are not equal, even if they do not appear in any edit rules, which is different from the situation considered before.

2.5.2 Soft constraints

The adjustment towards the register values due to a separate component in the objective function can also be interpreted as adding “soft” constraints to the optimisation problem. These soft constraints express that $\tilde{\mathbf{x}}_R$ should be approximately equal to the register values $\mathbf{x}_{0,R}$ but need not “fit” these data exactly as was required before.

The notion of soft constraints opens up a number of possibilities for further modelling the adjustment problem. Suppose, for instance, that the total amount of wages paid (*Wages*) is known from an administrative source and treated as fixed while the number of employees (*Employees*) is a free variable. Furthermore, assume that before adjustment the wages are 20,000 Euros per employee and that it is plausible that this ratio should hold approximately for the record after adjustment. This can be formulated as a “soft” ratio constraint on *Employment* and *Wages*: $Wages / Employment \approx 20,000$. This soft constraint can be handled by the optimisation problem by adding to the loss function the component $D(x_{wages}, 20000 \times x_{employment})$. This soft constraint is often more reasonable than using hard upper and lower bounds on the adjusted value for *Employment*. In fact we can do both, for instance to bound *Employment* within certain hard limits and use the soft constraint to draw the value of *Wages* within these bound towards the expected value of 20,000 times the number of employees.

3. Preparatory phase

4. Examples – not tool specific

4.1 Comparison of distance functions using the example record

The different methods (LS, WLS and KL) have been applied to make the two composite records consistent that are in the example of table 1 in the module “Micro-Fusion – Reconciling Conflicting Microdata”. For the WLS method we used as weights the inverse of the \mathbf{x}_0 -values so that the relative differences between \mathbf{x} and \mathbf{x}_0 are minimised and the adjustments are proportional to the size of the \mathbf{x}_0 -values.

The optimisation methods were implemented by an iterative method which is a special case of the so-called row-action algorithms treated in Censor and Zenios (1997) (see also, De Waal et al., 2011, Ch. 10). For the (weighted) least squares adjustments an R-package is available (van der Loo, 2012).

The results for the different methods are in table 1 below. The solutions for the KL- and WLS-adjustments appeared to be the same in all digits shown and were therefore combined into a single column. With the weights used here these solutions should be similar in practice. The register values that are treated as fixed are shown in bold; the other values may be changed by the adjustment procedure.

Table 1. Example business record: two composite versions and adjusted values.

Variable	Name	Composite record II			Composite record I		
		Unadj.	LS	WLS/KL	Unadj.	LS	WLS/KL
x_1	Profit	330	282	291	330	260	249
x_2	Employees	20	20	20	25	25	25
x_3	Turnover main	1000	960	922	1000	960	922
x_4	Turnover other	30	-10	28	30	-10	28
x_5	Turnover	950	950	950	950	950	950
x_6	Wages	500	484	470	550	550	550
x_7	Other costs	200	184	188	200	140	151
x_8	Total costs	700	668	658	700	690	701

Unadj. = Unadjusted values.

LS = adjusted values according to the LS criterion.

WLS/KL = adjusted values according to the WLS or KL criterion.

For both composite records, the LS adjustment procedure leads to one negative value for *Turnover other*, which is not allowed for this variable. Therefore the LS-procedure was run again with a non-negativity constraint added for the variable *Turnover other*. This results simply in a zero for that variable and a change in *Turnover main* to ensure that $Turnover = Turnover\ main + Turnover\ other$. Without the non-negativity constraint, the LS-results clearly show that for variables that are part of the same constraints (in this case the pairs of variables x_3, x_4 and x_6, x_7 that are both appearing in one constraint only), the adjustments are equal: -40 for x_3, x_4 and -16 for x_6, x_7 . *Total costs* (x_8) is part of two constraints and therefore the total adjustment to this variable consists of two additive components. One component to adjust to the constraint $e_1: x_1 - x_5 + x_8 = 0$ ($Profit = Turnover - Total\ Costs$) and one component to adjust to $e_3: x_8 - x_6 - x_7 = 0$ ($Total\ Costs = Wages + Other\ costs$). For the composite

record II, the first component is minus 48 – which is also the single adjustment component for *Profit* – and the second component is 16 – which is also the single adjustment component for *Wages* and *Other costs* (with opposite sign). These two components add up to the adjustment of –32.

The results for the WLS/KL solution show that for this weighting scheme the adjustments are larger, in absolute value, for large values of the survey variables than for smaller ones. In particular, the adjustment to *Turnover other* is only –2.3 – so that no negative adjusted value results in this case – whereas the adjustment to *Turnover main* is 77.7. The multiplicative nature of these adjustments (as KL-type adjustments) also clearly shows since the adjustment *factor* for both these variables is 0.92 (for both composite records). The adjustment factor for *Wages* and *Other costs* in composite record I is also equal (to 0.94) because these variables are in the same single constraint and so the ratio between these variables is unaffected by this adjustment. However the ratio of each of these variables to *Total Costs* is not unaffected because *Total Costs* has a different sign in the constraint e_3 and, moreover, *Total Costs* is also part of constraint e_1 so that it is subject to two adjustment factors.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Censor, Y. and Zenios, S. A. (1997), *Parallel Optimization. Theory, Algorithms, and Applications*. Oxford University Press, New York.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons Inc., Hoboken, New Jersey.
- Luenberger, D. G. (1984), *Linear and Nonlinear programming, second edition*. Addison-Wesley, Reading.
- Pannekoek, J. (2011), Models and algorithms for micro-integration. In: *Report on WP2: Methodological developments*, ESSNET on Data Integration, available at <http://www.cros-portal.eu/content/wp2-development-methods>.
- Pannekoek, J. and Zhang, L.-C. (2011), Partial (donor) imputation with adjustments. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing.
- van der Loo, M. (2012), *rspa: Adapt numerical records to fit (in)equality restrictions with the Successive Projection Algorithm*. R package version 0.1-1.
- Thomson, K., Fagan, J. T., Yarbrough, B. L., and Hambric, D. L. (2005), Using a Quadratic Programming Approach to Solve Simultaneous Ratio and Balance Edit Problems. Working paper 32, UN/ECE Work Session on Statistical Data Editing, Ottawa.

Specific section

8. Purpose of the method

The purpose of the method is to adjust the values of some variables in a data record to remove edit violations to ensure consistency of the data values obtained from different sources.

9. Recommended use of the method

1. The method should be used after detection and treatment of errors and missing values.

10. Possible disadvantages of the method

1. When inconsistencies arise due to large errors in some values, these errors may propagate to other values due to adjustment. Influential errors should therefore be treated before the method is applied.

11. Variants of the method

1. Least squares adjustments
2. Weighted least squares adjustments.
3. Kullback-Leibler adjustments.

12. Input data

1. Data records with possibly inconsistent values and edit rules.

13. Logical preconditions

1. Missing values
 1. Missing values are allowed but edit rules involving variables with missing values cannot be checked and no adjustment with respect to these edit rules will take place.
2. Erroneous values
 1. Influential erroneous values should be treated before the method is applied.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. The amount of change applied to individual variables can be controlled by specifying weights for the variables

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. The output consists of the same individual records as the input, with values adapted when needed to ensure consistency with the edit rules.

17. Properties of the output data

1. The output data are ensured to be consistent with all specified edit rules that do not involve variables with missing values.

18. Unit of input data suitable for the method

The input consists of individual records that are treated one-by-one, independently.

19. User interaction - not tool specific

- 1.

20. Logging indicators

- 1.

21. Quality indicators of the output data

- 1.

22. Actual use of the method

- 1.

Interconnections with other modules**23. Themes that refer explicitly to this module**

1. Micro-Fusion – Data Fusion at Micro Level
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Automatic Editing
4. Statistical Data Editing – Editing Administrative Data
5. Imputation – Main Module

24. Related methods described in other modules

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Micro-Fusion – Prorating
3. Micro-Fusion – Generalised Ratio Adjustments

25. Mathematical techniques used by the method described in this module

1. Optimisation of convex functions with linear (in)equality constraints.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

1. The R-package rspa of van der Loo (2012) can be used to apply adjustment according to the (weighted) least squares criterion.

28. Process step performed by the method

GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

29. Module code

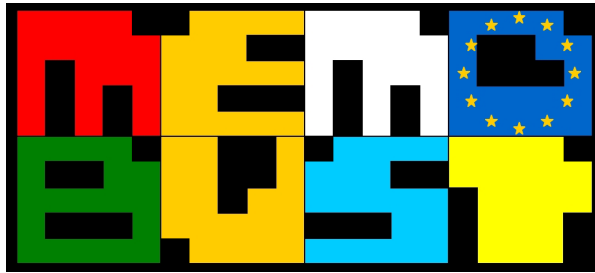
Micro-Fusion-M-Minimum Adjustment Methods

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-03-2013	first version	Jeroen Pannekoek	CBS (Netherlands)
0.2	17-04-2013	second version	Jeroen Pannekoek	CBS (Netherlands)
0.2.1	09-09-2013	preliminary release		
0.3	20-12-2013	improvements based on the EB-review	Jeroen Pannekoek	CBS (Netherlands)
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:01



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Generalised Ratio Adjustments

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	5
4. Examples – not tool specific.....	5
4.1 Generalised ratio adjustment compared with WLS/KL-adjustments	5
5. Examples – tool specific.....	6
6. Glossary.....	6
7. References	6
Specific section.....	7
Interconnections with other modules.....	8
Administrative section.....	10

General section

1. Summary

Generalised ratio adjustment is a method to reconcile conflicting information as described in the module “Micro-Fusion – Reconciling Conflicting Microdata”. The method uses multiplicative adjustments, just as the methods prorating (see the module “Micro-Fusion – Prorating”) and the KL-adjustments (see the module “Micro-Fusion – Minimum Adjustment Methods”). The generalised ratio adjustments method aims to make the adjustments as uniform as possible. Furthermore, and in contrast with the other adjustment methods, the method can result in adjustments to variables that are not involved in any of the constraints.

2. General description of the method

The generalised ratio adjustments are multiplicative adjustments applied to variables that are “free” variables which means that they are designated to be adjustable. These variables may or may not be involved in edit constraints. The adjustments methods considered in the modules “Micro-Fusion – Prorating” and “Micro-Fusion – Minimum Adjustment Methods” were only meant to resolve violations of edit rules, therefore only free variables involved in edit rules were adjusted since variables not appearing in the edit rules are irrelevant because they cannot violate edit rules.

However, there may be reasons other than the violation of edit rules to change the values of some variables. Consider, for instance, the business record shown in section 4 below (and in table 1 of the module “Micro-Fusion – Minimum Adjustment Methods”). In the column denoted by *Survey*, values for the variables are shown that are obtained from a survey. Two scenarios are assumed for additional data: (I) from administrative sources, values are available for the variables *Employees*, *Turnover* and *Wages* (the values in bold in the columns *Adjusted Composite (I)*) and (II) an administrative source is only available for the variable *Turnover*. Suppose that the administrative data are treated as fixed, for instance because they are more recent (although less detailed) and / or more accurate than the survey data. Adjusting the values of the survey variables *Turnover main* and *Turnover other* can then be seen as extrapolating the (slightly) outdated survey values to the more recent administrative data. Apparently, according to the available data, this unit’s turnover has been reduced (from 1030 to 950) and multiplicative adjustments for *Turnover main* and *Turnover other* are easily obtained by reducing them by the same ratio of 1030/950. In this case one may be tempted to apply this rescaling to all variables, also those not involved in constraints, which can be justifiable if it is assumed that these variables are related to *Turnover* in approximately the same way as in the original survey record; in some sense the “size” of the business has decreased by a factor 1030/950 and all variables are scaled with this factor to reflect this change. In the newly created consistent record their ratio to *Turnover* would be preserved by this rescaling. This intuitive and simple solution becomes difficult if more variables are obtained from administrative sources leading to multiple adjustment factors. It is then not obvious how to take these different factors into account and to ensure that constraints are satisfied.

One possible solution is to use the minimum adjustment approach described in the module “Micro-Fusion – Minimum Adjustment Methods” and to add the ratios of the variables not involved in the constraints to each of the administrative variables as “soft” constraints to the optimisation problem (see section 2.5.2 of that module). However, this approach leads to a non-trivial modelling effort and a more complicated loss function. As a method that can be applied more routinely, using only the

already specified edit constraints, Pannekoek and Zhang (2011) suggested a generalised ratio adjustments method. As in the modules “Micro-Fusion – Reconciling Conflicting Microdata” and “Micro-Fusion – Minimum Adjustment Methods” a composite record is considered, consisting of values obtained from different sources, that may violate some linear edit constraints. The task is to make adjustments to a subset of the variables in the composite record such that the resulting record becomes consistent with the edit rules. The variables that are allowed to be adjusted are named *free* variables and the other variables are the *fixed* variables. For instance in scenario (I) the unadjusted composite record consists of values from the administrative source for variables x_2 , x_5 and x_6 and these variables are treated as fixed. The remaining variables in the composite record have values from the survey; these variables are treated as free and will be adjusted to meet the edit constraints. The generalised ratio adjustments method finds multiplicative adjustments such that the resulting adjusted values meet the following two requirements: (1) the edit-constraints are satisfied and (2) the changes with respect to the original survey record are as uniform as possible (resembling a uniform overall ratio adjustment as much as possible).

The generalised ratio adjustments method focusses on the changes between the values in the *original survey record* and the final *adjusted composite record*. These changes can be expressed as factors δ_i , defined by

$$\delta_i = \tilde{x}_i / x_{s,i}, \text{ for } i=1, \dots, n, \quad (1)$$

with n the number of variables, \tilde{x}_i the values of the variables in the adjusted composite record and $x_{s,i}$ the survey values for these variables. By definition, the values of the fixed variables in the composite records are the same before and after adjustment. For these variables, the change factors δ_i represent the change between the survey value and the administrative value. For free variables the changes δ_i are adjustment factors that adjust the survey values such that the edit constraints are satisfied.

Before adjustment, the composite record consists of values $x_{0,i}$ which are equal to the administrative values if these are available and equal to the survey values otherwise. The record x_0 differs from the original survey record in the administrative values only. Since the administrative values are treated as fixed, these values will not be changed by the adjustment procedure and thus the change factors for the fixed variables can be expressed as

$$\delta_i = \tilde{x}_i / x_{s,i} = x_{0,i} / x_{s,i}, \text{ for } i \in I_{fix}, \quad (2)$$

with I_{fix} the set of indices corresponding to the fixed (administrative) variables. For the fixed variables the δ_i are given by (2) but for the other variables, the free variables with index set I_{free} , the δ_i need to be determined such that the edit rules are satisfied and all change factors (including those for the fixed variables) are as uniform as possible. Specifically, the δ_i will be obtained by minimising the following objective function (Δ) over the δ_i corresponding to the free variables:

$$\min_{\delta_i | i \in I_{free}} \Delta = \sum_{i \in I_{free}} (\delta_i - \bar{\delta})^2, \text{ where } \bar{\delta} = \frac{1}{n} \sum_{i=1}^n \delta_i, \quad (3)$$

with constraints as in the module “Micro-Fusion – Minimum Adjustment Methods”, i.e., $\mathbf{A}\tilde{\mathbf{x}}_{free} = \mathbf{b}$ with $\tilde{\mathbf{x}}_{free}$ the vector with adjusted free variables. Notice that the minimum is taken over the free variables only but the mean is taken over all variables, both free and fixed. The objective function (3) can be viewed as a function of the change factors δ_i for the free variables but also as a function of the adjusted values \tilde{x}_i (since $\delta_i = \tilde{x}_i / x_{s,i}$) for these variables. In either case the variation in the change factors is minimised subject to the linear edit constraints on the adjusted values. A possible generalisation of (3) that differentiates between the effects of the different δ_i on the objective value is to use weights similar to the WLS loss-function in “Micro-Fusion – Minimum Adjustment Methods”.

The fact that the objective function makes the changes (with respect to the original survey record) for all variables in the record as uniform as possible results in two properties of the generalised ratio adjustments not shared by the minimum adjustment methods. Firstly, adjustments are defined for all free variables, whether they are involved in edit constraints or not. This is because minimising the variation in the δ_i will, in general, lead to values for δ_i unequal to 1 (and hence to adjustment) even for survey values not involved in edit constraints. Secondly, the information from the changes between the survey values and administrative values of the fixed variables is used in the adjustment procedure. This is because the mean of all changes, $\bar{\delta}$, is partly determined by these changes in the fixed variables and therefore these changes influence the adjustment factors for the free variables since they are made to vary as little as possible around $\bar{\delta}$.

3. Preparatory phase

4. Examples – not tool specific

4.1 Generalised ratio adjustment compared with WLS/KL-adjustments

In this example we show the results of the generalised ratio method and compare these results with the WLS/KL-adjustments described in the module “Micro-Fusion – Minimum Adjustment Methods”. Both methods use multiplicative adjustments but the WLS/KL-adjustments apply only to variables that are involved in constraints whereas the generalised ratio method can also adjust variables that are not involved in constraints and, in addition, this last method will result in adjustments that are as uniform as possible. Both methods will result in a record that satisfies all linear constraints.

The data for this example are the values of a business record shown in table 1 of module “Micro-Fusion – Reconciling Conflicting Microdata” and repeated in Table 1 below. Two versions of an adjusted composite record are shown¹, one for a record with three values obtained from an administrative source (which are shown in bold) that is denoted by *Adjusted Composite (I)* and another with only *Turnover* obtained from an administrative source, denoted by *Adjusted Composite*

¹ Values are rounded to the nearest integer.

(II). The other values are from a survey, see the column *Survey*. The administrative values are treated as fixed while the survey values are free, i.e., they can be adjusted.

The composite record (II) with only *Turnover* from the administrative source violates two edit rules:

$$e_1: x_1 - x_5 + x_8 = 0 \text{ (Profit = Turnover - Total Costs);}$$

$$e_2: -x_3 + x_5 - x_4 = 0 \text{ (Turnover = Turnover main + Turnover other);}$$

The survey value of *Turnover* is 1030 and, as expected, the generalised ratio adjustments for this record reduce to a global proportional adjustment of all the survey values by a ratio of 0.922 (=950/1030) including the variable *Employee*. That this last variable is adjusted is a difference with the minimum-adjustment approaches that only adjust variables that are involved in constraints.

Table 1. Example business record with survey values and adjusted values for the WLS/KL and generalised ratio methods.

Variable	Name	Survey	Adjusted Composite (I)		Adjusted Composite (II)	
			WLS/KL	Gen. Ratio	WLS/KL	Gen. Ratio
x_1	Profit	330	249	239	291	304
x_2	Employees	20	25	25	20	18
x_3	Turnover main	1000	922	921	922	922
x_4	Turnover other	30	28	29	28	28
x_5	Turnover	1030	950	950	950	950
x_6	Wages	500	550	550	470	461
x_7	Other costs	200	151	161	188	184
x_8	Total costs	700	701	711	658	646

For composite record (I) with *Turnover*, *Wages* and *Employees* obtained from administrative sources, three edit rules are violated: in addition to e_1 and e_2 also the rule

$$e_3: -x_6 - x_7 + x_8 = 0 \text{ (Total Costs = Wages + Other costs).}$$

is violated. Also in this case, the generalised ratio adjustments are close to the WLS/KL solution. The empirical variance of the multiplicative factors (i.e., proportional to the value of the loss function Δ) is 0.0270 for the generalised ratio adjustments, which is a little bit less than the value 0.0276 obtained for the WLS/KL solution.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Pannekoek, J. and Zhang, L.-C. (2011), Partial (donor) imputation with adjustments. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing.

Specific section

8. Purpose of the method

The purpose of the method is to adjust the values of some variables in a data record to remove edit violations to ensure consistency of the data values obtained from different sources. The generalised ratio adjustments method aims to make the adjustments as uniform as possible. Furthermore, and in contrary to the other adjustment methods, the method can result in adjustments to variables that are not involved in the constraints.

9. Recommended use of the method

- 1.

10. Possible disadvantages of the method

- 1.

11. Variants of the method

- 1.

12. Input data

- 1.

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

- 1.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

- 1.

17. Properties of the output data

- 1.

18. Unit of input data suitable for the method

19. User interaction - not tool specific

- 1.

20. Logging indicators

- 1.

21. Quality indicators of the output data

- 1.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Data Fusion at Micro Level
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Editing Administrative Data
4. Imputation – Main Module

24. Related methods described in other modules

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Micro-Fusion – Prorating
3. Micro-Fusion – Minimum Adjustment Methods

25. Mathematical techniques used by the method described in this module

1. Quadratic optimisation

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

1. There are no specific tools available that implement this method. However, the method can be applied using quadratic programming routines.

28. Process step performed by the method

GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

29. Module code

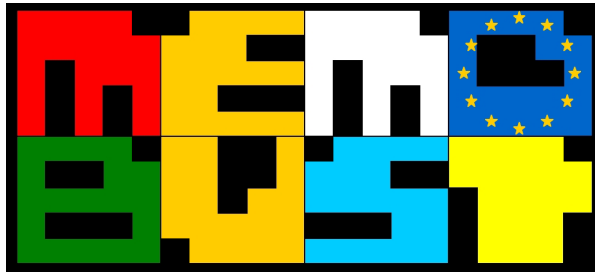
Micro-Fusion-M-Generalised Ratio Adjustments

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-03-2013	first version	Jeroen Pannekoek	CBS (Netherlands)
0.2	17-04-2013	second version	Jeroen Pannekoek	CBS (Netherlands)
0.3	09-07-2013	third version	Jeroen Pannekoek	CBS (Netherlands)
0.3.1	09-09-2013	preliminary release		
0.4	20-12-2013	improvements based on the EB-review	Jeroen Pannekoek	CBS (Netherlands)
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:01



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Coding – Main Module

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction	3
2.2 Elaboration	4
2.3 Comments about classifications and coding.....	7
2.4 Misconceptions about coding	9
3. Design issues	10
4. Available software tools.....	11
5. Decision tree of methods	11
6. Glossary.....	12
7. References	12
Interconnections with other modules.....	13
Administrative section.....	14

General section

1. Summary

The main theme of this document concerns the methods for automatic or semi-automatic (interactive) coding of answers to open questions. These are short descriptions (typically less than 10 words) in a respondent's own words formulated about the person's occupation, education followed, work performed, goods and services produced, etc. The code that is assigned to a description (if successful) originates from a classification. The classification itself is too complicated for respondents to directly search it for an answer. It is easier to let the respondent answer in his or her own words, and then to try to interpret this answer. Nowadays, this interpretation usually employs a computer if the material is delivered electronically, as we will assume. In the past, this coding was done completely 'manually'. That manual coding process is expensive, slow and non-transparent. Nowadays, the goal is to have the bulk of the coding work done by computer running special coding software. The remaining 'difficult cases' are then resolved more or less 'manually' as in the past.

2. General description

2.1 Introduction

Coding is an activity in the statistical process. It can be considered as a special type of derivation, and a rather difficult one. The purpose of coding is to match a code derived from a classification to textual information. The goal in this process is to reduce the large variety of answers to a convenient number, and to organise these answers (the classification used offers this option by means of its structure).

We view this matching as an interpretation of a description (the textual information) in the light of the classification concerned. An example is a description of an economic activity (in a respondent's own words) that is interpreted taking into account the NACE. Other examples concern descriptions of goods, descriptions of education that people have, illnesses suffered by people, and causes of death.

Coding is also very similar to a doctor's diagnosis of patients who present him or her with various complaints and symptoms. The task of a doctor is to diagnose an illness or abnormality based on a number of observations, answers from the patient and possibly additional tests (blood tests, for instance).

The main reason why variables with open answers are used is that this is convenient for the respondent. Also, there is far less influence on the answer. For such variables, the person can answer with a personally formulated text. If the respondent were required to give an answer in the form ultimately needed by NSIs to create statistics, then he would have to know the classification that serves as the basis for such a variable, such as the Standard Industrial Classification (SIC). However, this is much too difficult, and from a practical point of view, impossible to expect from a non-specialist.

In the past, coding these open-text answers invariably was done by human coders, specialists in coding of occupations, education, business activities, etc. The problem with this 'manual coding' is that it is time-consuming, expensive and not standardised. Consequently, over time, computers have been used increasingly to assist in the coding. This ranges from computer-supported applications, where the computer is used to provide search facilities in a file with codes and their descriptions, to a fully

automatic data processing application ('automatic coding'). To date, however, automatically coding all answers correctly has not been feasible, and the question is whether it ever will be. But it is not a requirement that coding should be fully automated. Partially processing such information can already result in substantial efficiency gains. Another benefit is that automatic coding is bound to increase, for cases that are not too difficult, consistency of the answers (codes), without a loss of quality and possibly even with an improvement of quality; for computer-aided coding, an audit trail can render the process well-defined. Obviously, special efforts are required to make the coding software suitable for this purpose.

In automatic coding, there are two big problems that must be dealt with:

1. To interpret the natural language descriptions, and
2. To link these descriptions to the classification that is used.

What is meant in the first point is primarily that the text is alphanumeric, not so much that it could be handwritten if a paper questionnaire is used. In fact, it is preferable that the text is not handwritten, as this is an additional complicating factor. A computer program must choose which code best fits a description. The problem with coding open text is that many complications can arise, such as:

- Spelling problems
- Grammatical problems (relationships between words, syntax)
- Semantic problems (meaning of words, concepts, sentence fragments, a single sentence, several sentences)
- Interpretation problems (which code from the classification best fits a description).

A complication that can arise in conjunction with this last point is that, viewed from the classification perspective, a description may be incomplete, or that it may relate to two or more different codes. These complications may be due to the fact that a respondent is not likely to be familiar with the classification used, and therefore can provide ambiguous or irrelevant information, or information that lacks detail or is too detailed. Furthermore, it is possible that the classification has been set up purely from a theoretical perspective, without taking into account how to map descriptions to these codes.

In this document, *coding* refers to the activity with the goal of converting descriptions (which are represented as strings of symbols) to a code, originating from a classification. *Coding* often refers to coding by a specialist coder. In this document, this is called *interactive coding*. Coding with the help of a computer program is referred to as *automatic coding* (if the decisions about individual records are not taken by a person) and *computer-supported coding* (if, in a large part of the cases, the computer/an algorithm does not make any coding decisions but only presents suggestions to a human coder, or acts as an electronic reference file or index). *Coder* refers to a person that concentrates on coding according to one or several classifications. This could be a full-time coder at the statistical office, or *pecially trained* interviewer in the field.

2.2 *Elaboration*

Coding an open-text question is a process of interpreting an answer in terms of a predefined set of possible answers. This choice is sometimes made by respondents, during an interview or when filling in a questionnaire, possibly with an interviewer's assistance. However, this choice can also be made

afterwards by coders, at the statistical office, lacking the feedback of the respondent. Because this manual coding is a rather time (and money)-consuming process, automating the process is extremely worthwhile. This is known as automatic coding. In this process, descriptions (answers from respondents in their own words) are the input, and the output is a set of codes related to a certain classification.

When respondents are permitted to give an answer in their own words, this gives them a lot of freedom. In addition, this prevents a situation where the respondents have to know the classification (which often requires specialist knowledge to be understood and used) or where respondents do not agree with the answer selection provided. A disadvantage, however, is that this must be followed by a rather expensive, time-consuming and error-prone coding process in order to code these answers. For that matter, it is highly questionable whether the answers provided always contain the precision and details that are desired or needed in order to code according to a given classification. To sidestep this problem, it is also possible to attempt to use a number of simple closed questions, and then to arrive at a desired code using a derivation scheme. As a result, it is possible to exert influence on the desired type of information and the detail level of the answers (see, for example, Hacking et al., 2006).

Before answers to questions can be used to produce statistical results, coding is indispensable. As a matter of fact a kind of coding is also applied if a closed question is used, but in that case, it is the respondent who does the coding, and has to decide which answer is best among the possible ones. As a rule, coding can be done at different places in the data collection or throughput process steps, as indicated in Table 1.

In practice, combinations of the four options provided in Table 1 are generally always used. The selection of the options is often based on shifting the effort involved and the difficulties of the coding. The ‘most convenient’ approach depends on a large number of preconditions, such as:

- The *domain* or *area of application* of the question (including the ‘hardness’ / ‘softness’ of the question). ‘Gender’ concerns a harder piece of data than ‘opinion about the government’. The first is more stable than the second and, furthermore, generally easier to indicate;
- The expertise of the respondent or the interviewer;
- The structure and complexity of the classification;
- The desired stability of the coding, i.e., how much or how often does the classification change over time?
- The number of respondents (or the net sample size);
- The input medium;
- The form of the source material: separate words, statements, short sentences, paragraphs;
- The desired balance between quality, output level and efficiency of the coding method;
- The desired speed (‘throughput time’) of the processing;
- The available budget;
- The desired detail of the coding results;
- The desire to make the coding process reproducible and transparent.

Table 1. Possible places to code and by whom/what?

Coder?	Where?	Type of survey	Advantages	Disadvantages
Respondent	Field	CAWI	<ul style="list-style-type: none"> direct feedback 	<ul style="list-style-type: none"> no knowledge of the classification
Interviewer	Field, NSI	CAPI (field), or CATI (NSI)	<ul style="list-style-type: none"> direct feedback 	<ul style="list-style-type: none"> superficial knowledge of the classification¹
Coding expert	NSI	PAPI, CAWI, CAPI, CATI	<ul style="list-style-type: none"> expert knowledge of the classification can also use extra information that was included 	<ul style="list-style-type: none"> direct feedback not always possible (sometimes possible for businesses) feedback is very time-consuming coding may be inconsistent not (always) transparent
Automatic coding tool	NSI	PAPI, CAWI, CAPI, CATI	<ul style="list-style-type: none"> fast, consistent coding coding knowledge is specified in a system and is therefore transferable can be made transparent (audit trail) can operate day and night 	<ul style="list-style-type: none"> no direct feedback only the relatively simple cases are coded (but that is often the bulk)

When descriptions are being coded, errors can be made, either by the coders or by the coding program used. Insight into this can be gained through experiments (double blind coding), possibly depending on the detail level of the classification used.

In coding, both interactive and automatic, an optimum must continually be found between maximising the yield (the coding percentage) and maximising the quality (that is, minimising the number of errors). There is also a third maximisation to consider: the smallest possible effort (from the employer's perspective, to control costs, etc.). An important means of preventing incorrect coding is by establishing a *doubt category*. Traditionally, human coders were not permitted to have a doubt category (or only a very small one)², but this is allowed for an automatic coding program. The records that are rejected by such a program because of difficulties encountered, are subsequently presented to human coders for coding. In addition, using an interactive coding module (based on an informative

¹ The amount of knowledge of the classification that an interviewer must have depends very much on the interactive coding tools as used in the CAPI/CATI tool. More knowledge may increase coding accuracy and/or rate, but may also increase costs as the interviewers must be (re)trained.

² A lot of classifications contain a code "other ..." at many places in the classification tree, which allows the human coder to "code" not sufficiently specified answers.

base) during CAPI, CAWI or CATI allows an escape from the conflict between yield and quality: such a module can give feedback during the interview to help and reduce ambiguities or vague answers (for example, see Hacking, 2006).

Experience has shown that nearly every source and every coding contains a large fraction of easy records to code, and a smaller fraction of difficult records to code (this situation is often referred to by the 80% / 20% rule, but these percentages should not be taken too literally). Automatic coding focuses mainly on the easier fraction of records to code, which represents the bulk of the material to be coded.

The automatic classification techniques can generally be divided into two groups:

1. **Language-based:** Here, we really look at the meaning of the words, and make use of language-specific attributes, such as grammar and the relationships between words and concepts (such as synonyms, hyponyms, hyperonyms, etc.)
2. **Statistical:** Here, descriptions are only viewed as a collection of words, which are often described by a sparse vector $Z = \{w_1, \dots, w_n\}$, where n is equal to the number of words occurring in the vocabulary, and w_i the frequency of word i in the description. As a rule, the word order is not included as input for the classification. We could view this approach as classifying a house by first breaking it down and then looking at the stones in the pile of rubble. The assumption used here is known as the *bag-of-words*³ assumption.

These two approaches – language-based and statistical – are extremes. It is very well possible that, in practice, a mixed form will be selected. This could involve, for example, an approach with some ‘light grammatical pre-processing’, followed by automatic coding based on statistical techniques.

2.3 *Comments about classifications and coding*

Here we want to take a moment to examine classifications in this section. A classification provides the codes that should be associated with the descriptions provided by respondents, (if this is possible, which is not guaranteed). Some examples of large hierarchical classifications are:

- NACE – Standard Industrial Classification
- NSTR, PRODCOM – classifications of goods

Mostly, classifications must be considered as given, only to be changed by special committees responsible for their maintenance.

In coding, use can be made of classifying principles that form the basis for a classification, such as the different dimensions that could play a role. Often, these dimensions can be mapped onto the different hierarchies in a classification tree: in figure 1, level 2 (codes 1.1 and 1.2) may relate to the concept inside or foreign trade, e.g. It would be a good idea to explicitly describe these classifying principles with a classification. Unfortunately, in practice, these kinds of principles are not always explicitly formulated, which means that one has to make guesses about them. It is also possible that a classification is set up based on clear principles, but that the practical situation forces compromises to

³ The assumption that, for a description, only the separate words that occur play a role, and not the order and the combinations of these words in the description.

be made, or even forces some principles to be violated. These inconsistencies in the classification will hinder the coding of text towards itself.

In a classification based on a tree structure, it is possible to assign codes to the nodes (or: vertices) such that they reflect this structure.

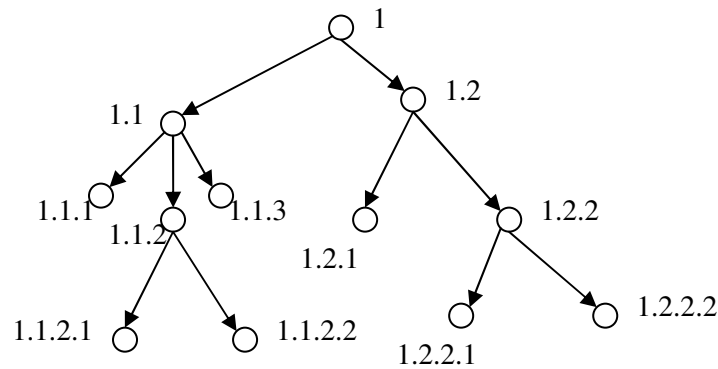


Figure 1. Example of a directed tree with labels for the nodes

Classifications consist of a set of categories, which also have a relationship among themselves. This relationship moves from general to more specific (i.e., in the direction of the arrows).

To clarify the difficulties that may arise from inconsistent classifications we will describe a few peculiarities that can occur in classifications. In the current Dutch standard industrial classification, we have a category ‘clothing’ that can be split in different ways, depending on the context. The ‘manufacture of clothing’ is split into the ‘manufacture of outerwear’ and ‘manufacture of underwear’. However, in the clothing retail trade, this category is split into: ‘retail trade of women’s clothing’, ‘retail trade of men’s clothing’, ‘retail trade of children’s clothing’ (and, perhaps, also the ‘retail trade of baby clothing’).

In the Dutch standard industrial classification (SBI-93), different splits of clothing are possible:

- According to age: clothing for babies, toddlers, children, teenagers and adults.
- According to gender: women’s and men’s clothing.
- According to how it is worn: underwear and outerwear.
- According to use: work clothing (including uniforms), leisure clothing, clothing for going out, clothing for formal events (for weddings, for academic events, such as receiving a PhD, for a fancy ball, receiving a medal, etc.), and everyday clothing.

Depending on the industry sector, the above splits may or may not apply.

Another example, also from the Dutch standard industrial classification (SBI-93), concerns agricultural products. The activity associated with these products determines the splitting level of these agricultural goods:

- *Cultivation* of vegetables. (Additional detail about these vegetables is not necessary.)

- *Wholesale trade* in potatoes for seed and potatoes for the retail market. (A split into the type of potato is necessary in order to code the type of wholesale trade.)
- *Processing* of potatoes. In other cases, different splits can occur.

Another problem arising from the classification definition is the following: how far apart, conceptually, are the categories? If there are two categories that are rather close together, then, in practice, it will be difficult to make a distinction between the two, based on descriptions. In this case, the descriptions must be quite precise. This can be difficult for a respondent who is not considered to be familiar with the classification, because this person will not be aware of a – probably quite subtle – difference between the two categories.

Finally, coding problems may arise from the lack of examples for certain codes: such a lack makes it difficult to construct an informative base or train a machine-learning approach for these codes.

These different problems require different solutions that, however, are not always available. After all, there are more issues that play an important role in official classifications than these methodology-related matters. Usually, a classification was already officially established at an earlier date. This must be viewed as given for the coding process. Changes to a classification generally are made with regards to the subject matter itself and not with observation / measurement in mind nor the coding problems that the classification poses when used in practice. It would be preferable if a classification was set up also addressing these issues. Experiences could then be used to adapt a classification and make it useful and applicable. There is little sense in retaining a theoretically ideal classification that cannot be used in practice due to observational or coding problems.

2.4 *Misconceptions about coding*

Here we want to point out several widespread (and persistent) misconceptions.

1. *‘Low quality input versus high quality output’*. This misconception conflicts with the truism: ‘garbage in, garbage out’. A source of the trouble may be that the classification distinguishes codes/subjects that seem the same to a ‘naive’ respondent. In this case input data are obtained that do not offer adequate information for correct and sufficiently detailed coding. The remedy for this could be as follows (while interviewing): in the event of vague / ambiguous texts, one can permit an appropriate code (a ‘doubt category’, or a less detailed code) or multiple (detailed) codes instead of a single code, possibly with probabilities assigned to the possible codes. Other solutions are better wording of the questions to elicit more precise answers or the use of interactive coding modules (when using CAPI, CATI or CAWI) to refine initial answers through feedback using further questions.
2. *‘Less detailed codes and therefore a higher yield’*. This may not be true in case the classification used is skewed at various levels, in the sense that the distribution of its scores in the population is skewed. An extreme example is shown in Figure 2, which is skewed on all hierarchical levels. In general, the link between a description and a less detailed code is not necessarily less ambiguous: if we would code all the occupations in the government by a code ‘occupation in the government’, this would not necessarily simplify the coding. For the practical situation, such a skewed distribution may imply that we can obtain a reasonable coding result with relatively little effort (i.e., by coding the most frequently occurring codes),

while a relatively large amount of effort must be put into the remaining part. If the coded corpus has a skewed distribution, then, typically, there are classes in the tail of the distribution for which too few examples are known to make a reliable and complete classification or classification model. Coding less detailed, i.e., at a higher level in the classification tree, doesn't present a solution, since this skewedness is often present on multiple levels. In Figure 2, for example, the skewedness occurs at both lowest levels. This skewedness may be due to classifications that are designed in a rather unbalanced way. It is also possible that they become more skewed due to changes in the population: certain NACE codes gradually disappear, while others start to occur more often.

3. '80% automatic coding is attainable'. The general opinion is that coding is an easy task. However, in our experience a yield of 40% of automatically coded records is a more realistic figure than the 80% claimed⁴. The yield strongly depends on the complexity of the coding problem. This also involves a difference in definition: if the classification literature for example refers to, for example, a percentage of 70% coded records, then this means that, of the 1000 texts, some 700 were correctly coded automatically. However, when the system must not only code, but also 'guarantee' some quality level of coding, the coding yield drops significantly. Another frequently occurring reason for overly optimistic estimations in the literature is that experiments have only been performed with codable descriptions, and that the non-codable descriptions have been ignored. It also plays a role in the validation of a coding system.

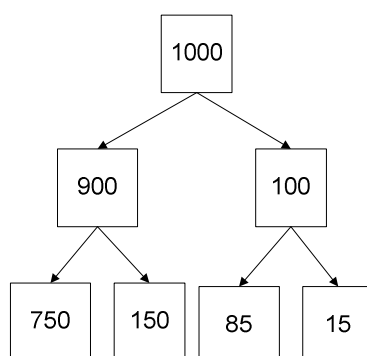


Figure 2. Example of an asymmetrically distributed corpus at all levels

3. Design issues

The approach to a new coding problem is driven by the following aspects:

- *The material that is available*: Is there already coded material in electronic form? Methods based on already coded material (as described in "Coding – Automatic Coding Based on Pre-coded Datasets") use this to train their machine-learning model. Methods that are applicable when there is no coded material (as described in "Coding –

⁴ The coding rate of 40% applies to the more challenging coding problems, e.g., when coding occupation with more than 1000 codes. If the respondents are experts at the classification, this rate may increase. In addition, coding very simple classifications (Municipalities or Country) may result in coding rates over 95%.

Automatic Coding Based on Semantic Networks”) construct an informative base to guide the coding of texts; having coded answers may help the construction of such informative bases.

- *The available software*: this point is strongly related to the previous point. Depending on the available software, one may start to construct an informative base or code a representative set of descriptions to feed a machine-learning method.
- *The coding method used*: manual, interactive or automatic / in batch. These methods can be combined into a strategy, e.g., start with interactive coding in the field followed by automatic coding at the statistical office (see “Coding – Different Coding Strategies”). The different individual approaches have been described in “Coding – Manual Coding”, “Coding – Computer-Assisted Coding”, “Coding – Automatic Coding Based on Pre-coded Datasets” and “Coding – Automatic Coding Based on Semantic Networks”.
- *The intended quality of the coding*: is it important that a lot of descriptions are coded, and that errors are accepted in some cases? Or does one take a more cautious attitude and should every code be correct with high probability? Besides measuring the quality, this may also give input for the enhancement/extension of the informative base or retraining of the machine-learning model. For more information see “Coding – Measuring Coding Quality”.
- *Maintenance*: How well can a coding strategy be kept up-to-date? The classification may alter its form year to year, new answers may be used to enrich the informative base. How to construct and to maintain such an informative base is described in “Coding – How to Build the Informative Base”.

4. Available software tools

Many of the methods described in the literature and their implementations are still in an academic phase, making their real-world application not (yet) really feasible. The following generic coding tools have been developed at several statistical offices or companies:

- Blaise: The Blaise suite contains several possibilities to search through classification(tree)s.
- SICORE from INSEE (see Rivière, 1994): This is based on decision trees.
- GCODE (successor of ACTR) from Statistics Canada (see Wenzowski, 1988): This is based on a kind of Nearest Neighbour technique.
- StafS from SPSS.
- Cascot from the Warwick Institute for Employment Research of the University of Warwick, UK. (See <http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/>.)

In addition to these there is an abundance of specific coding tools, geared at a particular application. Every NSI has probably a few of them. They are usually not supported and are of limited use outside the NSI where they are used.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Hacking, W. J. G, Michiels, J., and Janssen-Jansen, S. (2006), Computer Assisted Coding by Interviewers. Blaise Users Conference 2006.

Hacking, W. and Willenborg, L. (2012), *Coding – interpreting short descriptions using a classification*. Contribution to the CBS Methods Series, Statistics Netherlands, The Hague and Heerlen.

Rivière, P. (1994), The SICORE automatic coding system. Working Paper, Conference of European Statisticians, Cork.

Wenzowski, M. J. (1988), ACTR – A Generalised Automated Coding System. *Survey Methodology* **14**, 299–308.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Object Matching (Record Linkage)
2. Coding – How to Build the Informative Base
3. Coding – Different Coding Strategies
4. Coding – Measuring Coding Quality
5. Derivation of Statistical Units – Derivation of Statistical Units

9. Methods explicitly referred to in this module

1. Coding – Manual Coding
2. Coding – Automatic Coding Based on Pre-coded Datasets
3. Coding – Automatic Coding Based on Semantic Networks
4. Coding – Computer-Assisted Coding

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 5.2 Classify and code
2. 5.5 Derive new variables and statistical units

12. Tools explicitly referred to in this module

1. Blaise
2. ACTR / GCODE
3. Cascot
4. SICORE
5. StafS

13. Process steps explicitly referred to in this module

1. Input
2. Throughput

Administrative section

14. Module code

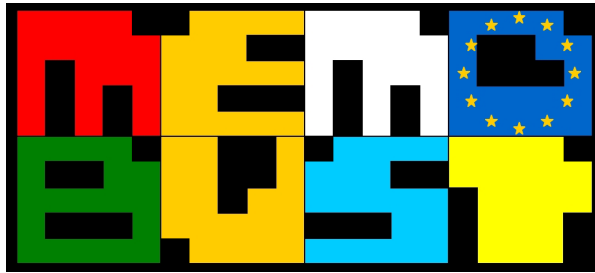
Coding-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	03-06-2012	first version	Leon Willenborg	CBS
0.2	31-10-2013	revised version	Wim Hacking	CBS
0.3	24-01-2014	revised version	Wim Hacking	CBS
0.4	19-02-2014	revised version	Wim Hacking	CBS
0.4.1	20-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:05



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: How to Build the Informative Base

Contents

General section.....	3
1. Summary	3
2. General description.....	3
3. Design issues	6
4. Available software tools.....	6
5. Decision tree of methods	7
6. Glossary.....	7
7. References	7
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

Coding of verbal responses of a statistical survey could be defined as assigning numeric codes to statements according to a manual of official classification. Performing this activity manually is costly, time consuming and error prone, so the computer support for this process is increasing. The informative base to be used for this purpose is the fundamental part of any computerised approach, because it must fulfil at least two requirements: make the computer rely on knowledge similar to that inside the human mind and, starting from the content of the official classification manual, process and enrich it with selected descriptions and/or synonyms derived from empirical responses given in previous surveys, so as to make the language closer to the spoken one. Logical steps to be carried out to build informative bases are described, as well as particular aspects to be taken into consideration either when coding is done in a completely automated way or with human support.

2. General description

Coding of verbal responses of a statistical survey could be defined as assigning numeric codes to statements according to a manual of official classification. The knowledge concerning these official classifications is usually contained in the classification manuals which describe, using appropriate words, the meaning of each concept and its code, providing definitions, specific details and exceptions.

When the coding process is done manually, the coder must be trained on using the classification and on how to find the information in the manual to assign the correct code corresponding to the textual response. But, as confirmed by the experience of a lot of NSIs, manual coding is time consuming, costly and error prone, so computer support for this activity is desired. It is also evident that when the computer is used, it must be given additional information from human experts lacking from the classification manual. As a matter of fact there are at least two aspects a human mind can consider while reading that a computer cannot, if not trained:

- * the grammar and syntax rules (singular/plural, masculine/feminine, verbs declinations, ...);
- * semantic knowledge (the computer does not know the real meaning of words, it does not know, for instance, that an *orange* is a *citrus fruit*).

For these reasons a computer tool, in order to be used to code text responses, should:

- * have an informative base supplying the computer knowledge similar to that from the human classification experts;
- * have a search engine able to perform a text standardisation so as to identify a word independently from all the variable parts of it.

Both these aspects are even more important in two situations:

- * when the coding process is made automatically, with a batch procedure and without any human intervention;
- * when the coding process is made with computer support and directly by the respondent (in self-administered interviews).

As a matter of fact, in the first situation the computer must reason as a human mind does, because it has no other input at its disposal apart from its informative base and the responses to be coded. In the second situation it must be considered that the respondent, differently from the interviewer, is not an expert of the classification and might not know technical words used in manuals.

Regarding the informative base, this is the fundamental part of any computerised approach for coding. It is mainly constituted of a *dictionary* containing words or phrases associated with numeric codes, that represent the possible values to be assigned to the variables entering the coding process. The dictionary has to contain the definitions of official classifications – that constitute the starting point for the construction of the database itself – as well as the empirical responses coming from previous surveys or pilot studies. This mixture of official and empirical definitions helps the coding procedure to take into account both the official and the common language. Besides, a continuous update of the dictionary is necessary to cover the variability of the spoken language – a lot of different words to express the same concept – and also to take into account its continuous changes.

As far as the search engines for text processing are concerned, they can be more or less sophisticated, but it must be considered that text responses of statistical surveys to be coded according to official classifications are generally not too long and usually do not consist of very complex syntax constructions.

Several studies have been made at NSIs to identify or develop suitable tools to process texts in order to perform coding (Lyberg and Dean, 1992): in the late sixties, the US Census Bureau realised different coding systems, called “*dictionary algorithms*”, that build the dictionary on the base of a large sample of verbal responses manually coded by experts. The simplest algorithms for automated coding software build the dictionary searching for an exact match, that is, searching for the verbal description in the expert coded file that perfectly corresponds to the verbal response to be coded. Other dictionary algorithms include in the dictionary a description belonging to the expert-coded file if it contains a “*classifier*”, that is to say, a word or a set of words corresponding to a specific code and whose occurrence is not lower than a defined level.

Other coding systems use a so-called “*weighting algorithms*”, that are a bit more complex than the previous ones. They assign to each single word of the input statement a weight that indicates how much a word is informative; the calculation of the weight is based on the occurrence frequency of each word in the dictionary. Afterwards, the computer searches for the input verbal response inside the dictionary: if no exact match is found then it analyses those descriptions that are “similar” to the input one and chooses the one with the highest weight, thus realising a “*partial match*”¹. This feature – *partial match* – represents the main difference between the *dictionary* and the *weighting algorithms*.

More articulated coding systems have been developed subsequently. Some of them - like BLAISE, Netherlands CBS - perform a partial match for both entire word and sub-strings, that is, for groups of consecutive letters of a word, thus widening the possibility of assigning the right code.

Other more sophisticated instruments use the so-called “*artificial intelligence*”. One of these is the “Connection Machine” – Thinking Machine Corp. – that is a computer working with thousands of

¹ The system mainly used in Istat in several surveys, ACTR -Automatic Coding by Text Recognition, produced by Statistics Canada (Wenzowski, 1988), is based on a weighting algorithm. The new release of ACTR is called GCode.

processors in parallel (each representing a category – group of codes – of the official classification) that search for a code simultaneously. The peculiarity of the Connection Machine relays in its *memory based reasoning*: when searching for a match for a new input verbal response, the PC recalls codes that were attributed to similar past descriptions (Appel and Hellerman, 1983).

Whichever coding system is adopted, the problems faced when building these dictionaries are the same: first of all the official classification manual must be transformed so as to be ‘processable’ by computerised systems and then lots of sources must be integrated in order to make it closer to spoken language used by respondents. As a matter of fact, textual descriptions of classifications are designed for manual coders, who assign codes making deductions and referring to their specific knowledge of the matter or to their personal cultural background (Knaus, 1987). A ‘processable’ dictionary, on the contrary, should include only **synthetic**, **analytical** and **unambiguous** descriptions (while two or more different descriptions can be associated to the same code, the same description must never be associated to different codes).

In general, the following logical steps can be defined in the activity of construction of dictionaries (D’Orazio and Macchia, 2002):

- **Simplifying descriptions** → often a description which summarises more than one concept is associated to a single code, while the typical respondent is used to refer to a single concept (for example: the Istat classification on Occupation assigns a single code to ‘*mathematicians and statisticians*’, while the respondent will presumably answer only ‘*mathematician*’ or ‘*statistician*’, according to his specialisation). In these cases, it is necessary to split the phrase in two or more descriptions and to associate each of them to the same code.
- **Defining synonyms** → classifications contain generic words relating to categories, while people answer using specific words (for example, the Economic Activity classification considers the ‘*production of cereals*’, while the respondent might answer only ‘*production of wheat*’ or ‘*production of corn*’). Here it is necessary to list all the specific synonymous words to which the generic word refers to.
- **Eliminating exception clauses** → automated coding software do not usually reason in terms of exclusion, so they cannot understand the meaning of ‘apart from...’ and of similar clauses used to exclude certain categories from the class. In this case, it is necessary to take the ‘apart from...’ away and to verify that the ‘excluded’ concepts are included in other classes.
- **Treating open classes** → classifications usually include open descriptions, that is ‘Other ...’ which means ‘other than the concepts already specified’ (for example: ‘*Other specialised clerks*’, where different kinds of specialised clerks have already been listed in the preceding classes). Also in this case it is necessary to list all the explicit descriptions to which the open description refers. To make the list as complete as possible, it is advisable to use the responses given in previous surveys which have been coded as ‘Other...’ by expert coders.
- **Integrating with reference material** → the ‘processable’ dictionary can be usefully widened with descriptions coming from other related classifications. For example, each time the classification of Economic Activity has an element regarding the production of a certain ‘*category of products*’ (summarising an implicit list) the specific classification of products can be used to enumerate the explicit list of products.

- **Integrating with empirical responses** → official classifications texts are often not very similar to the way people speak and their updating is slower than real world changes. Thus it is advisable to include in the dictionary selected descriptions derived from empirical responses, given in previous surveys, that had already been coded by classification experts.

As the dictionary is extended according to these criteria, its performance will increase, especially in the case of automated coding.

As far as Istat's experience is concerned, for instance, an application to code with the Economic Activity classification has been set up since 1998 and was updated following the new classification releases and used for several surveys. The informative base was built starting from the classification manual and enriched through the analysis of results of its use in each survey. As a matter of fact, after using this application to code the textual responses of a survey, non-coded responses were examined to find the cause, e.g., an ambiguity in the original text or the response contained some synonyms not present in the informative base. In the latter case, the missing synonym was added to the informative base.

In order to give an idea about the impact of the size of the informative base on the results of automatic coding, the dictionary grew from 27,306 descriptions to 34,180 and the average percentage of coded texts (on the total number of texts to be coded) from 50% to 71% in the last surveys (Macchia, Murgia, and Vicari, 2010).

Finally, it must be mentioned that, when the coding process is made interactively with the computer support (during the interview, either by the respondent or by the interviewer, or after the interview by coders), computer tools often provide functions to navigate inside the informative base.

When the classifications have a hierarchic structure it is advisable that the software tools allow to navigate inside the dictionary according to the classification tree.

Blaise for instance, developed by Statistics Netherlands, manages navigation in its coding database according to three different methods:

- through the textual matching → the respondent/coder enters the text and Blaise extracts from the database the texts which have one or more trigrams in common with the keyed text;
- according to the classification tree → the respondent/coder selects the classification branch of the highest level and then goes deeper in the sub-branches towards the lower levels;
- with a mixed method → the respondent/coder selects the classification branch of the highest level and then enters the text to perform textual matching among descriptions belonging to the selected branch.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Appel, M. and Hellerman, E. (1983), Census Bureau Experience with Automated Industry and Occupation Coding. *Proceedings of Section on Survey Research Methods*, American Statistical Association, 32–40.

BLAISE for Windows 4.5 Developer’s Guide (2002).

D’Orazio, M. and Macchia, S. (2002), A system to monitor the quality of automated coding of textual answers to open questions. *RESEARCH IN OFFICIAL STATISTICS (ROS)*, N.2.

Knaus, R. (1987), Methods and problems in coding natural language survey data. *Journal of Official Statistics* **1**, 45–67.

Lyberg, L. and Dean, P. (1992), Automated Coding of Survey Responses: an international review. Conference of European Statisticians, Work session on Statistical Data Editing, Washington DC.

Macchia, S., Murgia, M., and Vicari, P. (2010), Integration between automatic coding and statistical analysis of textual data systems. Journée d’Analyse des Données Textuelles JADT, Rome.

Wenzowski, M. J. (1988), ACTR – A Generalised Automated Coding System. *Survey Methodology* **14**, 299–308.

Interconnections with other modules

8. Related themes described in other modules

1. Coding – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Sub-process 5.2 Classify and code

12. Tools explicitly referred to in this module

1. BLAISE for Windows
2. GCode, new release of ACTR -Automatic Coding by Text Recognition

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

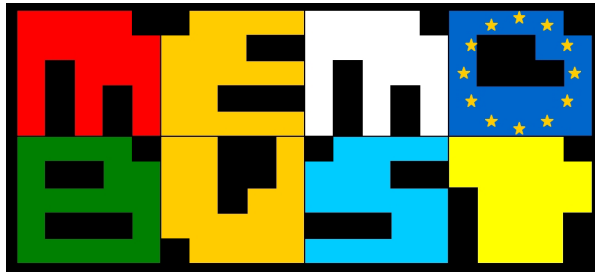
Coding-T-Informative Base

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	20-07-2012	first version	Stefania Macchia	Istat (Italy)
0.2	21-11-2012	second version (following first revision)	Stefania Macchia	Istat (Italy)
0.3	17-01-2014	third version (following EB review 08-01-2014)	Stefania Macchia	Istat (Italy)
0.3.1	21-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:05



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Manual Coding

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	3
4. Examples – not tool specific.....	4
5. Examples – tool specific.....	4
6. Glossary.....	4
7. References	4
Specific section.....	5
Interconnections with other modules.....	7
Administrative section.....	8

General section

1. Summary

We will briefly describe some aspects related to the manual coding of open text answers. In this era dominated by the use of computers, most of the coding is done either by computers or at least computer-assisted. However, there always remains a small part that cannot be coded and needs the attention of an expert.

2. General description of the method

This module will focus on the organisational aspects related to manual coding. Nowadays, manual coding without computer support seems almost unthinkable: for aspects related to the computer-assisted part, see the module “Coding – Computer-Assisted Coding”. Apart from that a number of issues remain:

- Administrative tool(s): when a group of coders is working on the coding of text, it may be useful to have some sort of administrative (workflow) tool to distribute the workload amongst the coders. If there are more classifications to be coded it becomes even more convenient. Such a tool may also include import (from the survey) and export (to the subsequent processing of the survey) options. Also, the coding itself could be done using such a tool: displaying the text(s) relevant to code the answer and allow for searching/browsing the classification.
- Knowledge sharing: many descriptions will be coded, but a small fraction cannot, either because the information is vague or ambiguous. These cases can be discussed and rules can be established to code these difficult descriptions. These discussions will enhance the standardisation of the coding process and help to share knowledge.
- Educating new coders: educating a coder is mostly training on the job. In practice they will code the easier codes at first and leave the ambiguous descriptions to the experienced coders. In our experience it may take many months before they code at the level of the existing coders, for the more complex codes.
- Interaction with code designers: often, the coders are not the ones that design or maintain the classification. Therefore, some amount of interactions is wanted: on the one hand the “code designers” need to explain the philosophy of the classification (e.g., by what criterion are certain occupations grouped together). On the other hand, the coders need to give feedback to the “code designers”: for example, certain distinctions in the classifications may be too subtle which makes it hard to code.

3. Preparatory phase

The preparation consist mainly of the training of the coders. This is a continuous process, since classification systems change over time and there are frequent changes of coders in the coding teams. Therefore, the (new) coders must build up experience with new codes.

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Hacking, W. and Willenborg, L. (2012), *Coding – interpreting short descriptions using a classification*. Contribution to the CBS Methods Series, Statistics Netherlands, The Hague and Heerlen.

Specific section

8. Purpose of the method

The purpose of manual coding method is to describe relevant (e.g., organisational) aspects when coding open text answers from surveys.

9. Recommended use of the method

1. When trying to code open texts from surveys, the order should preferably be:
 - a. Code the answers (semi-)automatically during the interview.
 - b. Code the answers automatically at the statistical office.
 - c. Code the answers manually at the statistical office.

The idea is that the most expensive step is done last.

2. Another reason may be the difficulty of the texts and the necessity to use other variables to arrive at a valid code. Such complex coding processes are much more difficult to automate. Hence, manual coding can be used:
 - for all the texts collected;
 - only for those texts which could not be coded with a computer-assisted method (automatic coding or assisted coding).
3. Due to the disadvantages implied with manual coding, it should be better to:
 - use this method only for texts not coded with the computer assistance;
 - when the text to be coded is not sufficient to assign a code, use other variables to arrive at a valid code.

10. Possible disadvantages of the method

1. Compared to other coding scenarios, manual coding is rather expensive. Other disadvantages are:
 - this method is error prone when the knowledge/experience of the coder is not sufficient;
 - manual coding results in less standardisation of the process (each coder, even if well trained, has his own knowledge and can make deductions according to his interpretation of the text).

11. Variants of the method

- 1.

12. Input data

1. A text to be coded, possibly combined with a number of other variables correlated with the classification at hand, e.g., the kind of goods when coding economic activity.

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

- 1.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. For each input text, a code is added unless the text has not an informative content sufficient to assign a code and there are no other variables which could help to arrive at a valid code.

17. Properties of the output data

- 1.

18. Unit of input data suitable for the method

Incremental processing

19. User interaction - not tool specific

- 1.

20. Logging indicators

1. The coder may log which variables were used to arrive at a code. Such a scenario would only be feasible if the manual coding is supported by a computer program, though: the program could trace the interactions of the coder while trying to arrive at a proper code. For example, the coder may use additional variables if the text is ambiguous.

21. Quality indicators of the output data

1. The quality of coding can be measured with two indicators (as described in “Coding – Measuring Coding Quality”):
 - Coding rate (efficacy) → percentage of coded texts on the total of texts to be coded;

- Precision rate (accuracy) → percentage of *correct* coded texts on the total of coded texts.

The verification of coding can be performed by having a different team of coders recode a sample of the texts. If the original code and the verification code differ, the ‘correct’ code can be decided by expert coders by a reconciliation process. The set of correct codes can then be used to estimate the values for coding rate and precision rate.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Coding – Main Module
2. Coding – Measuring Coding Quality

24. Related methods described in other modules

1. Coding – Automatic Coding Based on Pre-coded Datasets
2. Coding – Automatic Coding Based on Semantic Networks
3. Coding – Computer-Assisted Coding

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. 5.2 Classify and code

27. Tools that implement the method described in this module

- 1.

28. Process step performed by the method

Coding

Administrative section

29. Module code

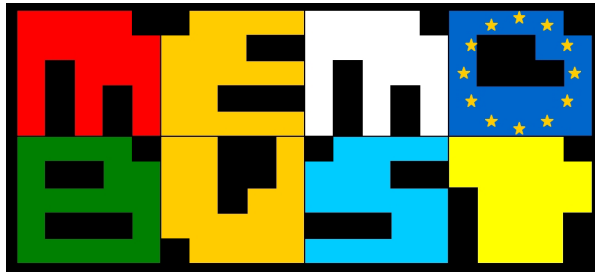
Coding-M-Manual Coding

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	02-04-2013	first version	Wim Hacking	CBS
0.2	20-01-2014	following review by Stefania Macchia	Wim Hacking	CBS
0.3	30-01-2014	following review from EB	Wim Hacking	CBS
0.3.1	30-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:06



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Automatic Coding Based on Pre-coded Datasets

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	6
4. Examples – not tool specific.....	6
5. Examples – tool specific.....	6
6. Glossary.....	6
7. References	6
Specific section.....	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

For a number of variables in questionnaires, one wants the answer in closed form, e.g., “city”; this is a relatively simple classifying task. Sometimes this task is much harder, e.g., when trying to get a code for occupation. One approach is to ask an open question (“what is your occupation”) and then try and code this text at the statistical office. For the sake of efficiency, that coding process will start by an automatic step.

Here we will describe the coding of open text answers based on existing sets of correctly coded answers. We will briefly look at some existing techniques and then focus on one method in more detail as an example.

2. General description of the method

We will first discuss briefly the general approach for (short) text classification in the literature, as described more extensively in Sebastiani (2001) and Joachims (2002).

The literature describes several techniques to classify text if a corpus is available, that is, previously coded (and verified) descriptions. Most of the literature concerns numerical rather than textual data, and is known as ‘pattern recognition’, ‘data mining’ or ‘business intelligence’. Recently, the application of these techniques and new ones to text has received much more attention due to the data explosion taking place at the internet; the terms used are ‘text-mining’, ‘web-mining’, etc. A number of these have been described in Sebastiani (2001). These are techniques based on data mining techniques, such as *K-Means*, *Naïve Bayes* and *Support Vector Machines*. For most of these techniques a description is represented by a very large sparse vector, where a 1 means that the word is present in the input and 0 that it is not; for this model, the order of the words is of no importance (the so called bag-of-words assumption). In order to reduce the size of these vectors, extra pre-processing techniques are used, such as Latent Semantic Indexing (Sebastiani, 2001). These classification problems are close to but not the same as the coding problem considered in the present document. The descriptions used in coding usually do not contain more than 10 words.

At statistical offices, automated coding based on pre-coded datasets (so called coding dictionaries) has been implemented using different matching algorithms. In the late sixties, the US Census Bureau realised different coding systems, based on “dictionary algorithms”, that build the dictionary on the base of a large sample of verbal responses manually coded by experts (Lyberg and Dean, 1992). The simplest algorithms searched for an exact match, while other ones were based on a classifier, that is to say, a word or a set of words corresponding to a specific code and whose occurrence is not lower than a defined level. Other coding systems use the so-called ‘*weighting algorithms*’, which are based on a measure of similarity between the text to be coded and those of the coding dictionary. In this way, these methods consider not only ‘*perfect matches*’, but also ‘*partial matches*’ between input texts and texts of the dictionary¹; this approach is comparable to *K-Means*.

¹ For this type of (so called fuzzy) string matching, see Hall and Dowling (1980) and Navarro (2001).

To illustrate the coding process based on pre-coded data, we will look at a method as applied at Statistics Netherlands, which belongs to the ‘*weighting algorithms*’.

As mentioned, this method is a nearest-neighbour technique, combined with a choice of a specific distance measure between two descriptions: first, each word (or combination of two words) is assigned a weight that indicates how specific that word is in the training set. This can be illustrated using Figure 1.



Figure 1. Examples of conditional distributions over (sorted) occupation classes given the key words 'lawyer' (a) and 'employee' (b).

Figure 1 shows histograms of $P(\text{Code}_i | \text{Word})$ (the probability of Code_i , given that Word is in the description). Subfigure (a) of Figure 1 shows the probability distribution (sorted by frequency) for $\text{Word} = \text{'lawyer'}$, and (b) depicts the histogram for $\text{Word} = \text{'employee'}$. The asymmetry of the distribution indicates how specific a word is. Following Chen et al. (1993), this specificity is quantified as²:

$$F(W) = \frac{\sqrt{\sum_{i=1}^n P(C_i | W)^2}}{n}, \quad (1)$$

where n is the number of codes C_i where W occurs in the description.

Based on this definition, a word such as 'lawyer' is assigned a higher weight than a word such as 'employee', when comparing two descriptions. Note that, in this way, words with little meaning such

² Other measures for this are: entropy and the skewness of the distribution.

as ‘and’, ‘the’, etc. (stop words) naturally have a minimal effect, because they are given a low weight if they had not been filtered out earlier in the pre-treatments. The pre-processing step in which stop words are removed could, in principle, be omitted.

Based on formula (1), defined *per word* (or *word combination*), we can define a measure for the similarity of two *descriptions* D_1 and D_2 (after removing the words that occur multiple times in both descriptions):

$$\text{Similarity}(D_1, D_2) = \sum_{x_i \in D_1 \cap D_2} F(x_i)$$

where

$$D_1 = \{a_1, \dots, a_n\} \text{ and } D_2 = \{b_1, \dots, b_m\} .$$

In other words, the similarity between two descriptions is determined by adding up the weights $F(x_i)$ of all shared words³. A new description D is compared with all the descriptions present in the training set, and the best N descriptions are retained. The code that occurs most often among the codes associated with the N best fitting descriptions is either selected (provided that it occurs frequently enough) or rejected, i.e., the algorithm cannot assign a code and the description will be presented to an expert.

Any matching algorithm, not just the example algorithm described above will return a list of possible codes along with some score. For automatic coding to work, this list must be reduced to either

- zero, i.e., even the top score code doesn’t have enough “confidence”;
- one, i.e., the description gets classified.

The first choice is important for practical implementations: a coding algorithm cannot classify every description, so some fraction of the set of descriptions must be passed on coding experts. This means that the automatic coding algorithm must do two things:

- try to classify the description;
- try to assign a measure of confidence in the assigned classification.

In the current example above the selection is done as follows. Let C be the most frequently occurring class among the N descriptions. C is selected unconditionally if $\#C \geq f_{GOOD} * N$, where $\#C$ is the frequency score of C among the scores associated with the N descriptions. If it is true for $\#C$ that $(f_{BAD} * N \leq \#C \leq f_{GOOD} * N)$, then the selection of C is doubtful, and is presented to a specialist coder on this topic, who must then decide whether or not to assign C . If $\#C < f_{BAD} * N$, the selection of C is rejected: it simply does not occur frequently enough. The choices of f_{GOOD} and f_{BAD} are empirically determined using the data; for a higher quality (but less coded answers), these can be higher than for a lesser quality (but more coded answers).

³ The use of synonyms, hyponyms and hypernyms could further increase the returns of the matchings; this has not yet been studied.

3. Preparatory phase

To prepare this method, one needs a sufficiently large pre-coded dataset. Not only the number of records is important, but one also has to check if each code has a sufficient amount of records; otherwise the classification becomes rather unreliable for those codes.

4. Examples – not tool specific

5. Examples – tool specific

To our knowledge, there is only one general coding system, based on pre-coded texts, that is currently available: ACTR (Wenzowski, 1988); this system allows the coding of texts based on data-mining techniques and has many pre- and post-processing options to make it suitable for any given classification (ACTR has been enhanced and is currently called GCODE). In addition, there are two other systems that are designed generically, but are only used at the statistical office where they were created (Hacking and Janssen-Jansen, 2009; Hacking and Willenborg, 2012; Rivière, 1994).

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Creecy, R. H., Masand, B. M., Smith, S. J., and Waltz, D. L. (1992), Trading MIPS and memory for Knowledge Engineering. *CACM* **35**, 48–63.
- Chen, B., Creecy, R., and Appel, M. (1993), Error control of automated industry and occupation coding. *Journal of Official Statistics* **9**, 729–745.
- Hacking, W. J. G. and Janssen-Jansen, S. (2009), The coding of economic activity based on spreading activation. Report, Statistics Netherlands, Heerlen.
- Hacking, W. and Willenborg, L. (2012), *Coding – interpreting short descriptions using a classification*. Contribution to the CBS Methods Series, Statistics Netherlands, The Hague and Heerlen.
- Hall, P. V. and Dowling, G. R. (1980), Approximate string matching. *Computing Surveys* **12**, 381–402.
- Joachims, T. (2002), *Learning to classify text using support vector machines*. Kluwer.
- Lyberg, L. and Dean, P. (1992), Automated Coding of Survey Responses: an international review. Conference of European Statisticians, Work session on Statistical Data Editing, Washington DC.
- Navarro, G. (2001), A guided tour to approximate string matching. *ACM Computing Surveys* **33**, 31–88.
- Rivière, P. (1994), The SICORE automatic coding system. Working Paper, Conference of European Statisticians, Cork.

- Sebastiani, F. (2001), Machine learning in automated text categorization. *ACM Computing Surveys* **34**, 1–47.
- Wenzowski, M. J. (1988), ACTR – A Generalised Automated Coding System. *Survey Methodology* **14**, 299–308.

Specific section

8. Purpose of the method

The automatic coding step takes place just after the data have been collected, in most cases data from interviews. These interviews contain a few fields that are input for the coding step, e.g., “production of wooden crates” as input for the coding of economic activity. For simple classifications, e.g., “nation of birth”, a closed question can suffice in the interview; for more complex classifications such as education or occupation, a closed question will lead to long lists and the quality of the response will decrease rapidly. For that reason, it is often more logical to use an open question in the interview and code this answer at the statistical office; for reasons of efficiency, it is better to start coding automatically followed by a manual or a computer-assisted coding step (texts not coded automatically can be analysed by expert coders manually or with the computer support)

The method described here can be used to do the automatic coding step.

9. Recommended use of the method

1. Recommendations on the use of the different methods for coding (automatic or assisted) are given in the module “Coding – Different Coding Strategies”: the decision about which is the most suitable coding approach to be adopted in a survey depends on different correlated factors. If it has been decided to use automatic coding, one needs a training set of coded descriptions that is available in electronic form, and a correct code (after verification) that is assigned to each description.

10. Possible disadvantages of the method

1. The main disadvantage occurs when the classifications changes (which is not uncommon). Especially, if it is a large change, many coded records need to be recoded and often there does not exist a 1:1 mapping between old and new codes. As a result, a large portion of the pre-coded material needs to be recoded. If one uses a semantic network, the resulting amount of rework is much less; of course this depends on the way how the network is constructed.

11. Variants of the method

- 1.

12. Input data

1. During the coding phase, the input is quite simple: a textual description, in most cases no more than 10 words. During the construction of the “coding machine” the input consists of pre-coded datasets that are used to train the coding algorithm, i.e., a set of records containing a description and a correct code.

13. Logical preconditions

1. Missing values

- 1.

2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. The input text to be coded should not be too large; in general, this will result in many possible codes for this input text by the method.

14. Tuning parameters

1. In general, all practical automatic coding algorithms need a *score cut-off value*, to make a selection which descriptions are coded and which need manual or assisted coding. In our experience this parameter is rather robust.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. Per description the following is derived by the method:
 - a score;
 - a classification code.

17. Properties of the output data

- 1.

18. Unit of input data suitable for the method

Incremental processing

19. User interaction - not tool specific

1. None

20. Logging indicators

1. To monitor the quality of the coding process, all relevant parameters influencing the classification must be stored for later analysis, i.e., comparing (a subset of) the automatically coded texts with correctly coded.

21. Quality indicators of the output data

1. The method is tested by splitting a pre-coded data set into two parts: 10% test set and 90% learning set. After training the method with the learning set, the method is tested by feeding it the descriptions from the test set; after coding, both set of codes (N) from the algorithm and the test set are compared.

A small part of the test set will be rejected ($N_{rejected}$) by the algorithm (e.g., because it's too vague), and the non-rejected matching part ($N_{Coded} = N - N_{rejected}$) is used for the comparison; the number of descriptions that were coded correctly is $N_{CorrectlyCoded}$. As described in the module “Coding – Measuring Coding Quality” we can use two measures to quantify the quality of the automatic coding method:

$$coding_rate = \frac{N_{Coded}}{N}, \text{ i.e., what fraction was coded;}$$

$$precision_rate = \frac{N_{CorrectlyCoded}}{N_{Coded}}, \text{ i.e., what fraction was coded correctly.}$$

22. Actual use of the method

1. This method was used by the US Census Bureau already in the nineties (Creecy et al., 1992). It is also used at INSEE (the SICORE system; Rivière, 1994), at Statistics Canada and ISTAT (ACTR, now G-CODE; Wenzowski, 1988). This method is also used at Statistics Netherlands (Hacking and Willenborg, 2012).

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Coding – Main Module
2. Coding – Different Coding Strategies
3. Coding – Measuring Coding Quality

24. Related methods described in other modules

1. Coding – Manual coding
2. Coding – Automatic Coding Based on Semantic Networks
3. Coding – Computer-Assisted Coding

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. 5.2 Classify and code

27. Tools that implement the method described in this module

1. G-CODE: see Wenzowski (1988)
2. SICORE: see Rivière (1994)

28. Process step performed by the method

Coding

Administrative section

29. Module code

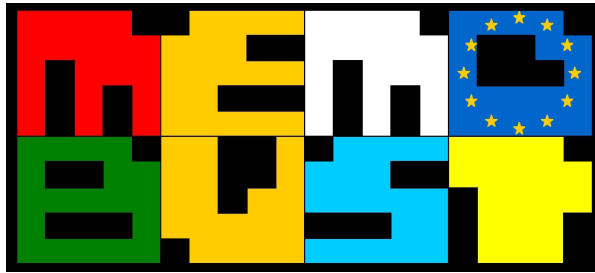
Coding-M-Automatic Coding Based on Pre-coded Datasets

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	02-04-2013	first version	Wim Hacking	CBS
0.2	20-01-2014	following review by Stefania Macchia	Wim Hacking	CBS
0.3	30-01-2014	following review from EB	Wim Hacking	CBS
0.3.1	30-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:06



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Automatic Coding Based on Semantic Networks

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Spreading activation	4
2.2 Algorithm	5
3. Preparatory phase	6
4. Examples – not tool specific.....	6
5. Examples – tool specific.....	6
6. Glossary.....	7
7. References	7
Specific section.....	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

For a number of variables in questionnaires, one wants the answer in closed form, e.g., “city”; this is a relatively simple classifying task. Sometimes this task is much harder, e.g., when trying to get a code for occupation. One approach is to ask an open question (“what is your occupation”) and then try and code this text at the statistical office. For the sake of efficiency, that coding process will start by an automatic step.

In some cases, no previously coded material is available in electronic form. The starting point then consists of the data to be coded and a classification with a textual description per code. In this situation, we can either build the informative base transforming the classification manual so as to be ‘processable’ by a computerised system and ensuring pre-coded descriptions or one must try and code open text answers based on the texts themselves and the associated semantics, to enable the approach from the module “Coding – Automatic Coding Based on Pre-coded Datasets”.

Although an informative base can be constructed based on expert knowledge, pre-coded answers may also be added to the informative base to enhance the coding rate. This makes the distinction with the module “Coding – Automatic Coding Based on Pre-coded Datasets” less strict. The main distinction between the latter module and this one is the amount of manual work to construct an informative base: the methods in the other module are based on machine-learning requiring much less manual work. As described in the module “Coding – How to Build the Informative Base”, the informative base can contain:

- the classification manual descriptions, transformed so as to be ‘processable’ by a computerised system;
- pre-coded descriptions collected in previous surveys;
- different kinds of synonymous, hypernyms and hyponyms.

There are general systems (ACTR, now G-CODE (Wenzowski, 1988) and Cascot (Cascot)) that use the elements above to code text in a number of steps, like pre-processing the text, replacing words and finally assign a code. Alternatively, most of these steps can be combined into a so-called semantic network (Hacking and Janssen-Jansen, 2009). In the following section we will describe the “spreading activation” search method in the semantic network in more detail as an example; at certain points we will describe the link with the “processing approach” in the ACTR tool.

2. General description of the method

Coding methods based on semantic networks (in its simplest form a search table) have in common that they are based on a number of relations between words or combinations of words; there is also a relation between combinations of words and classification codes. The most common relations are hypernyms, e.g., an apple is a kind of fruit; this kind of relationship allows the coding system to reduce the variation of words before performing the final coding step. Other relationships are synonym and hyponym (described in the next subsection). To code words are linked to classification codes (e.g.,

“carpenter” & “building site” → code 12345, “carpenter” & “factory” → code 12344, ...) or there is a more advanced algorithm (e.g., Cascot¹) that derives codes from a pre-processed description.

Here we describe an algorithm to use all of the semantic information in a single semantic network called ‘Spreading activation’. This method is described in more detail below.

2.1 Spreading activation

For the coding of SBI codes (the Dutch version of the NACE codes), a technique called ‘spreading activation’ is used, where coding is performed based on a semantic network (which may have been created manually). This is a directed graph, also called a digraph, where the nodes represent words, and where the edges or directed edges (or arcs) indicate relationships between words (the exact relationship is stated by listing that next to an arc). For example:

- greenhouse vegetables $\xrightarrow{\text{hypernym}}$ tomato: greenhouse vegetables include tomatoes.
- tomato $\xrightarrow{\text{hyponym}}$ greenhouse vegetables: tomatoes are a kind of greenhouse vegetables.
- Agatha $\xrightarrow{\text{synonym}}$ potato: for the classification, the potato varieties like ‘Agatha’, ‘Anya’, ‘Fingerling’, ‘Jersey Royal’, ‘Kerr’s pink’, etc. are not important, and if they do occur in a description they can be considered synonyms to ‘potato’ which can be used instead.
- sale_of_childrens_clothing $\xrightarrow{\text{Code}}$ 12345, because the description ‘sale of children’s clothing’ unambiguously leads to the code ‘12345’.

These relationships² form a semantic network, of which a small part is shown for illustrative purposes in Figure 1.

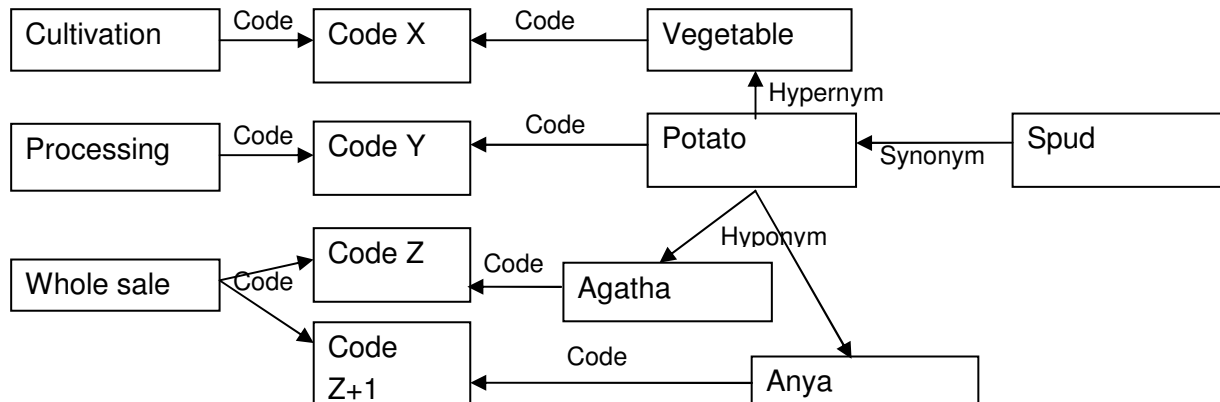


Figure 1. A fragment from a semantic network, as used in the coding of the SBI (Agatha and Anya are two varieties of potato).

¹ Unfortunately the actual coding/scoring algorithm is not described.

² Synonym, hyponym and hypernyms more or less correspond to the ACTR preprocessing step of replacing words. The ACTR step *remove words* simply corresponds to non-existing nodes in the semantic network. The code relationship corresponds to assigning certain codes a score, given the input description (the *weighting step* in ACTR).

Consider the description ‘cultivation of spud’ will give code X a score 2, whereas the other codes will get a score of 1 at the most: the word ‘spud’ leads to code X³, through ‘potato’ and ‘vegetable’); the word ‘cultivation’ directly leads to this code. Hence, using hypernyms allows the description for code X (‘cultivation of vegetables’) to be found, even though ‘spud’ is too specific.

2.2 Algorithm

Text In this network, we see interrelated words, due to certain semantic relationships. The words in a semantic network are also called nodes, for which an associated tag is ‘activation’; this serves to quantify the extent to which a word correlates with the terms from the search string. The binary or other relationships that exist between the nodes can (formally) be recorded in an adjacency matrix.⁴

In brief, the algorithm amounts to the following:

1. Let the set of nodes be denoted as $\{n_1, \dots, n_m\}$. Call the activation values during iteration round k of the nodes $A_k = (a_{k1}, \dots, a_{km})$. Call the adjacency matrix $P = (p_{ij})$.

This means:

- $p_{ij} = 1$ if there is a link between two nodes n_i and n_j ,
- $p_{ij} = 0$ if that is not the case.

2. Next: for each word stated in the description:

- a. Set the activity of the node linked with word l from the description to 1: $a_{ll} := 1$.
- b. After this, all nodes that can be reached by an arrow from ‘activated’ nodes are also activated by means of the following relationship:

$$a_{k+1,i} = \sum_{j} a_{k,j} \cdot p_{i,j}$$

This must only be done for nodes not yet visited. In addition, there is a special restriction for the *hypernym* and *hyponym* relationships (the *parents* and *children*): if a path has already run along a *hypernym relationship*, then it may not run along any other *hyponym* relationships, and vice versa.⁵

3. The ‘expansion’ of the activity stops because all paths ultimately ‘collide’ on a code node, or because there are no more unvisited nodes near a node. The codes then contain an activity as described in 2a and 2b. All codes with an activity > 0 , in order of activity, form the result of the search operation. In order to increase its effectiveness, one can only select those codes that have the same score as the top scoring node, e.g., if there is only 1 node with score 2 and 5 other having a score 1, the search method results in exactly 1 code, which is the desired for an automatic coding method.

³ Actually, in this example, it leads to all codes.

⁴ The actual implementation is likely to be different from the description given here, as this would be very inefficient. It is only for the sake of the explanation of the algorithm that an adjacency matrix is used.

⁵ If this restriction were not present, then all the nodes in the classification tree would be visited, and this is not intended. We only want parents, grandparents, etc., and the ‘subtree’ of a classification node to be visited.

For more details, see Hacking and Janssen-Jansen (2009). For another application, see Berger et al. (2004). For a discussion of the use of semantic networks in coding, see Willenborg (2012).

3. Preparatory phase

In order to start coding with the approach as described in the previous section, one needs a so-called “informative base” (see “Coding – How to Build the Informative Base”). Such a base must be constructed “by hand” by experts of the classification. It is a process of trial-and-error: changes to the network may enhance the accuracy of some codes and, at the same time, decrease the accuracy of others. In order to keep the overall accuracy sufficiently large, one must use a test set to detect the changes in coding after changes in the informative base:

- descriptions that get coded correctly due to alterations or additions;
- descriptions that are no longer coded correctly.

These changes (especially the latter one) serve as a good feedback.

On the internet a number of general semantic networks can be found, such as “WordNet” or “OpenCyc”. These networks contain many concepts and relationships as described earlier. When using such network as a basis much work still remains as most classifications are rather domain-specific compared to these general networks.

4. Examples – not tool specific

5. Examples – tool specific

The Spreading Activation method has been applied to the classification of the SBI code. The provisional results are as follows: 80% correctly coded codes, and in 15% of the cases, multiple codes. For more details, see Hacking and Janssen-Jansen (2009). Some practical numbers: the total number of nodes was approximately 5200, the number of relationships was approximately 17200, and the search time in the implementation was around 0.23 sec on a 1.5 GHz machine.

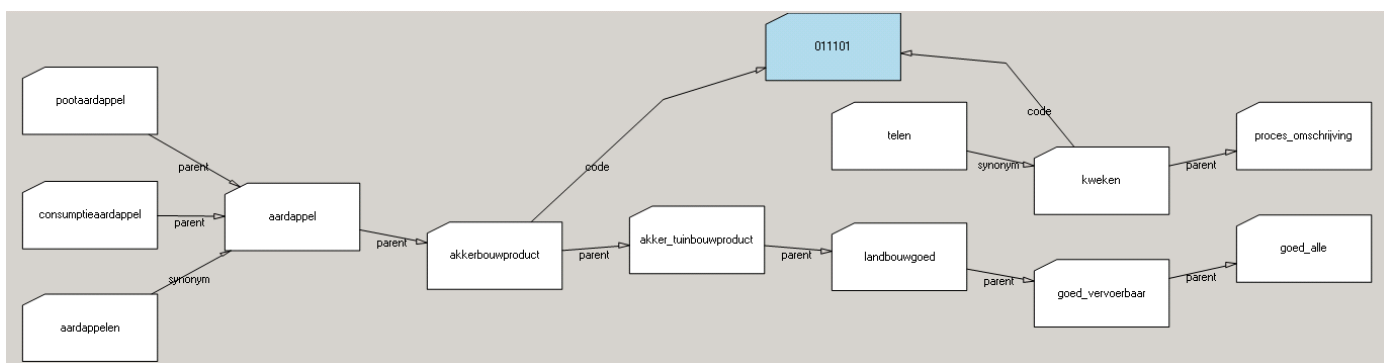


Figure 2. A screenshot that shows a small part of the semantic network (in Dutch) that was visited after the search string ‘telen van aardappelen’(‘cultivation of potatoes’) was provided to the spreading activation algorithm.

In Figure 2, a screenshot is shown of the proof-of-concept that was used for the coding of SBI. This shows a part of the semantic network that is ‘visited’ after providing the search string ‘telen van aardappelen’ (‘cultivation of potatoes’) Note that ‘aardappelen’ (‘potatoes’) (via the classification) leads to ‘akkerbouwproduct’ (‘agricultural product’); combined with ‘telen’ (‘cultivation’) this leads to code 011101 having a score of 2; all other codes (not shown) that were visited received a score of 1.

The software described here implementing spreading activation (currently used for the coding of economic activity) can be used for other kinds of classifications, by creating a different semantic network files. Also, by translating the terms in the semantic network files into another language, one could have a system for other national statistical institutes⁶. The spreading activation program is currently being upgraded to a web version and we intend to offer a web interface or web page to various parties that need to code economic activity, especially the chambers of commerce.

For the ACTR and the Cascot tool, see Wenzowski (1988) and Cascot, respectively.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Berger, H., Dittenbach, M., and Merkl, D. (2004), An accommodation recommender system based on associative networks. In: Frew, A. J. (ed.), *Proceedings of the 11th International Conference on Information Technologies in Tourism (ENTER 2004), Cairo, Egypt, January 26-28, 2004*, Springer-Verlag, 216–227.
- Cascot (a program to semi-automatically classify descriptions):
www2.warwick.ac.uk/fac/soc/ier/software/cascot/.
- D’Orazio, M. and Macchia, S. (2002), A system to monitor the quality of automated coding of textual answers to open questions. *RESEARCH IN OFFICIAL STATISTICS (ROS)*, N.2 2002.
- Hacking, W. J. G. and Janssen-Jansen, S. (2009), The coding of economic activity based on spreading activation. Report, Statistics Netherlands, Heerlen.
- Hacking, W. and Willenborg, L. (2012), *Coding – interpreting short descriptions using a classification*. Contribution to the CBS Methods Series, Statistics Netherlands, The Hague and Heerlen.
- Willenborg, L. C. R. J. (2012), Semantic networks for automatic coding. Report, Statistics Netherlands, The Hague.
- Wenzowski, M. J. (1988), ACTR – A Generalised Automated Coding System. *Survey Methodology* **14**, 299–308.

⁶ There is a complication however: national versions of the NACE classification are allowed to add a 5th digit to some NACE codes; this country-specific part needs to be redone.

Specific section

8. Purpose of the method

Automatic coding takes place just after the data have been collected, in most cases data from interviews. These interviews contain a few fields that are input for the coding step, e.g., “production of wooden crates” serves as input for the coding of economic activity. For simple classifications, e.g., “nation of birth”, a closed question can suffice in the interview; for more complex classifications such as education or occupation, a closed question will probably lead to long lists of possible code descriptions and the quality of the response will decrease rapidly. For that reason, it is often more logical to use an open question in the interview and code this answer at the statistical office; also, for reasons of efficiency, it is better to start coding automatically followed by a manual or a computer-assisted coding step (texts not coded automatically can be analysed by expert coders manually or with the computer support).

The method described here can be used to do the automatic coding step.

9. Recommended use of the method

1. Recommendations on the use of the different methods for coding (automatic or assisted) have been given in the module “Coding – Different Coding Strategies”: the decision about which is the most suitable coding approach to be adopted in a survey depends on different correlated factors. If it has been decided to use automatic coding, one needs a training set of coded descriptions that is available in electronic form, and a correct code (after verification) that is assigned to each description.

10. Possible disadvantages of the method

1. The initial construction of the information base requires a lot of work.

11. Variants of the method

- 1.

12. Input data

1. During the coding phase, the input is quite simple: a textual description, in most cases no more than 10 words. During the construction of the “coding machine” the input consists of the expertise from the classification experts.

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 - 1.
3. Other quality related preconditions

- 1.
4. Other types of preconditions
 1. The input text to be coded should not be too large; in general, this will result in many possible codes for this input text by the method.
- 14. Tuning parameters**
 1. In general, all practical automatic coding algorithms need a *score cut-off value*, to make a selection which descriptions are coded and which need manual or assisted coding. In our experience this parameter is rather robust.
- 15. Recommended use of the individual variants of the method**
 - 1.
- 16. Output data**
 1. Per description the following is derived by the method:
 - a score;
 - a classification code.
- 17. Properties of the output data**
 - 1.
- 18. Unit of input data suitable for the method**

Incremental processing
- 19. User interaction - not tool specific**
 1. None
- 20. Logging indicators**
 1. A number of things may be logged during coding operations: for each coded text all (intermediate) results can be stored for further analysis. This logging can be used when analysing the coding results for a given test set; for each text that was coded correctly before and incorrectly coded now, one can look at logging associated with that text.
- 21. Quality indicators of the output data**
 1. The indicators described in the module “Coding – Measuring Coding Quality” (coding rate and precision rate) can be used to quantify the quality of the method. Quality testing is a continuous process when using semantic networks. During the development of the network, each alteration (or addition or removal) meant to enhance the classification towards code A may deteriorate the classification towards code B. For that reason one needs to check and record the codes assigned by the network based on an incorrectly coded test set. By comparing the assigned codes before and after changes, one can assess the good the change was.

22. Actual use of the method

1. This semantic network method has been used since 2006 until now at the Dutch Chambers of Commerce (in collaboration with Statistics Netherlands) for the coding of economic activity. The ACTR tool has been used by Statistics Canada (who made it; Wenzowski, 1988) and IStat (D’Orazio and Macchia, 2002). Cascot is used by ONS; Statistics Netherlands currently uses it for the coding of occupation.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Coding – Main Module
2. Coding – How to Build the Informative Base
3. Coding – Different Coding Strategies
4. Coding – Measuring Coding Quality

24. Related methods described in other modules

1. Coding – Manual Coding
2. Coding – Automatic Coding Based on Pre-coded Datasets
3. Coding – Computer-Assisted Coding

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. 5.2 Classify and code

27. Tools that implement the method described in this module

1. ACTR (Wenzowski, 1988)
2. Cascot (Cascot)

28. Process step performed by the method

Coding

Administrative section

29. Module code

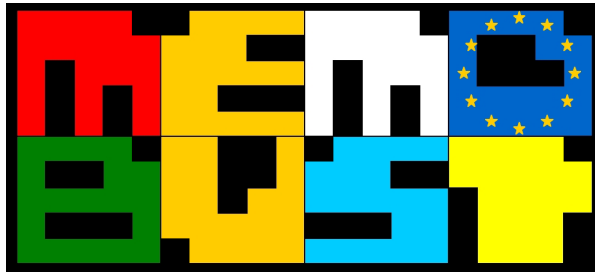
Coding-M-Automatic Coding Based on Semantic Networks

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	02-04-2013	first version	Wim Hacking	CBS
0.2	20-01-2014	following review by Stefania Macchia	Wim Hacking	CBS
0.3	30-01-2014	following review by EB	Wim Hacking	CBS
0.3.1	30-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:06



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Computer-Assisted Coding

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Simple interaction (for the knowledgeable coder)	3
2.2 More interaction (for the less knowledgeable coder)	4
3. Preparatory phase	6
4. Examples – not tool specific.....	6
5. Examples – tool specific.....	6
5.1 Example 1: Interactive coding during CAPI/CATI at Statistics Netherlands	6
5.2 Example 2: More interaction	6
6. Glossary.....	7
7. References	7
Specific section.....	9
Interconnections with other modules.....	11
Administrative section.....	12

General section

1. Summary

First we should define what we mean by computer-assisted coding: it is a situation where a person codes an answer using the computer to search for possible classifications based on some search text. Compared to automatic coding the demands for such a program are less strict: the program may return multiple results, ordered by relevance.

Computer-assisted coding can be used:

1. During the interview if a question arises that requires coding: the coding can be done either by the respondent (e.g., CAWI) or by the interviewer (e.g., CAPI or CATI).
2. After the interview has taken place and some of the variables need to be coded at the statistical office by a coding expert.

Obviously, these situations require different approaches depending on the knowledge of the person that codes the question: if a respondent fills in a coding question, one must assume he has little or no knowledge of the targeted classification. On the other hand, a coding expert trying to code an open answer from the interview just has the information supplied in the open text answer as a basis for coding. Both situations require a different interaction with the computer: an unknowledgeable respondent needs to be taken by the hand to arrive at the classification, whereas the expert needs to be able to formulate a detailed search.

In the following sections we will describe two situations with a different degree of interaction. It will depend on the underlying search system how much interaction is possible.

All the pre-processing steps, such as stop word removal are applicable here as well, but will not be described; for more detail on these steps see Hacking and Willenborg (2012) and Sebastiani (2001).

2. General description of the method

Note that the computer-assisted coding task is less strict compared to automatic coding, where the computer not only has to search for possible codes, but also has to make a decision. That is not necessary in the method described in the present section: it is sufficient if the computer gives the N most probable classifications. The user makes the final choice.

2.1 *Simple interaction (for the knowledgeable coder)*

Performing assisted coding using an informative base constituted only by the classification manual would not be efficient, above all if coding is made directly by the respondent who has not the knowledge of the classification to be used. It would be better to build an informative base, integrated by pre-coded descriptions and/or other materials like synonymous, hypernyms, hyponyms (Macchia and Murgia, 2002)

When used interactively the search program only needs to supply a number of codes plus scores given a text from the user. The descriptions corresponding to these codes are shown in a list, sorted by score in descending order. As a search program, any of the automatic coding programs mentioned in the module “Coding – Automatic Coding Based on Pre-coded Datasets” can be used.

Interaction with the user

In this situation, the automatic coding program only supplies a list of possible codes; especially when this list is large (e.g., due to a vague description), a respondent may be overwhelmed. A possible solution might be to show only the list if the number of items is less than N items. If larger, the computer may ask for a rephrase of the search text. Alternatively, one may cut off the list at N items; but this may be dangerous, because one might throw away the correct code this way.

2.2 More interaction (for the less knowledgeable coder)

By its nature, this method is suitable to be used during the interview in both cases: if the person who codes is the respondent (self-interviewing) and if the person who codes is the interviewer (CATI/CAPI).

If more interaction is needed, the search program must be able to pose additional questions in case of a vague or ambiguous text. The automatic coding programs mentioned in the module “Coding – Automatic Coding Based on Pre-coded Datasets” could be altered to accommodate this. However, to our knowledge, this has not been done yet, except for the automatic coding program using “spreading activation”. We will describe this in more detail later in this section. Blaise¹ (Blaise) also offers more interaction than just a list: it can also show the classification tree or part of it corresponding to the result of the search. We will now describe an extension of the “spreading activation” method, allowing the program to pose further question in case of vague or ambiguous answers.

Detailed description

The semantic networks described in “Coding – Automatic Coding Based on Semantic Networks” can serve as the basis for an interactive search technique as well. By assigning a dimension with the associated question text to every word in the network, we can use the network for interactive questioning. To illustrate: for the coding of ‘education’, we can add the dimensions ‘level’, ‘is a teacher training’, ‘subject’, etc.

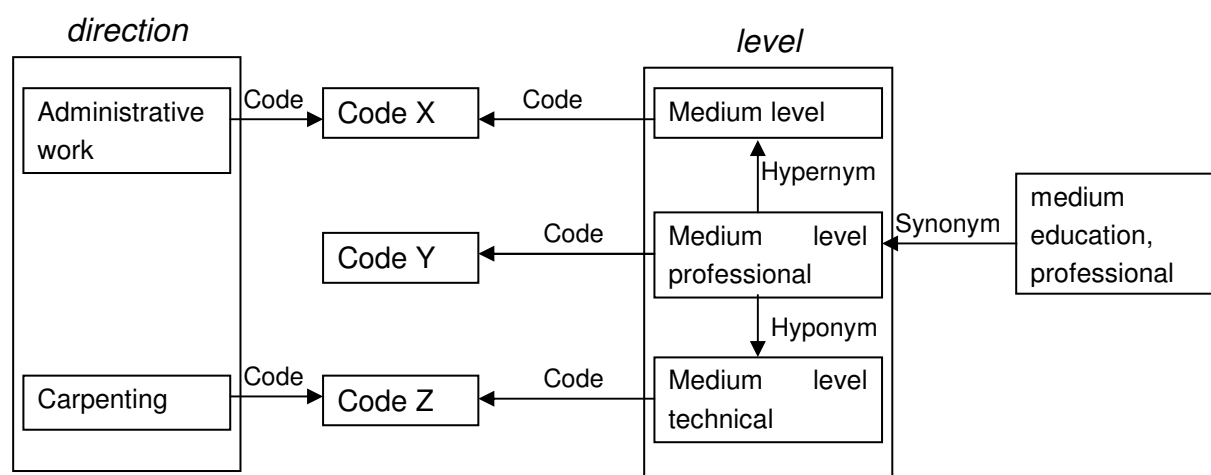


Figure 1: A fragment of the semantic network for the coding of ‘education’ to illustrate the use of a semantic network with dimensions (direction and level).

¹ Blaise is a general package for designing and doing electronic interviews (see www.blaise.com).

The interactive coding process starts with an open question about, for example, education. Based on the answer, which is used as a search string, a number, say N , of codes are selected in the semantic network (see Figure 1).

1. Open question \rightarrow Codes $C = \{C_1, \dots, C_N\}$ with the associated scores $S = \{S_1, \dots, S_N\}$, sorted based on score ($S_{i-1} \geq S_i$); N is the total number of hits. Call the number of codes with the same highest score M .
2. If $M = 0$ (in other words, no suitable codes have been found): either stop or ask for another description.
3. If $1 \leq M \leq_{MAX}$: show the codes found and let the user choose one.
4. If $M \geq M_{MAX}$: select the (next²) dimension D_i and make a list of all words W_k , which are both linked with the dimension D_i and with the codes in C ; now also add the synonyms. Show a question or additional question associated with D_i and let the user make a choice from the list W_k . Each word in this list leads to a sub-selection $S_k \subseteq S$.³
5. The user makes a selection and reduces the set of possible codes: $S := S_k$; now continue with step 2.

Interaction with the user

As described above, semantic networks allow for much more user interaction: this makes it possible to guide a respondent towards a description that increasingly fits any of the classifications:

1. The respondent starts with an open text answer (starting with a closed answer would influence the answer too much, in general)
2. Then, either
 - a. Very little codes apply: let the user make a selection from a list
 - b. Too many codes apply: ask the next closed question (as described above) to try and reduce the number of codes.

Especially when a respondent uses such a system, user-friendliness is very important. There is no general method for this, but many little details contribute to the user-friendliness when constructing a program for computer-assisted coding. Note that this method is much more intended for the untrained user, in contrast with the previous method. To enhance the user-friendliness one may use so called fuzzy string-matching techniques (such as trigrams, Levenshtein, etc.; see Hall and Dowling (1980) and Navarro (2001)); these techniques allow the coding system to recognise incorrectly spelled words, e.g., ‘aple’ instead of ‘apple’.

² This order is predefined.

³ This is therefore a ‘hard’ sub selection. If this is not desired, we can, for example, also reduce the set of records by repeatedly expanding the search string with the selected string from list W_k .

3. Preparatory phase

The preparation for both approaches of computer-assisted coding are almost identical to the preparations for both automatic coding methods. The main difference with automatic coding is the user interaction part, which needs to be added and thought about. Especially when dealing with a system that is intended for use by respondents, user-friendliness is very important to obtain good responses. In the case of computer-assisted coding based on a semantic network, there is also an additional piece of information that needs to be added to the semantic network: how are the words grouped and linked to a dimension and what is the question text associated with that dimension.

4. Examples – not tool specific

5. Examples – tool specific

5.1 Example 1: Interactive coding during CAPI/CATI at Statistics Netherlands

We will now look at an example of the interactive coding of occupations (based on the method described above) as realised at Statistics Netherlands⁴. This coding is performed during the electronic interview process for CAPI and CATI where one of the answers must be coded. To this end, in the Blaise interview, use is made of the option of adding a so called external plugin, which makes it possible to integrate external programs during the interview process. This plugin reads information from the Blaise interview and, on this basis, starts a coding session in which one or more questions are asked to arrive at a classification code. After the coding session, the selected classification code is written back to the Blaise form, and the interviewer or the respondent continues with the interview.

The method as described in section 2.1 is used to offer the respondent several options: sometimes one option in the case of a specific search string, and sometimes several options (e.g., via a drop-down list) in the case of a vague search string.

The interested user is referred to Michiels and Hacking (2004) for more information on the implementation details.

5.2 Example 2: More interaction

As described in section 2.2, coding based on a semantic network offers more possibilities to construct a computer program that interacts with the user.

To illustrate this, Table 1 shows an extract from the semantic network for education, in the form of a search table, to emphasise the link between the words and their associated dimensions (columns) and codes (rows). The extract is based on the initial answer 'English'.

⁴ In addition, a comparable module was developed for the coding of business activities (see Hacking et al., 2009).

Table 1. A small extraction from the table for the coding of education

<i>C</i>	<i>Level</i>	<i>Subject</i>	<i>IsTeacher- Training</i>	<i>University</i>	<i>TeacherType</i>
1	Senior secondary vocational education (MBO)	English	No		
2	Higher professional education (HBO)	English	No		
3	University	English literature	No	Master's	
4	Higher professional education (HBO)	Interpreter English	No		
5	Higher professional education (HBO)	Translator English	No		
6	University	English	Yes	Master's	First level teaching qualification
7	Higher professional education (HBO)	English	Yes		Second level teaching qualification

To further reduce the number of possible codes, we select the dimension 'IsTeacherTraining' with the associated question 'Is this a teacher training?' and answer set $W_k = \{yes, no\}$ with $S_{yes} = \{6, 7\}$ and $S_{no} = \{1, 2, 3, 4, 5\}$. If, for example, we had chosen 'level', then the question would have been 'What is the level of the education?', $W_k = \{HBO, university, academic\}$ (*academic* is a synonym of *university* in the network) and $S_{hbo} = \{2, 4, 5, 7\}$, $S_{university} = \{3, 6\} = S_{academic}$.

This continued questioning technique is currently used for both the coding of education and the coding of the economic activity⁵.

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Blaise, www.blaise.com.

Hacking, W. J. G. and Janssen-Jansen, S. (2009), The coding of economic activity based on spreading activation. Report, Statistics Netherlands, Heerlen.

⁵ For the coding of economic activity, the subselection is slightly more subtle: instead of a hard subselection, there is a repeated search action based on an increasingly expanding search string.

- Hacking, W. J. G., Michiels, J., and Janssen-Jansen, S. (2006), Computer assisted coding by Interviewers. IBUC2006.
- Hacking, W. and Willenborg, L. (2012), *Coding – interpreting short descriptions using a classification*. Contribution to the CBS Methods Series, Statistics Netherlands, The Hague and Heerlen.
- Hall, P. V. and Dowling, G. R. (1980), Approximate string matching. *Computing Surveys* **12**, 381–402.
- Joachims, T. (2002), *Learning to classify text using support vector machines*. Kluwer.
- Macchia, S. and Murgia, M. (2002), Coding of textual responses: various issues on automated coding and computer assisted coding. Journée d’Analyse des Données Textuelles JADT, Saint Malo.
- Michiels, J. and Hacking, W. (2004), Computer assisted coding by interviewers. European Conference on Quality and Methodology in Official Statistics, Mainz, Germany.
- Navarro, G. (2001), A guided tour to approximate string matching. *ACM Computing Surveys* **33**, 31–88.
- Sebastiani, F. (2001), Machine learning in automated text categorization. *ACM Computing Surveys* **34**, 1–47.

Specific section

8. Purpose of the method

The purpose of computer-assisted coding is to guide a person when trying to classify an open text answer.

9. Recommended use of the method

1. Recommendations on the use of the different methods for coding (automatic or assisted) have been given in the module “Coding – Different Coding Strategies”: the decision about which is the most suitable coding approach to be adopted in a survey depends on different correlated factors.

There are two possible situations when CAC (computer-assisted coding) is useful:

- a. During an electronic interview: the method/program will facilitate the interviewer (CAPI,CATI) or the respondent (CASI) to arrive at a code corresponding to the description of the initial text.
- b. After all data have been collected, at the statistical office: coding experts can use the method/program to code the texts more quickly.

10. Possible disadvantages of the method

1. In some cases the computer-assisted coding method may show codes not suitable to the context, leading to incorrect codes. This is particularly so, when coding takes place by someone who does not know enough about the classification, e.g., a respondent. Another source of problems with selecting a description from a list, is that respondents or interviewers often select the first answer, without looking at or scrolling through all possible descriptions.

11. Variants of the method

- 1.

12. Input data

1. During the coding phase, the input is quite simple: a textual description, in most cases no more than 10 words.
2. During the construction of the “coding machine” the inputs can be:
 - pre-coded datasets and lists of different kinds of synonymous (hypernyms, hyponyms) that are used to train the coding algorithm;
 - the knowledge from experts to construct the semantic network.

13. Logical preconditions

1. Missing values

- 1.

2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. The input text to be coded should not be too large; in general, this will result in many classifications by the method. This can be understood, since most methods do take word order into consideration and many different subsets from a large description may fit many classifications.

14. Tuning parameters

1. Often each result returned by the method has a score and the descriptions of the resulting codes are shown by score in descending order. Sometimes there are many low score codes at the end of this list that are probably not relevant. For that purpose, there is a score threshold value T: only codes with score > T are shown.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. For each input description, the method returns a set of codes, each with a score.

17. Properties of the output data

- 1.

18. Unit of input data suitable for the method

Incremental processing.

19. User interaction - not tool specific

1. This has been described in more detail above.

20. Logging indicators

- 1.

21. Quality indicators of the output data

1. The quality indicators have been described in the module “Coding – Measuring Coding Quality”:
 - Coding rate (efficacy) → percentage of coded texts on the total of texts to be coded.
 - Precision rate (accuracy) → percentage correct coded texts on the total of coded texts.

The verification of coding can be performed by a (different) team of coders on a sample of texts. If the original code and the verification code differ, the ‘correct’ code can be decided by expert coders by a reconciliation process.

22. Actual use of the method

1. The methods described above are used at Statistics Netherlands.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Coding – Main Module
2. Coding – Different Coding Strategies
3. Coding – Measuring Coding Quality

24. Related methods described in other modules

1. Coding – Manual Coding
2. Coding – Automatic Coding Based on Pre-coded Datasets
3. Coding – Automatic Coding Based on Semantic Networks

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

1. 5.2 Classify and code

27. Tools that implement the method described in this module

- 1.

28. Process step performed by the method

Coding

Administrative section

29. Module code

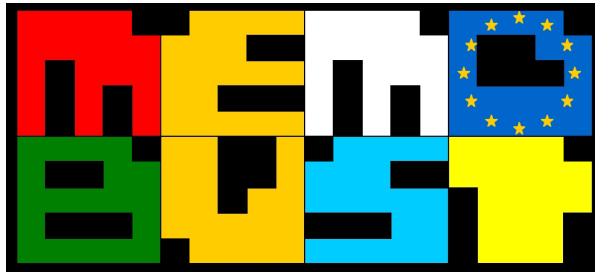
Coding-M-Computer-Assisted Coding

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	02-04-2013	first version	Wim Hacking	CBS
0.2	20-01-2014	following review by Stefania Macchia	Wim Hacking	CBS
0.3	30-01-2014	following review from EB	Wim Hacking	CBS
0.3.1	30-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:07



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Different Coding Strategies

Contents

General section	3
1. Summary	3
2. General description.....	3
2.1 Coding phase during data collection	5
2.2 Coding phase after data collection.....	6
3. Design issues	6
4. Available software tools.....	6
5. Decision tree of methods	6
6. Glossary.....	7
7. References	7
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

Coding of textual responses of statistical surveys, if not made completely manually, can be done in a completely automated way (“automated coding” or batch coding – AUC) or with computer support (“computer-assisted coding” or interactive coding – CAC). The decision which is the most suitable coding approach to be adopted in a survey depends on four correlated factors: the survey technique, the amount of data to be coded, the interview length and the structure of the classification. The combination of these factors can be analysed in two alternative situations, deriving from the moment of the implementation of the coding activity: coding phase during data collection (possible only for CAC) or coding phase after data collection. Elements to define a strategy are provided.

2. General description

Generally speaking, the coding activity, if not made completely manually, can be performed according to two coding procedures (Lyberg and Dean, 1992), using computers in two possible ways:

1. “automated coding” or batch coding (AUC);
 2. “computer-assisted coding” or interactive coding (CAC).
-
1. (AUC). The computer assigns codes to the verbal responses working in ‘batch’ processing. As this technique cannot be expected to assign a code to all the input statements, a manual coding or an assisted coding procedure is required after this step to assign codes to the non-coded responses.
 2. (CAC). The operator assigns codes working interactively with the computer, supporting him in ‘navigating’ the dictionary while searching for codes to be assigned to the input descriptions. For example, when the operator fills in the verbal response on the PC, the machine will show him all dictionary descriptions that could match the input statement (only one description is shown if an exact match exists); the operator should choose one of them, assigning the most suitable code. Thus a CAC system combines the human mind with the computer potential.

The difference between the two procedures lies in their final aim and coding approach. The final aim of AUC procedure is to maximise the number of unique codes assigned automatically to the input statements, whereas the CAC aims at providing the operator with as much assistance as possible. As a consequence, the coding approach of the two systems is different:

- AUC aims at extracting a single description from the dictionary matching the input statement;
- CAC shows different descriptions (also slightly different from each other); it is important to remember that the operator works interactively with the PC and can navigate through the descriptions shown, choosing the most suitable one. Besides, CAC allows the usage of other survey information to support the assignment of codes.

These two procedures allow to manage the coding activity at two different moments of the data collection phase:

- AUC can be used after the interview, that is, when data collection is over;

- CAC can be used both after the interview (by coders) or during the interview (by the interviewer or by the respondent).

The decision which is the most suitable coding approach to be adopted in a survey depends on different correlated factors (Macchia and Murgia, 2002) that is:

1. the survey technique:
 - computer-assisted with the interviewer (CATI – *Computer Assisted Telephone Interviewing*, CAPI – *Computer Assisted Personal Interviewing*);
 - computer-assisted without the interviewer (CASI – *Computer Assisted Self-Interviewing*);
 - traditional Paper and Pencil Technique (PAPI);
2. the amount of data to be coded (Appel and Hellerman, 1983):
 - a large number (e.g., like a census);
 - a small number (like sample surveys on a few thousands of units);
3. the interview length in terms of time necessary to fill in the questionnaire:
 - short interview (less than 15 minutes);
 - long interview (more than 15 minutes);
4. the structure of the classification in conjunction with the variability of the verbal responses:
 - simple classification structure;
 - complex classification structure and high variability of verbal responses.

The structure of a classification can be represented as a tree with branches, sub-branches and leaves. Branches represent general levels of classification that are hierarchically higher than sub-branches and leaves, that represent detailed levels of classification. Therefore, a simple classification structure means a tree with branches, none or few sub-branches and no leaves, whereas a complex structure corresponds with a tree with all these components. Examples of a simple and a complex classification structure are the “*Country Classification*” and the “*Classification of economics activities*”, respectively.

Table 1: Example of classifications with different levels of complexity

Simple Classification	Complex Classification
Country	Economic activities
1. France	01. Crop and animal production, hunting and related service activities
2. Germany	01.1 Growing of non-perennial crops
3. Great Britain	01.11 Growing of cereals (except rice), leguminous crops and oil seeds
4. Italy	01.12 Growing of rice
5. Spain	01.13 Growing of vegetables and melons, roots and tubers

Combining the above-mentioned factors, it is possible to see whether one procedure is more suitable than the other. This combination can be analysed in two alternative situations, based on the moment of the implementation of the coding activity:

1. coding phase during data collection;
2. coding phase after data collection.

2.1 Coding phase during data collection

The following table shows which is the most appropriate coding solution to adopt when computer data capturing is performed by an interviewer.

Table 2: Survey technique: computer-assisted with the interviewer (CATI, CAPI)

Classification structure	Interview length	
	Short	Long
• Simple	CAC	CAC
• Complex & high response variability	CAC	No data coding (coding after data collection)

In general, as can be seen, it is advisable to use CAC during the interview with the interviewer because:

- coded data are available for processing as soon as data collection is over;
- a higher quality of the coded data is also guaranteed by the contact with the respondent who can provide the interviewer with further explanations on the given answer, if needed;
- the previous point implies that, during this activity, the interviewer will ‘train himself’ in getting an answer with sufficient information to be coded.

But, if the interview is long and the coding activity during the interview would increase its duration, it is better not to use CAC and code the data at the end of data collection (even more so if the classifications are complex). In this way the following can be avoided:

- too large a number of uncompleted interviews – respondents deny their co-operation to the operator;
- errors in coding, due to the interviewer’s need to speed up the interview.

As shown in table 3, the situation is different when a computer-assisted technique *without interviewer* is adopted for data capturing.

Table 3: Survey technique: computer-assisted without the interviewer (CASI)

Classification structure	
• Simple	CAC
• Complex & high response variability	No data coding (coding after data collection)

In this case, the coding activity chosen during the interview – done by the respondent himself, not being an expert of the classification – strictly depends on the classification structure. It is advisable to use CAC only if:

- the classification structure is simple;
- the codes to be assigned belong to only one branch of the classification, that is to a high hierarchical level.

2.2 Coding phase after data collection

Whatever technique is used to collect data (CATI, CAPI, CASI, or PAPI), when they are stored in a database, the amount of data to be coded plays a fundamental role in deciding which coding procedure can be adopted:

- for a large amount of data it is advisable to use AUC and subsequently CAC for the non-coded cases;
- for a small amount of data and simple classification it is better to apply AUC;
- for a small amount of data, complex classification and high response variability it is more convenient to adopt CAC.

The following table summarises what was stated before.

Table 4: Coding activity after data collection

Classification structure	Quantity of data/statements to be coded	
	Large number	Small number
• Simple	AUC + CAC	AUC
• Complex & high response variability	AUC + CAC	CAC

3. Design issues

4. Available software tools

Different tools have been developed by statistical offices to be used to code their survey data. The two mentioned here are completely generalised, meaning that they do not depend neither on the language used nor on the classification:

- for automatic coding – ACTR from Statistics Canada (Wenzowski, 1988), recently replaced by GCode;
- for computer-assisted coding – Blaise from CBS for interactive coding.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Appel, M. and Hellerman, E. (1983), Census Bureau Experience with Automated Industry and Occupation Coding. *Proceedings of Section on Survey Research Methods*, American Statistical Association, 32–40.

BLAISE for Windows 4.5 Developer’s Guide (2002).

Lyberg, L. and Dean, P. (1992), Automated Coding of Survey Responses: an international review. Conference of European Statisticians, Work session on Statistical Data Editing, Washington DC.

Macchia, S. and Murgia, M. (2002), Coding of textual responses: various issues on automated coding and computer assisted coding. *Journée d’Analyse des Données Textuelles JADT*, Saint Malo.

Wenzowski, M. J. (1988), ACTR – A Generalised Automated Coding System. *Survey Methodology* **14**, 299–308.

Interconnections with other modules

8. Related themes described in other modules

1.

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM sub-process 5.2

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

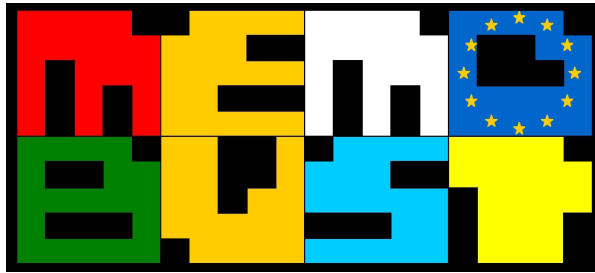
Coding-T-Different Coding Strategies

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	20-07-2012	first version	Stefania Macchia	Istat (Italy)
0.2	21-11-2012	second version (following first revision)	Stefania Macchia	Istat (Italy)
0.3	25-10-2013	third version (following EB review 04-10-2013)	Stefania Macchia	Istat (Italy)
0.3.1	28-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:07



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Measuring Coding Quality

Contents

General section.....	3
1. Summary	3
2. General description.....	3
3. Design issues	4
4. Available software tools	4
5. Decision tree of methods	4
6. Glossary.....	4
7. References	4
Interconnections with other modules.....	5
Administrative section.....	6

General section

1. Summary

Two indicators are usually adopted to measure quality of coding: coding rate (percentage of coded texts on the total of texts to be coded) and precision rate (percentage correct coded texts on the total of coded texts). The quality analysis is usually based on coding texts multiple times and reconciling different codes assigned to the same texts, usually based on a sample of the coded descriptions. The expected values of these rates are different depending on some factors like the complexity of the classification and the detail level of the codes to be assigned.

2. General description

The quality of coding can be measured with two indicators:

- Coding rate (efficacy) → percentage of coded texts on the total of texts to be coded;
- Precision rate (accuracy) → percentage correct coded texts on the total of coded texts.

These rates are suitable either if coding is made automatically (both AUC or CAC) or manually.

The results of the analysis of coding quality requires different approaches depending on which of these two is selected: in the first case, when the results do not fulfil the expectations, the software application and/or the informative base must be updated, while in the second one the further training of interviewers/coders can be necessary.

The quality analysis is usually based on the verification of the coding, which means coding again texts and reconciling different codes assigned to the same texts. Naturally:

- if automatic coding was used, texts will be coded again by human coders (manually or with assisted coding);
- if texts were coded by human coders (manually or with assisted coding) they will be coded again automatically or with the intervention of different coders.

For the precision rate, it is assumed that when the original code and the verification code are equal, the code is correct, otherwise the reconciliation process must be performed by a different expert coder.

Concerning the expected values of these rates, different factors must be considered such as the complexity of the classification and the detail level of the codes to be assigned (Macchia and Murgia, 2002). On the other hand, it also has been noticed that different types of respondents, using the same classification, can have an impact on the final coding rate. For instance, in the experience of Istat, the coding rate of the economic activity responses has always been higher in business surveys than in households or individuals. This is due to the fact that the concept of economic activity is closer to respondents of the first type of surveys than to the latter one; as a result, less precise responses are given (Colasanti et al., 2009).

Finally, the quality analysis is usually conducted on a sample of texts. The sample for verification of coding can be selected in different ways.

Statistics Sweden, for instance, conducts the verification process for at least five percent of the coded records (this threshold of five percent is not statistically motivated, but a requirement for fulfilment of ISO 20252) (Svensson, 2012).

This quality control is made in each relevant survey for data coded through a computer-assisted manual procedure, while once every three years for data coded through an automatic coding procedure.

In Istat a different method is used for the verification of automated coding results, when the amount of processed texts is big: a sample of 'different' texts is checked (D'Orazio and Macchia, 2002). In practice, in order to avoid analysing more than once the same texts, "different" texts are identified through a kind of "raw normalisation", so to delete from descriptions the articles, the conjunctions, the prepositions and the suffixes (in practice all the elements that determine the gender of words, the singular/plural, etc.). Then the occurrence of 'equal' texts is calculated and classes of occurrences are defined (texts are considered 'equal' after a process of raw normalisation). Then texts are stratified according to their frequency of occurrence; then, within each stratum, a simple random sample (without replacement) of texts is selected. The strata coincide with the previously defined classes of occurrences. This implies that the sample contains only different texts, but each of them has a different weight according to its class of occurrence. In this way the work of expert coders is reduced because they will never analyse more than once a text, which, on the other hand, could correspond to a certain number of collected responses.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Colasanti, C., Macchia, S., and Vicari, P. (2009), The automatic coding of Economic Activities descriptions for Web users. NTTS 2009.

D'Orazio, M. and Macchia, S. (2002), A system to monitor the quality of automated coding of textual answers to open questions. *RESEARCH IN OFFICIAL STATISTICS (ROS)*, N.2 2002.

Macchia, S. and Murgia, M. (2002), Coding of textual responses: various issues on automated coding and computer assisted coding. Journée d'Analyse des Données Textuelles JADT, Saint Malo.

Svensson, J. (2012), Quality control of coding of survey responses in Statistics Sweden. European Conference on Quality in Official Statistics Q2012.

Interconnections with other modules

8. Related themes described in other modules

1.

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM sub-process 5.2

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

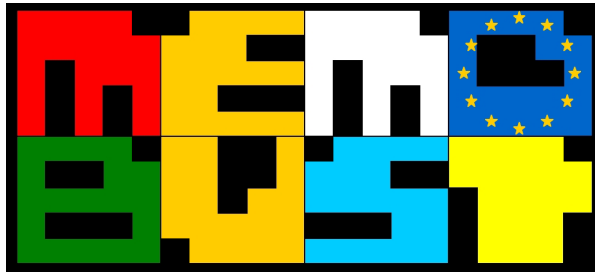
Coding-T-Measuring Coding Quality

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	20-07-2012	first version	Stefania Macchia	Istat (Italy)
0.2	21-11-2012	second version (following first revision)	Stefania Macchia	Istat (Italy)
0.3	25-10-2013	third version (following EB review 04-10-2013)	Stefania Macchia	Istat (Italy)
0.3.1	29-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:08



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Data Editing – Main Module

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction to statistical data editing	3
2.2 Types of errors.....	5
2.3 Edit rules.....	6
2.4 Overview of methods for statistical data editing	7
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

Data that have been collected by a statistical institute inevitably contain errors. In order to produce statistical output of sufficient quality, it is important to detect and treat these errors, at least insofar as they have an appreciable influence on publication figures. For this reason, statistical institutes carry out an extensive process of checking the data and performing amendments. This process of improving the data quality for statistical purposes, by detecting and treating errors, is referred to as statistical data editing.

2. General description

2.1 Introduction to statistical data editing

Errors are virtually always present in the data files used by producers of statistics. This is true for both data obtained by means of surveys and data originating from external registers. Insofar as these errors result in inaccurate estimates of publication figures, it is important for statistical institutes to detect and treat these errors.

Errors can arise during the measurement process; if this is the case, there will be a difference between the reported value and the actual value. This can occur because the respondent does not know the actual value exactly or at all, or has difficulty finding this value and therefore makes an estimate. Another possible cause is a difference in definitions between the accounting records of businesses and the statistical institute, for example because the financial year differs from the calendar year. Furthermore, it is possible that businesses simply do not have all the information requested by the statistical institute on file. In this case, the respondent will again estimate certain values or not answer all questions. Finally, respondents may also read or understand questions incorrectly. For example, they may report in euros, while they were actually asked to report in thousands of euros (this is an example of a so-called *unit of measurement error*).

Errors may also arise during data processing. At a statistical institute, the collected data typically go through different processes, such as entering, coding, detection, imputation, weighting, and tabulation. All of these processes can introduce errors into the data. An example of this is that the manual entry of data can result in misinterpretations, for example, a '1' is taken for a '7' or vice versa. Similar mistakes can occur when optical character recognition is used to process survey forms automatically. Additionally, there may be errors in the processing software, and good values may incorrectly be seen as errors during the editing process.

The process of detecting and treating errors in a data file to be used for statistical purposes is called *statistical data editing*. Other commonly used terms are *data validation* and *data cleaning*. In traditional survey processing, data editing was mainly a manual activity, intended to check and correct all data items in every detail. Inconsistencies in the data were investigated and, if necessary, adjusted by subject-matter experts, who would consult the original questionnaires or recontact respondents to verify suspicious values. Overall, this was a very time-consuming and labour-intensive procedure. According to estimates in the literature, statistical institutes would spend up to 25% or 40% of their total budget on data editing (Federal Committee on Statistical Methodology, 1990; Granquist, 1995; Granquist and Kovar, 1997).

According to Granquist (1997), statistical data editing should have the following objectives, in descending order of priority:

1. To identify possible sources of errors so that the statistical process can be improved in the future;
2. To provide information about the quality of the data collected and published;
3. To detect and correct influential errors in the collected data.

In EDIMBUS (2007), a fourth objective is added:

4. If necessary, to provide complete and consistent microdata.

In line with the first objective mentioned above, the main aim of recontacts with respondents should not be to merely resolve individual observed errors, but rather to collect information on the causes of these errors. By collecting and analysing this information, a statistical institute has the opportunity to identify potential measures for improving the quality of incoming data in the future. Examples of such measures include improving the design of the questionnaire and, in particular, changing the wording of a question that many respondents found difficult to answer. In the words of Granquist (1997), “editing should highlight, not conceal, serious problems in the survey vehicle.”

Currently at most statistical institutes, statistical data editing is used primarily with the third and fourth of the above goals in mind: correcting errors that have a significant influence on publication totals and providing complete and consistent data. Although it is widely acknowledged in the data editing literature that the information obtained during editing could and should also be used to improve aspects of the statistical process for a repeated survey, the development of practices to achieve this goal still appears to be a rather neglected area. Some statistical institutes have had good experiences with standardised debriefings of editing staff as a device for identifying possible improvements in questionnaire design (Rowlands et al., 2002; Hartwig, 2009; Svensson, 2012). An overview of indicators for assessing the quality of the data before and after editing is given in EDIMBUS (2007).

Over the past decades, statistical institutes have recognised that it is usually not necessary to correct all data in every detail. Several studies have shown that reliable estimates of publication totals can also be obtained without removing all errors from a data set (see, e.g., Granquist, 1997, and Granquist and Kovar, 1997). The main output of most statistical processes consists of tables of aggregated data, which are often estimated from a sample of the population. Hence, small errors in individual records can be accepted, provided that (a) these errors mostly cancel out when aggregated, and (b) insofar as they do not cancel out when aggregated, the resulting measurement error in the estimate is small compared to the total error – in particular the natural variation in the estimate due to sampling.

The notion that not all errors need to be corrected in every detail has led to the development of more efficient editing approaches: in particular selective editing, automatic editing and macro-editing. Section 2.4 introduces these approaches, and also illustrates how they may be combined into an effective data editing process. Before that, we discuss different types of errors in Section 2.2 and edit rules in Section 2.3.

We refer to De Waal et al. (2011) and EDIMBUS (2007) for a more comprehensive description of statistical data editing.

2.2 Types of errors

Different editing methods have been developed for different types of errors. We will consider here the distinction between influential and non-influential errors and the distinction between systematic and random errors.

Influential errors include the errors that have a significant influence on the final publication total. An error can be influential because it was made by a business that naturally has a strong influence on the estimate, i.e., either by a large business or by a smaller one with a large sampling weight. In addition, sometimes an error is so large that it will strongly influence the total, regardless of the size of the business for which the error occurred. A notorious example of a type of error that is usually influential is the above-mentioned unit of measurement error.

It is clear that errors that have a large influence on a publication total can lead to significant bias. For this reason, it is crucial to treat these errors as effectively as possible. An efficient and timely data editing process will have to focus mainly on the detection and treatment of influential errors. The distinction between influential and non-influential errors is particularly useful in business surveys, because these often contain variables with a skew distribution in the population, such as *Turnover*.

Another distinction that is often made is that between *systematic* and *random* errors.¹ These terms do not have universally accepted definitions. In particular, UN/ECE (2000) defines a systematic error as “an error reported consistently over time and/or between responding units”, while EDIMBUS (2007) defines it as “a type of error for which the error mechanism and the imputation procedure are known.” The first definition refers in particular to errors that are caused by persistent response problems, which are ‘not random’ in the sense that they would likely be observed again if the data collection process were repeated. Examples include: the unit of measurement error mentioned in Section 2.1; different definitions used by the statistical institute and the respondent (e.g., gross turnover versus net turnover); persistent problems with data entry or coding at the statistical office. The second definition focuses on the fact that, in many cases, errors of this kind are relatively easy to detect, precisely because they are made in a consistent way. Thus, in many cases, these two definitions of systematic errors agree. In practice, the only systematic errors that can be treated as such are those for which the error mechanism is understood, i.e., errors that are systematic according to the definition of EDIMBUS (2007).

Although the above definitions of systematic errors do not mention bias, it does hold that systematic errors often produce a systematic bias in estimated figures. This is true because these errors are often made in the same way by several respondents. For random errors – i.e., errors that are not systematic as defined in the previous paragraph – the risk of a bias is smaller. On the other hand, random errors are more difficult to detect and correct reliably, precisely because little is known about the underlying causes.

It should be noted that systematic errors may or may not be influential. For instance: the unit of measurement error is usually influential, but an error where a small business with a moderate sampling

¹ Here, the terms ‘systematic’ and ‘random’ are supposed to refer to the mechanism that *causes* an error. This differs from the use of these terms in measurement error models, where they refer to the *effect* of an error on an estimator (an error being systematic to the extent that it introduces bias and random to the extent that it introduces noise). As explained in the main text, these two meanings of ‘systematic’ do overlap to some extent.

weight reports gross turnover instead of net turnover will usually be non-influential. The same holds for random errors.

2.3 Edit rules

To detect errors in observed data, *edit rules* are widely used. These are rules that indicate conditions that should be satisfied by the values of single variables or combinations of variables in a record. Edit rules are also commonly known as *edits* or *checking rules*. If a record does not satisfy the condition specified by an edit rule, the edit rule is said to be failed by that record. Inspection of data items that fail an edit rule is an important technique for finding errors in a data file.

A conceptual distinction should be made between so-called *hard* and *soft* edit rules. Hard edit rules (also known as *fatal* edit rules or *logical* edit rules) are edit rules that must hold by definition, such as

$$\text{Turnover} = \text{Profit} + \text{Costs}.$$

If a hard edit rule is failed by an observed combination of values, then it is certain that at least one of those values contains an error. Soft edit rules (also known as *query* edit rules) indicate whether a value, or value combination, is suspicious. For instance, the soft edit rule

$$\text{Profit} / \text{Turnover} \leq 0.6$$

states that it is unusual for the value of *Profit* to be higher than 60% of the value of *Turnover*. In contrast to hard edit rules, soft edit rules can be failed by unlikely values that are in fact correct. Thus, soft edit failures should trigger a closer investigation of the data items involved, to assess whether the suspicious values are erroneous or merely unusual.

Typically, business surveys involve (mainly) numerical data. For this type of data, some commonly encountered classes of edit rules include the following:

- *Univariate edits / Range restrictions.* These edit rules restrict the range of admissible values for a single variable. A common example is the restriction that a numerical variable may attain only non-negative values, e.g., the edit rule “ $\text{Turnover} \geq 0$ ”. Depending on the context, edits of this type can be either hard or soft.
- *Ratio edits.* These edit rules are bivariate restrictions taking the general form $a \leq x / y \leq b$, where x and y are numerical variables and a and b are constants. An example could be that the ratio of *Turnover* and *Number of Employees* (i.e., the average contribution of one employee to the total turnover of a business) should be between certain bounds. The above-mentioned edit rule “ $\text{Profit} / \text{Turnover} \leq 0.6$ ” is another example of a ratio edit. As the latter example illustrates, some ratio edits contain only a lower bound a or an upper bound b , but not both. Typically, ratio edits are soft edit rules.
- *Balance edits.* These edit rules are multivariate restrictions that relate a set of variables through a linear equality. The above-mentioned edit rule “ $\text{Turnover} = \text{Profit} + \text{Costs}$ ” is an example of a balance edit. The general form of a balance edit is: $a_1x_1 + \dots + a_nx_n + b = 0$, where x_1, \dots, x_n are numerical variables and a_1, \dots, a_n, b are constants. Usually, but not always, balance edits are hard edit rules.

2.4 Overview of methods for statistical data editing

The data editing process that is considered here starts after the data have been collected and entered. It should be noted, however, that nowadays many business surveys use computer-assisted modes of data collection (see the topic “Data Collection”) which often involve electronic questionnaires. With computer-assisted data collection, it is possible to perform part of the editing already at the data collection stage, for instance by building certain edit rules into the electronic questionnaire. We refer to the theme module “Questionnaire Design – Editing During Data Collection” for a discussion of the possibilities.

The specific way that the data editing process is structured will vary by statistic and by statistical institute. However, there is a general strategy that is followed in broad lines in many processes. This general strategy is shown in Figure 1; similar strategies are discussed in De Waal et al. (2011, pp. 17-21) and EDIMBUS (2007, pp. 6-8). It consists of five steps:

1. Deductive editing;
2. Selective editing;
3. Automatic editing;
4. Interactive editing (manual editing);
5. Macro-editing.

In the remainder of this section, we give a brief outline of each of these steps. More detailed descriptions can be found in the accompanying modules on methods for statistical data editing.

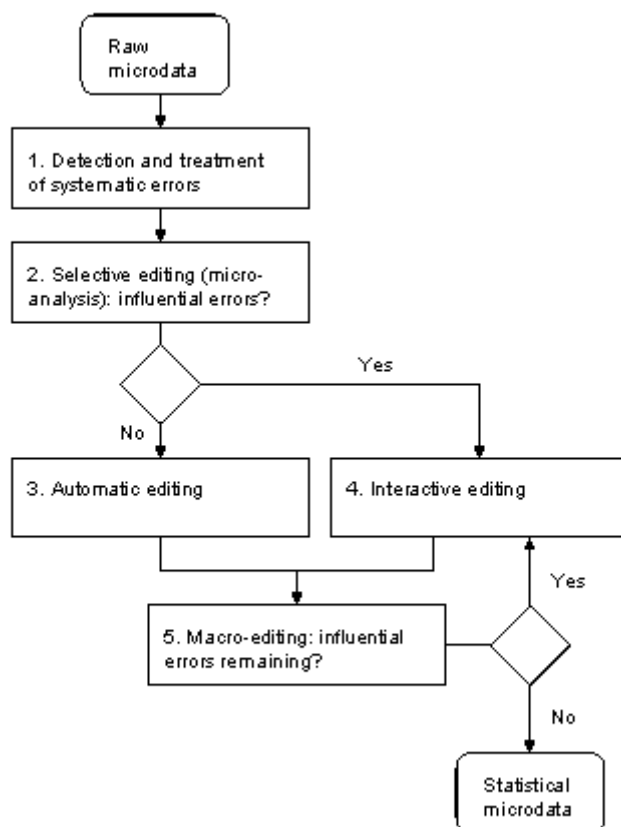


Figure 1. Example of a data editing process flow

In the first phase of the data editing process, identifiable systematic errors are detected and treated. As stated in Section 2.2, these systematic errors can lead to significant bias. Moreover, these errors can often be automatically detected and treated easily and very reliably. It is highly efficient to treat these errors at an early stage. In the remainder of the data editing process, it may then be assumed that the data contain only random errors. The detection and treatment of systematic errors is discussed in the method module “Statistical Data Editing – Deductive Editing”.

After the identifiable systematic errors have been edited automatically, a decision can be taken to begin *manual editing*, i.e., manual detection and treatment of errors. This process step is performed by editors or analysts who are usually supported in this regard by software that allows, for example, edit rules to be applied to the data and values to be changed interactively. This form of editing (also known as *interactive editing*) is described in the method module “Statistical Data Editing – Manual Editing”.

As mentioned above, manual editing is usually expensive and time-consuming. It is therefore better to restrict the manual work only to records that likely contain influential errors, so that the specialists’ limited time can be used where it is most effective. The other records, with less important errors, can either be left unedited or, alternatively, be edited automatically (see below). Limiting interactive editing to those records that likely contain influential errors which cannot be reliably resolved automatically is known as *selective editing* or *micro-selection*. Methods that can be used in this step are discussed in the theme module “Statistical Data Editing – Selective Editing”. It should be noted that the selective editing step by itself does not treat any errors; it merely assigns records to different forms of further treatment.

Most selective editing methods make use of anticipated values for the variables in a record to identify the most suspicious values in the observed data. Observed values that deviate strongly from the anticipated values may be caused by influential errors. In determining the anticipated values, information is used from sources other than the actual data file. Oftentimes, edited data from a previous period for the same statistic is used for this purpose. As such, selective editing can proceed on a record-by-record basis, and hence it is possible to start the selection process for manual editing during the data collection period, as soon as the first records are received. This is in fact the main advantage of selective editing over macro-editing, a different selection method to be discussed below.

Records that are not selected for manual editing can be processed by *automatic editing* instead. The automatic treatment of random errors and other errors for which the cause cannot be established usually takes place in two steps. First, the best possible determination is made of what values in a record are incorrect. This is trivial if a value does not fall in the permissible range according to a univariate edit, such as a negative number of employees or an improperly missing value. As such, the value is then certainly incorrect. In many cases, however, inconsistencies can occur for which it is not immediately clear which value or values are responsible. If, for example, the hard balance edit

$$\text{Total Costs} = \text{Personnel Costs} + \text{Capital Costs} + \text{Transport Costs} + \text{Other Costs}$$

is not satisfied, then it is clear that (at least) one of the reported values must be erroneous, but it is usually not obvious which one. The problem of identifying the erroneous values in an inconsistent record is known as the *error localisation problem*.

In automatic editing of business survey data, the error localisation problem for random errors is usually solved by applying the *Fellegi-Holt paradigm*, which states: a record should be made

consistent by changing the fewest possible items of data (Fellegi and Holt, 1976). Methods for automatic error localisation based on the Fellegi-Holt paradigm are discussed in the method module “Statistical Data Editing – Automatic Editing”.

Once the erroneous values have been detected, they are replaced with better values by means of *imputation*. Automatic imputation relies on (explicit or implicit) mathematical models that use information from the correctly observed values to predict the values that were incorrectly observed or missing. We refer to the topic “Imputation” for a discussion of this subject.

Instead of applying automatic editing, one may also choose not to edit the records that are not selected for interactive treatment by the selective editing procedure. In fact, one may argue that it is not necessary to edit these records, because they will not contain any influential errors, assuming that the selective editing procedure works as intended. Nevertheless, there are reasons why automatic editing may be of use in practice (see also De Waal and Scholtus, 2011). Firstly, it is often desirable to resolve at least all obvious inconsistencies (values that fail hard edit rules), even when these are not influential as such. This is especially true if the microdata are to be released to external users. Secondly, automatic editing provides a relatively inexpensive way to test the quality of a selective editing procedure. If the selection procedure is working correctly, then the records that are not selected for interactive treatment should require only minor adjustments with little influence on a publication figure. Thus, if many influential adjustments are made during automatic editing, this may indicate that the design of the selective editing procedure needs to be improved.

In the final phase of the process in Figure 1, provisional publication figures are calculated and analysed using historical data or external sources. This analysis is called *macro-editing* or *output editing*. If the aggregate figures are implausible, the underlying individual records are examined by, for example, further analysing outliers or influential records and adjusting these as necessary. In Figure 1, this is indicated by the arrow leading back from macro-editing to interactive editing. The errors detected at this stage may be errors that were not found in earlier phases of the data editing process or errors that were actually introduced by the process. In macro-editing, the detection of errors begins at an aggregated level, but the adjustment always takes place in the underlying microdata, i.e., the records of individual respondents. As soon as the provisional figures are considered plausible, the statistical data editing process is completed. For more information on this step, see the module “Statistical Data Editing – Macro-Editing”.

In the macro-editing step, as well as during selective editing and manual editing, mathematical techniques for outlier detection are often applied. An extensive discussion of outlier detection in the context of statistical data editing can be found in EDIMBUS (2007).

The process in Figure 1 should be viewed as a prototype. In practice, not all of the steps will be undertaken for all statistics, or a different order of process steps may be used. For instance, it was already mentioned that automatic editing is not always included in the process. Another example is that the selection of records for manual editing is often partly based on other criteria than only whether a record contains influential errors. As such, important or complex businesses are frequently identified as crucial, meaning that their data are always inspected manually. Examples of such businesses could be those that are individually responsible for a significant portion of turnover in their sector. See, e.g., Pannekoek et al. (2013) for a further discussion of the design of an editing process.

Many business surveys have a longitudinal aspect. Sometimes, a panel of units is followed over time during multiple rounds of the same survey. Even for cross-sectional business surveys, the largest units in the population are usually observed in each survey round. This implies that during a particular survey round, at least for part of the responding units, historical data are available. These historical data may be used in various ways during several steps of the editing process; for example, they are often used to determine anticipated values for selective editing. We refer to the theme module “Statistical Data Editing – Editing for Longitudinal Data” for more details on this aspect of statistical data editing.

Finally, it should be noted that, traditionally, applications of statistical data editing have been aimed mainly at survey data. More recently, the use of administrative data for statistical purposes has become increasingly important. These data require an editing process that is in some respects different from the typical editing process for survey data. For instance, for statistics based on administrative data, often all the data (or a large proportion thereof) become available at the same time. In that case, it is not necessary to use micro-selection methods, and we can start immediately with output editing. We refer to the theme module “Statistical Data Editing – Editing Administrative Data” for a discussion of editing in the context of statistics based on administrative data.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.

De Waal, T. and Scholtus, S. (2011), Methods for Automatic Statistical Data Editing. Paper presented at the 2011 KSS International Conference on Statistics and Probability, Busan.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.

http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Federal Committee on Statistical Methodology (1990), *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18, U.S. Office of Management and Budget, Washington, D.C.

- Fellegi, I. P. and Holt, D. (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.
- Granquist, L. (1995), Improving the Traditional Editing Process. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, 385–401.
- Granquist, L. (1997), The New View on Editing. *International Statistical Review* **65**, 381–387.
- Granquist, L. and Kovar, J. (1997), Editing of Survey Data: How Much Is Enough? In: L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, 415–435.
- Hartwig, P. (2009), How to Use Edit Staff Debriefings in Questionnaire Design. Paper presented at the 2009 European Establishment Statistics Workshop, Stockholm.
- Pannekoek, J., Scholtus, S., and van der Loo, M. (2013), Automated and Manual Data Editing: A View on Process Design and Methodology. *Journal of Official Statistics* **29**, 511–537.
- Rowlands, O., Eldridge, J., and Williams, S. (2002), Expert Review Followed by Interviews with Editing Staff – Effective First Steps in the Testing Process for Business Surveys. Paper presented at the 2002 International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, South Carolina.
- Svensson, J. (2012), Editing Staff Debriefings at Statistics Sweden. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- UN/ECE (2000), *Glossary of Terms on Statistical Data Editing*. United Nations, Geneva.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Editing During Data Collection
2. Data Collection – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Statistical Data Editing – Editing Administrative Data
6. Statistical Data Editing – Editing for Longitudinal Data
7. Imputation – Main Module

9. Methods explicitly referred to in this module

1. Statistical Data Editing – Deductive Editing
2. Statistical Data Editing – Automatic Editing
3. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Statistical data editing

Administrative section

14. Module code

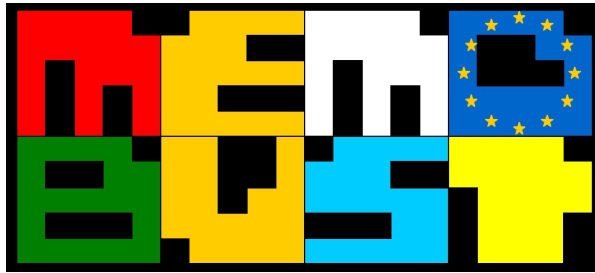
Statistical Data Editing-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	09-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	20-04-2012	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	19-06-2012	minor improvements	Sander Scholtus	CBS (Netherlands)
0.3.1	16-07-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.4	31-10-2013	minor improvements based on comments by Italian reviewer and Editorial Board	Sander Scholtus	CBS (Netherlands)
0.4.1	31-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:10



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Deductive Editing

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Introduction to deductive editing.....	3
2.2 Correction rules for subject-matter related errors.....	4
2.3 The unit of measurement error	5
2.4 Identifying new systematic errors	8
3. Preparatory phase	9
4. Examples – not tool specific.....	9
4.1 Example: Correction rules for the statistic Building Objects in Preparation.....	9
4.2 Example: Simple typing errors.....	10
5. Examples – tool specific.....	11
6. Glossary.....	13
7. References	13
Specific section.....	15
Interconnections with other modules.....	16
Administrative section.....	18

General section

1. Summary

Data collected for compiling statistics frequently contain obvious systematic errors; in other words, errors that are made by multiple respondents in the same, identifiable way (see “Statistical Data Editing – Main Module”). Such a systematic error can often be detected automatically in a simple manner, in particular in comparison to the complex algorithms that are needed for the automatic localisation of random errors (see the method module “Statistical Data Editing – Automatic Editing”). Furthermore, after a systematic error has been detected, it should be immediately clear which adjustment is necessary to resolve it. For we know, or think we know with sufficient reliability, how the error came about.

A separate deductive method is needed for each type of systematic error. The exact form of the deductive method varies per type of error; there is no standard formula. The difficulty with using this method lies mainly in determining *which* systematic errors will be present in the data, before these data are actually collected. This can be studied based on similar data from the past. Sometimes, such an investigation can bring systematic errors to light that have arisen due to a shortcoming in the questionnaire design or a bug in the processing procedure. In that case, the questionnaire and/or the procedure should be adapted. To limit the occurrence of discontinuities in a published time series, it can be desirable to ‘save up’ changes in the questionnaire until a planned redesign of the statistic, and to treat the systematic error with a deductive editing method until that time.

2. General description of the method

2.1 Introduction to deductive editing

In this module, we focus on methods for detecting and treating so-called systematic errors. As mentioned in “Statistical Data Editing – Main Module”, a systematic error is commonly defined as an error with a structural cause that occurs frequently between responding units. A well-known type of systematic error is the so-called *unit of measurement error* which is the error of, for example, reporting financial amounts in units instead of the requested thousands of units.

Systematic errors can introduce substantial bias in aggregates, but once detected, systematic errors can easily be treated because the underlying error mechanism is known. It is precisely this knowledge of the underlying cause that makes the treatment of systematic errors different from random errors. Treating systematic errors based on knowledge of the underlying error mechanism is called *deductive editing*. Systematic errors can often be identified by examining frequently occurring edit rule failures. Deductive methods are therefore mainly effective for data for which many edit rules have been defined.

Deductive editing of systematic errors is an important first step in the editing process. It can be done automatically and reliably at virtually no costs. Moreover, the rest of the editing process can proceed more efficiently after the systematic errors have been resolved. Deductive editing is in fact a very effective and probably often underused editing approach.

Any systematic error for which the cause is understood with sufficient certainty can be resolved deductively. In the case of incorrect assumptions about the error mechanism, however, deductive

editing may introduce a bias in the estimators. In practice, a deductive method might also be used to resolve certain random errors, for reasons of efficiency, provided that the introduced bias is negligible. An example of this is the deductive resolution of rounding errors (see Scholtus, 2011).

De Waal and Scholtus (2011) make a further distinction between *generic* and *subject-related* systematic errors. Errors of the former type occur for a wide variety of variables in a wide variety of surveys and registers, where the underlying cause is always essentially the same. Apart from the unit of measurement error, other examples include *simple typing errors*, such as interchanged or mistyped digits (Scholtus, 2009) and *sign errors*, such as forgotten minus signs or interchanged pairs of revenues and costs (Scholtus, 2011). For an example that involves a simple typing error, see Section 3.2 below. Generic errors can often be detected and treated automatically by using mathematical techniques.

Subject-related systematic errors are specific to a particular questionnaire or survey. They may be caused by a frequent misunderstanding or misinterpretation of some question such as reporting gross values rather than net values. Another example is that, for some branches of industry, staff is frequently classified as belonging to an incorrect department of the responding enterprise. Subject-related systematic errors are usually detected and treated by applying correction rules that have been specified by subject-matter experts.

The remainder of this text is organised as follows. Section 2.2 further discusses the use of correction rules for subject-related systematic errors. Section 2.3 discusses techniques that treat possibly the most notorious of generic systematic errors, the unit of measurement error. Section 2.4 discusses methods for identifying new systematic errors.

2.2 Correction rules for subject-matter related errors

Subject-matter related errors can often be detected and treated by means of deterministic checking rules. Such rules state which variables are to be considered erroneous when the edits are failed in a certain way. Often, deterministic checking rules also describe how the erroneous variables should be adjusted. In that case, these rules are commonly referred to as *correction rules*.

The general form of a correction rule is as follows:

if (*condition*) **then** (*correction*).

Here, *condition* indicates a combination of values in a record that is not allowed. Subsequently *correction* describes the adjustment that is made to the record to resolve the inconsistency.

An example of a correction rule is:

if (*Number of Temporary Employees* > 0 **and** *Costs of Temporary Employees* = 0)
then *Number of Temporary Employees* := 0. (1)

This rule detects an inconsistency that occurs when a business reports to have employed temporary staff without reporting associated costs. In this example, the inconsistency is treated deductively by making the number of temporary employees equal to zero.

In general, a correction rule is intended to resolve an inconsistency that can be resolved in a unique way on logical and/or content-related grounds, under a certain assumption. If the assumption is valid, the deductive editing method always reproduces the true values. For instance, the correction rule (1)

operates under the assumption that the variable *Costs of Temporary Employees* is reported more accurately than the variable *Number of Temporary Employees*. Making such assumptions in a valid way generally requires subject-matter knowledge and knowledge of the data collection process.

Correction rules are attractive because of their simplicity. However, they may only be used when no important nuances are lost with such a simple approach. If the data do not satisfy the assumptions made, then deductive editing may lead to biased estimators. For instance: if in the above example it happens that some businesses actually forget to report the costs of temporary employees, then, after applying the correction rule (1), we may underestimate the total number of temporary employees for businesses in the target population.

Another potential drawback of using correction rules is that a large collection of correction rules may be difficult to maintain, especially when the collection has grown over a long period of time. In particular, it then becomes difficult to grasp the effects of adding a new correction rule, or removing an old one, or changing the order in which the rules are applied to the data. For this reason, it is usually not recommended to try to treat all possible errors in a rule-based manner, because this would require a very complex set of correction rules. Broadly speaking, deductive editing should be limited to the treatment of systematic errors only. For the treatment of random errors, there exist other methods that are more powerful and less difficult to maintain (see “Statistical Data Editing – Automatic Editing”).

2.3 *The unit of measurement error*

Business surveys usually contain instructions to the reporter that all financial amounts must be rounded to thousands of euros (dollars, pounds, etc.), that all quantities must be rounded to thousands of units, et cetera. Some respondents ignore these instructions and, consequently, report values that are a factor 1000 larger than they actually mean. It is clear that, if these *thousand-errors* are not corrected, the resulting estimates for the figures to be published will be too high. The thousand-error is a commonly encountered special case of the more general unit of measurement error, which occurs whenever respondents report values that are consistently too high or too low by a certain factor.

We refer to a *uniform* unit of measurement error if all variables (of a certain type) in a record are too large by the same factor. It is known that, in practice, records with *partial* unit of measurement errors also occur. A partial unit of measurement error could arise, for instance, if several departments of a business each fill in part of a questionnaire independently. Partial unit of measurement errors are generally more difficult to detect than uniform ones.

Traditional methods for detecting unit of measurement errors usually work by comparing one or more reported amounts with reference values. The type of reference data used and the way in which the comparison takes place varies per statistic and per statistical office. Examples of reference data are: a statement from the same respondent from an earlier period, the median value of a number of similar respondents in an earlier period or the same period, and available register data about the respondent.

A widely used method computes the ratio of the unedited value and the reference value. If this ratio is larger than a lower bound, or lies between certain bounds, then it is concluded that the unedited value contains a unit of measurement error. Once a unit of measurement error has been detected, it is treated deductively by dividing all relevant amounts by an appropriate factor. It is often assumed for convenience that all unit of measurement errors are uniform.

For instance: in the Dutch Short Term Statistics, thousand-errors are detected as follows (Hoogland et al., 2011). The total turnover indicated by the respondent for period t , say x_t , is compared to the turnover from the most recent period for which a statement from the respondent is available, up to a maximum of six previous periods. The stated turnover for this earlier period must also not be equal to zero. A thousand-error is detected in x_t if the following applies:

$$|x_t| > 300 \times |x_{t-i}|, \quad \text{for some } i \in \{1, \dots, 6\}.$$

If no data from the respondent from an earlier period are available, then the median of the turnover from the previous period in the stratum of the respondent is used instead. The stratification is based on economic activity and number of employees. A thousand-error is detected in x_t if the following applies:

$$|x_t| > 100 \times \text{stratum median}(x_{t-1}).$$

If a thousand-error is detected by either formula, then it is resolved by dividing the total turnover and all the sub-items by 1000.

Table 1 shows an example of a record with a thousand-error that was found in this way.

Table 1. Example of a uniform thousand-error

	reference data	unedited data	data after treatment
<i>first sub-item turnover</i>	3,331	3,148,249	3,148
<i>second sub-item turnover</i>	709	936,142	936
<i>total turnover</i>	4,040	4,084,391	4,084

It should be noted that the above-described method assumes that the reference value is not affected by unit of measurement errors. Thus, the reference value should either be based on previously edited data, or it should be calculated in a way that is robust to the presence of (some) unit of measurement errors.

Clearly, the choice of bounds in the detection method for unit of measurement errors is important. There is a trade-off here between the number of missed errors (observations that are supposedly correct, but actually contain unit of measurement errors) and the number of false hits (observations that supposedly contain unit of measurement errors, but are actually correct). If previously edited data are available, then a simulation study can be conducted to experiment with different bounds. See Pannekoek and De Waal (2005) for an example of such a simulation study.

In manual editing, unit of measurement errors are often detected using a graphical aid. As an illustration, Figure 1 shows a scatter plot of unedited values of turnover (on the y axis) against reference values (on the x axis), with both variables plotted on a logarithmic scale (using the logarithm to base 10). A cluster of thousand-errors can clearly be identified near the line $y = x + 3$.

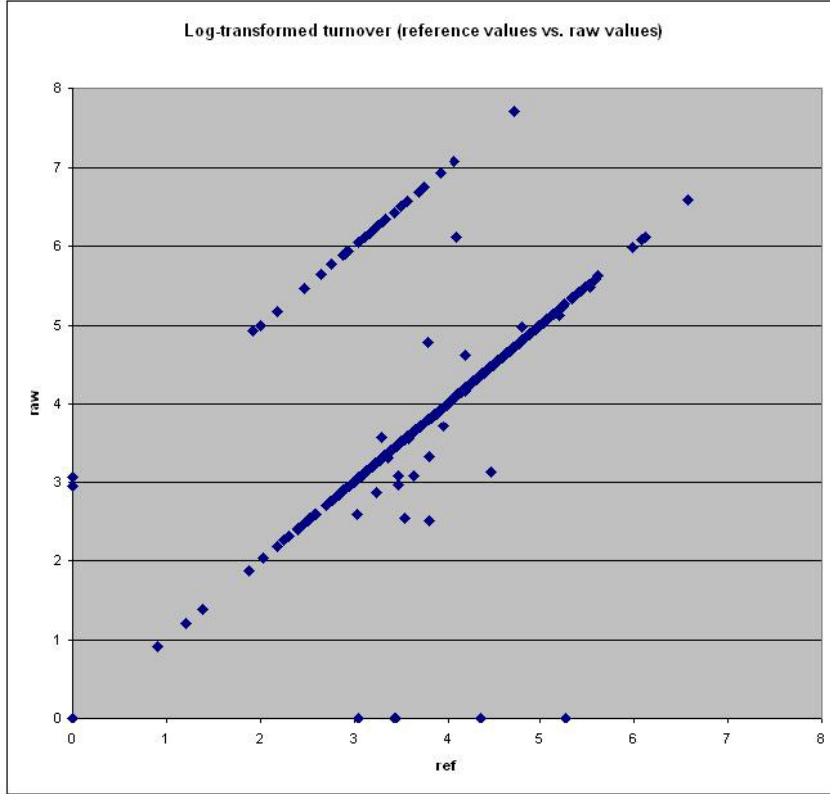


Figure 1. A scatter plot displaying thousand-errors on a logarithmic scale

Elaborating on this graphical approach, Al-Hamad et al. (2008) proposed an alternative automatic method for detecting unit of measurement errors. They considered the difference between the number of digits in the unedited value and the reference value:

$$diff = \left| \lceil \log_{10} x \rceil - \lceil \log_{10} x_{ref} \rceil \right|, \quad (2)$$

where $\lceil a \rceil$ denotes the smallest integer larger than or equal to a . Using (2), different types of unit of measurement errors may be detected by identifying records with a certain value of $diff$. For example, a thousand-error corresponds to $diff = 3$. It should be noted that this method can also detect unit of measurement errors in the reference data, because the absolute value is taken in (2).

Di Zio et al. (2005) proposed a more complex method for detecting unit of measurement errors, by explicitly modeling both the true data and the error mechanism. They used a so-called finite mixture model to identify different clusters within the data set. Each cluster contains records that are affected by a particular type of unit of measurement error; there is also one cluster of records without unit of measurement errors.

Compared with the traditional methods for detecting unit of measurement errors, the approach of Di Zio et al. (2005) has several interesting features. First, it does not require reference data, because the model is fitted directly to the unedited data. However, reference values may also be included in the model if they are available. Second, the method provides diagnostic measures of its own performance, which can be used to identify observations with a significant probability of being misclassified. A selection of doubtful cases may then be checked by subject-matter experts. Finally, this method provides a natural way to detect partial unit of measurement errors. A drawback of the method is that it

may not always be possible to fit an appropriate model to the data set, especially for data sets with many variables or irregular structures. Di Zio et al. (2007) consider an extension of this approach that can accommodate more general models.

2.4 Identifying new systematic errors

New systematic errors can be identified by analysing edit rule failures. If an edit rule is frequently failed, this can be an indication of the presence of a systematic error in the relevant variables. A further analysis of the records that fail the edit rule, in which the questionnaire is also examined, can bring the cause of the error to light. Once the error has been identified, it is generally quite simple to draw up a deductive method to automatically detect and treat the error.

Detecting new systematic errors can only take place once sufficient data have been collected. The results are therefore usually too late to be used in the production process of the current survey cycle. If the analysis produces new deductive editing methods, then these can be built into the editing process for the data in the next survey cycle.

As far as systematic errors are concerned, prevention is better than cure. Sometimes it is possible to improve the design of the questionnaire so that far fewer respondents make a certain type of error. If many respondents make the same kind of error, this can in fact be an indication that a certain question is not presented clearly enough. In some cases, it is also possible to adapt the processing procedure to ensure that a certain processing error no longer arises. In principle, this approach should be preferred to that of making deductive adjustments afterwards. However, because there are practical objections to the constant adaptation of the questionnaire, one may choose initially to build in a deductive editing method, and to use the accumulated knowledge of systematic errors later in a redesign of the questionnaire. (See also the module “Repeated Surveys – Repeated Surveys”.) Moreover, some systematic errors appear to be impossible to prevent, no matter how well the questionnaire is designed. This is, for instance, the case with the unit of measurement error.

To illustrate the identification of a new systematic error, we consider the data collected in 2001 for the Dutch Structural Business Statistics for Wholesale. In this data set, there are (among many other variables) five variables on labour costs, which should satisfy the following edit rule:

$$x_1 + x_2 + x_3 + x_4 = x_5. \quad (3)$$

Here, x_5 represents the variable *total labour costs*. The other four variables are the sub-items of this total. Table 2 shows several records that do not satisfy edit rule (3).

Table 2. Examples of inconsistent partial records in the Dutch SBS for Wholesale 2001

	record 1	record 2	record 3	record 4
x_1	1,100	364	1,135	901
x_2	88	46	196	134
x_3	40	34	68	0
x_4	42	0	42	0
x_5	170	80	306	134

It is striking that, for all records in Table 2, it holds that $x_2 + x_3 + x_4 = x_5$. This suggests that these reporters have ignored the first sub-item x_1 in the calculation of x_5 . A closer look at the questionnaire (see Figure 2) reveals why this could have happened: there is a gap between the answer box for x_1 and the other boxes. As a result, from the design of the questionnaire alone, it is ambiguous whether x_1 should be part of the sum or separate from the rest. Most respondents understand from the context what the intention is, but in several dozen records, we found the same error as in Table 2.

Arbeidskosten	
D.4	Brutolonen en -salarissen van het bij vraag B.1 opgegeven personeel
	Sociale lasten, bestaande uit:
D.5	Werkgeversaandeel sociale voorzieningen
D.6	Pensioenlasten
D.7	Overige sociale lasten
D.8	Totaal arbeidskosten

Figure 2. Part of the questionnaire used for the Dutch SBS Wholesale (until 2005)

We can draw up a deductive method that resolves this error. A more structural solution consists of removing the cause of the error by adapting the questionnaire. This has already been done: the questionnaire from Figure 2 was replaced for the Dutch Structural Business Statistics of 2006. On the new questionnaire, the answer boxes are spaced evenly.

3. Preparatory phase

4. Examples – not tool specific

4.1 Example: Correction rules for the statistic Building Objects in Preparation¹

The Dutch quarterly statistic Building Objects in Preparation (BOP) follows the development of the total construction value of new contracts at architectural firms in the Netherlands. In 2007, a new editing process was designed for this statistic.

When filling in the BOP questionnaire, the reporter must answer several questions about each building object separately. The reporter must tick a box indicating whether the building object concerns a residence (r), a combined-purpose building (c ; this means that the building is used for other purposes as well as residential purposes) or neither of these (o for other). Another question concerns n , the total number of dwellings in the building. For a combined-purpose building, the percentage of floor area intended for residential use (p) is also requested.

¹ This example is adapted from a report written in Dutch by Mark van der Loo and Jeroen Pannekoek (Statistics Netherlands).

The statement contains an error if zero, two, or three of the boxes for r , c , and o have been ticked. In that case, the type of building object has not been clearly specified. In certain situations, this error can be treated deductively based on the values of n and p .

If the value indicated for n is greater than zero and if, moreover, p is equal to 100% or is not filled in, then it is obvious that the building object is a residence. If n is larger than zero and furthermore if p is not equal to 0 or 100%, it is obvious that the building object is a combined-purpose building. And, finally, if neither n nor p has been filled in, or if they have been given the value of 0, then it is highly probable that the building object falls in the category ‘other’. These interpretations follow from the assumption that the statement must be rendered correct by changing as few values as possible.

We write $r = T$ if the box for residence has been ticked, and otherwise $r = F$, and we do the same for c and o . The following correction rule expresses the deductive assertions made in the previous paragraph in formal notation:

```

if  $(r,c,o) \in \{ (T,T,T) , (T,T,F) , (T,F,T) , (F,T,T) , (F,F,F) \}$ 
  then
    if  $( p = \text{'empty'} \text{ or } p = 100\% ) \text{ and } n > 0$ 
      then  $(r,c,o) = (T,F,F)$ 
    if  $0\% < p < 100\% \text{ and } n > 0$ 
      then  $(r,c,o) = (F,T,F)$ 
    if  $( p = \text{'empty'} \text{ or } p = 0\% ) \text{ and } ( n = \text{'empty'} \text{ or } n = 0 )$ 
      then  $(r,c,o) = (F,F,T)$ .

```

This is a small part of the editing process for the statistic BOP.

In the implementation of the editing process for BOP, the derivation of the correction always takes place separately from the actual application of the correction. Initially, in the above example, only an indicator is created that specifies for each record whether a deductive correction is applicable, and if so, which one. Only in the next step are the values of r , c and o changed in the record. As such, the editing process is transparent, so that it is clearly visible afterwards exactly what changes have been made to each record.

4.2 Example: Simple typing errors

We consider a fictitious survey in which the values of *Turnover*, *Costs*, and *Profit* are asked from businesses. By definition, these variables are related through the following edit rule:

$$\text{Turnover} - \text{Costs} = \text{Profit}. \quad (4)$$

The first column of Table 3 shows a record that is inconsistent with respect to (4). The inconsistency can be resolved by adapting any one of the three variables. Moreover, under the assumption that only one variable contains an error, its true value can be computed by inserting the observed values of the other variables into equation (4). The other columns of Table 3 show the three consistent versions of the original record that can be produced by adapting one of the variables (the adapted value is shown in bold in each column).

Table 3. Example of a record with a simple typing error

	record	adjustment 1	adjustment 2	adjustment 3
<i>Turnover</i>	252	315	252	252
<i>Costs</i>	192	192	129	192
<i>Profit</i>	123	123	123	60

Intuitively, the solution in which *Costs* is adapted is the most attractive, since it has the nice interpretation that two adjacent digits were interchanged by mistake. That is to say, it seems much more probable that the true value of 129 was changed to 192 at some point during the collection and processing of the data, than the case that 315 was changed to 252 or 60 to 123. Therefore, we could draw up the following rule for deductive editing: if a record does not satisfy (4), but it can be made consistent by interchanging two adjacent digits in one of the observed values (and, moreover, this can be done in a unique way), then the inconsistency should be treated in this way.

Interchanging two adjacent digits is an example of a simple typing error. Other examples include:

- adding a digit (for example, writing ‘1629’ instead of ‘129’);
- omitting a digit (for example, writing ‘19’ instead of ‘129’);
- replacing a digit (for example, writing ‘149’ instead of ‘129’).

Common features of all simple typing errors are that they only affect one value at a time, and that they produce an observed erroneous value which is related to the unobserved true value in a way that is easy to recognise.

In the example from Table 3, the simple typing error could be detected by using the fact that the variables should satisfy edit rule (4). In general, a survey may contain variables that are related by many equalities and also by other types of edit rules. Moreover, the equalities may be interrelated, so that variables have to satisfy different edit rules simultaneously. Scholtus (2009) described a deductive method for detecting and treating simple typing errors in this more general setting.

Simple typing errors are generic errors, because they occur in many different surveys and they are not content-related. This type of error is easy to make and can therefore occur frequently in practice. A review of data from the Dutch Structural Business Statistics for Wholesale in 2007 revealed, for example, that nearly 10% of all inconsistencies in linear equalities could be explained by one of the four typing errors mentioned above (Scholtus, 2009).

5. Examples – tool specific

The R package `deducorrect`, which can be downloaded for free at <http://cran.r-project.org>, contains an implementation of deductive editing methods for several generic errors:

- sign errors and interchanged values;
- simple typing errors (as defined in Section 3.2);
- rounding errors (very small inconsistencies with respect to equality constraints).

The underlying methodology is described by Scholtus (2011) for sign errors and rounding errors, and by Scholtus (2009) for simple typing errors. To illustrate the use of `deducorrect`, we work out an example. Consider a data set of 11 variables that should satisfy the following edit rules:

$$\left\{ \begin{array}{l} x_1 + x_2 = x_3 \\ x_2 = x_4 \\ x_5 + x_6 + x_7 = x_8 \\ x_3 + x_8 = x_9 \\ x_9 - x_{10} = x_{11} \end{array} \right.$$

The following record is inconsistent with respect to these edit rules; in fact, it does not satisfy the second, fourth, and fifth constraints:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1452	116	1568	161	323	76	12	411	19979	1842	137

We shall use the `deducorrect` package to treat this record for simple typing errors. First, we load the package:

```
> library(deducorrect)
```

Next, we create an object of type “editmatrix” containing the system of edit rules:

```
> E <- editmatrix( c("x1 + x2 == x3",
+                   "x2 == x4",
+                   "x5 + x6 + x7 == x8",
+                   "x3 + x8 == x9",
+                   "x9 - x10 == x11") )
```

We also have to read in the record that we want to treat as a data frame:

```
> x <- data.frame( x1 = 1452, x2 = 116, x3 = 1568, x4 = 161,
+                 x5 = 323, x6 = 76, x7 = 12, x8 = 411,
+                 x9 = 19979, x10 = 1842, x11 = 137 )
```

To check whether simple typing errors can be found in this record, we use the function `correctTypos` provided by the package:

```
> sol <- correctTypos(E, x)
```

The object `sol` is a list which contains the results of the search for simple typing errors. We first check the status of the record:

```
> sol$status
      status
1 corrected
```

The status ‘corrected’ means that the record could be made consistent with respect to all edit rules by only treating simple typing errors. Other possible statuses are: ‘valid’ for a record that was consistent in the first place, ‘invalid’ for an inconsistent record in which no typing error could be detected, and ‘partial’ for a record that could be made consistent with respect to some, but not all edit rules by treating simple typing errors.

The list `sol` also contains the adjusted version of the record and a table of the suggested adjustments:

```
> sol$corrected
      x1  x2   x3  x4  x5 x6 x7  x8   x9  x10 x11
1 1452 116 1568 116 323 76 12 411 1979 1842 137

> sol$corrections
      row variable   old  new
1     1         x4   161 116
2     1         x9 19979 1979
```

Thus, `correctTypos` has detected two simple typing errors in this example: the value of x_4 should be 116 instead of 161 (interchanged adjacent digits), and the value of x_9 should be 1979 instead of 19979 (added digit). By treating these errors, a consistent record is obtained with respect to all edit rules.

We refer to Van der Loo et al. (2011) for more details on the `deducorrect` package.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Al-Hamad, A., Lewis, D., and Silva, P. L. N. (2008), Assessing the Performance of the Thousand Pounds Automatic Editing Procedure at the ONS and the Need for an Alternative Approach. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.
- De Jong, A. (2002), Uni-Edit: Standardized Processing of Structural Business Statistics in the Netherlands. Working Paper, UN/ECE Work Session on Statistical Data Editing, Helsinki.
- De Waal, T. and Scholtus, S. (2011), Methods for Automatic Statistical Data Editing. Paper presented at the 2011 KSS International Conference on Statistics and Probability, Busan.
- Di Zio, M., Guarnera, U., and Luzi, O. (2005), Editing Systematic Unity Measure Errors through Mixture Modelling. *Survey Methodology* **31**, 53–63.
- Di Zio, M., Guarnera, U., and Rocci, R. (2007), A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error. *Computational Statistics & Data Analysis* **51**, 2573–2585.
- Hoogland, J., van der Loo, M., Pannekoek, J., and Scholtus, S. (2011), *Data Editing: Detection and Correction of Errors*. Methods Series Theme, Statistics Netherlands, The Hague.
- Pannekoek, J. and de Waal, T. (2005), Automatic Edit and Imputation for Business Surveys: the Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* **21**, 257–286.
- Scholtus, S. (2009), Automatic Correction of Simple Typing Errors in Numerical Data with Balance Edits. Discussion Paper 09046, Statistics Netherlands, The Hague.
- Scholtus, S. (2011), Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data. *Journal of Official Statistics* **27**, 467–490.

Van der Loo, M., de Jonge, E., and Scholtus, S. (2011), Correction of Rounding, Typing, and Sign Errors with the deducorrect Package. Discussion Paper 201119, Statistics Netherlands, The Hague.

Specific section

8. Purpose of the method

Detecting and treating errors in a deductive manner

9. Recommended use of the method

1. The method should be used, in principle, only for detecting and treating systematic errors.
2. Deductive editing is most effective when it is applied at the very beginning of the editing process, before any other form of editing has been used.

10. Possible disadvantages of the method

1. Deductive editing should only be used to treat errors for which the error mechanism is known with sufficient reliability. Deductive adjustments based on invalid assumptions can produce biased estimators.
2. It may be difficult to maintain a large collection of deterministic correction rules over a long period of time. In particular, it becomes difficult to grasp the consequences of adding or removing a correction rule, or changing the order in which the rules are applied, when faced with a large collection of rules.

11. Variants of the method

1. Each type of systematic error requires its own particular variant.

12. Input data

1. A data set containing unedited microdata.
2. If relevant, a data set containing reference data

13. Logical preconditions

1. Missing values
 1. Allowed, but an assumption has to be made on their interpretation (e.g., “consider all missing values to be equal to zero unless evidence to the contrary is found”).
2. Erroneous values
 1. Allowed; in fact, the object of this method is to detect and treat some of them.
3. Other quality related preconditions
 1. n/a
4. Other types of preconditions
 1. n/a

14. Tuning parameters

1. If relevant, a collection of edit rules for the microdata.

2. Other parameters, depending on the particular variant / type of error.

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. A data set containing partially edited microdata, which is an updated version of the first input data set.

17. Properties of the output data

1. Ideally, the data set should contain no more systematic errors, only random errors.

18. Unit of input data suitable for the method

Incremental processing by record

19. User interaction - not tool specific

1. User interaction is not needed during an execution of deductive editing.

20. Logging indicators

1. All adjustments that are introduced by each deductive editing method should be flagged as such. This helps to keep the editing process transparent and it also provides input for future analyses of the editing process itself.

21. Quality indicators of the output data

1. The quality of deductive editing can be assessed in a simulation study. This requires a data set that has been edited by experts to a point where the edited data may be considered error-free. In the simulation study, the original data are edited again using deductive editing methods. The quality of a deductive editing method may then be measured in terms of its success in detecting systematic errors in the original data set.
2. Alternatively, one could also perform a simulation study by introducing artificial systematic errors into an existing data file. The quality of a deductive editing method may then be measured in terms of its success in identifying these artificial errors.

22. Actual use of the method

1. Several forms of deductive editing are used in the production process for Structural Business Statistics at Statistics Netherlands (see De Jong, 2002).

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Repeated Surveys – Repeated Surveys
2. Statistical Data Editing – Main Module

24. Related methods described in other modules

1. Statistical Data Editing – Automatic Editing

25. Mathematical techniques used by the method described in this module

1. n/a

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1. R package `deducorrect`

28. Process step performed by the method

Statistical data editing

Administrative section

29. Module code

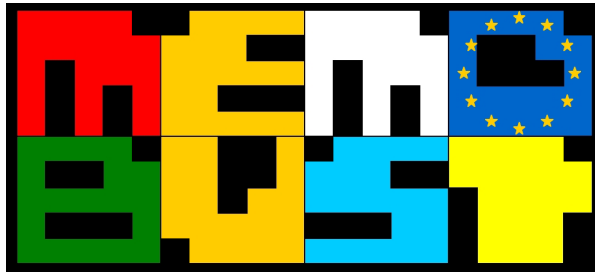
Statistical Data Editing-M-Deductive Editing

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	22-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	20-04-2012	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.2.1	16-07-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.3	04-09-2013	minor improvements	Sander Scholtus	CBS (Netherlands)
0.3.1	09-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:11



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Selective Editing

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Selective editing	3
2.2 Score function.....	3
2.3 The selection rule	4
2.4 How to compute the threshold.....	5
2.5 Dealing with errors remaining in data: a probability sampling approach to selective editing 7	
3. Design issues	8
4. Available software tools.....	8
5. Decision tree of methods	9
6. Glossary.....	9
7. References	9
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

The experience of NSIs in the field of correction of errors has led to assume that only a small subset of observations is affected by influential errors, i.e., errors with a high impact on the estimates, while the rest of the observations are not contaminated or contain errors having small impact on the estimates. Selective editing is a general approach to the detection of errors, and it is based on the idea of looking for important errors in order to focus the treatment on the corresponding subset of units to reduce the cost of the editing phase, while maintaining the desired level of quality of estimates. In this section a general description of the framework and the main elements of selective editing is given.

2. General description

2.1 *Selective editing*

The experience of NSIs in the field of correction of errors has led to assume that only a small subset of observations is affected by influential errors, i.e., errors with a high impact on the estimates, while the rest of the observations are not contaminated or contain errors having small impact on the estimates (Hedlin, 2003). This assumption and the fact that the interactive editing procedures, like for instance, recontact of respondents, are resource demanding, have motivated the idea at the basis of selective editing, that is to look for important errors (errors with an harmful impact on estimates) in order to focus the expensive interactive treatments (follow-up, recontact) only on this subset of units. This should reduce the cost of the editing phase maintaining at the same time an acceptable level of quality of estimates (Lawrence and McKenzie, 2000; Lawrence and McDavitt, 1994). In practice, observations are ranked according to the values of a *score function* expressing the impact of their potential errors on the target estimates (Latouche and Berthelot, 1992), and all the units with a score above a given threshold are selected.

2.2 *Score function*

The score function is an instrument to prioritise observations according to the expected benefit of their correction on the target estimates. According to this definition, it is natural to think of the score function as an estimate of the error affecting data. The estimate is generally based on comparing observed values with predictions (sometimes called *anticipated values*) obtained from some explicit or implicit model for the data. In the case of sample surveys, the comparison should also include the sampling weights in order to properly take into account the error impact on the estimates. An additional element often considered in the context of selective editing, is the *degree of suspiciousness*, that is an indicator measuring, loosely speaking, the probability of being in error. The necessity of this element arises from the implicit assumption of the intermittent nature of the error in survey data, i.e., the assumption that only a certain proportion of the data are affected by error, or, from a probabilistic perspective, that each measured value has a certain probability of being erroneous (Buglielli et al., 2011). Some authors do not introduce this element, others implicitly use it in their proposals. Norberg et al. (2010) state that several case studies indicate that procedures based only on the comparison of observed and predicted values without the use of a degree of suspiciousness tend to generate a large proportion of false alarm.

Several score functions are proposed in literature, the difference being mainly given by the kind of prediction and the use of ‘degree of suspiciousness’.

Among the different methods used to obtain predictions it is worthwhile to mention the use of information coming from a previous occasion of the survey (Latouche and Berthelot, 1992), regression models (Norberg et al., 2010), contamination models (Buglielli et al., 2011). A detailed review can be found in De Waal et al. (2011).

As far as the degree of suspiciousness is concerned, a common drastic approach consists in introducing it in the score function through a zero-one indicator that multiplies the difference between observed and predicted values, where zero and one correspond to consistency or inconsistency respectively with respect to some edit rules. In this case it is assumed that errors appear only as edit failures and observations that pass the edits are considered error-free without uncertainty (Latouche and Berthelot, 1992). More refined methods to estimate the probability of being in error can be found in Norberg et al. (2010) and Buglielli et al. (2011). In the first case a nonparametric approach based on quantiles is used, while in the second a latent model based on a mixture of normal (or lognormal) distributions is proposed.

Prediction and suspiciousness can be combined to form a score for a single variable, named *local score*. A local score frequently used for the unit i with respect to the variable Y_j is

$$S_{ij} = \frac{p_i w_i |y_{ij} - \tilde{y}_{ij}|}{\hat{T}_{Y_j}}$$

where p_i is the degree of suspiciousness, y_{ij} is the observed value of the variable Y_j on the i th unit, \tilde{y}_{ij} is the corresponding prediction, w_i is the sampling weight, and \hat{T}_{Y_j} is an estimate of the target parameter.

Once the local scores for the variables of interest are computed, a global score to prioritise observations is needed.

Several functions can be used to obtain the global score (see Hedlin, 2008); an example is the sum of squares $GS_i^{(2)} = \sum_j S_{ij}^2$.

In some cases, some variables can be considered to be more important than others. Such situations can be dealt with by multiplying the local scores by weights stating their relative importance.

2.3 The selection rule

Once the observations have been ordered according to their global score, it is important to build a rule in order to determine the number of units to be reviewed.

A first rule can be suggested by budget constraints. In this case, it is obvious to choose the first n^* observations, in the given ordering, such that the budget constraints are satisfied.

A more interesting and complex approach is to select the subset of units such that the impact on the target estimates of the errors remaining in the unedited observations is negligible, that is in fact the core of selective editing. Since the true values are unknown, this bias cannot be evaluated and an approximation is used. This approximation can be expressed in terms of the weighted differences

between the raw values y_{ij} and the anticipated values \tilde{y}_{ij} for the variable Y_j in the units i not selected for interactive treatment (EDIMBUS, 2007).

Let T_{Y_j} be the target quantity related to the variable Y_j (for instance the total), the estimated bias is given by

$$EB_j(t) = \frac{\left| \sum_{i \notin E_t} w_i (y_{ij} - \tilde{y}_{ij}) \right|}{\hat{T}_{Y_j}},$$

where w_i is the sampling weight of the i th unit, \hat{T}_{Y_j} is an estimate of the target quantity T_{Y_j} , and E_t is the set of units to be selected. This set is composed of all the units having a global score $GS > t$, where t is a threshold value such that $EB_j(t)$ is below a predefined value.

An alternative measure known as the *estimated relative bias* is obtained by replacing the estimate of the total at the denominator of EB with the standard error of the estimate \hat{T}_{Y_j} . With this measure, the error due to the non-sampling error left in data is compared with the sampling error. The reasoning underlying is that there is no need to edit observations because the ‘noise’ due to their errors is overwhelmed by the sampling error.

We remark that when edited values are available, they can be used as anticipated values, in this case the estimated bias and the estimated relative bias are the absolute pseudo bias and the relative pseudo-bias introduced by Latouche and Berthelot (1992) and Lawrence and McDavitt (1994), respectively.

It is worthwhile to note the similarity between the terms appearing in the sum defining the estimated bias and the local score function. The main difference is in the parameter related to the suspiciousness. In fact in the estimated bias all differences between observed values and corresponding predictions are considered as they were determined by errors, while in the score functions, where the degree of suspiciousness is included, this is not assumed with certainty.

2.4 How to compute the threshold

There are two approaches: 1) through a simulation study, 2) by using a model.

2.4.1 Simulation approach

This approach is based on the availability of raw and edited data comparable with the data on which selective editing has to be applied. The idea is to simulate the selective editing procedure considering the edited data as if they were the ‘true’ data. Often data from a previous cycle of the same survey are used for this purpose.

The approach can be described by the following steps (De Waal et al., 2011).

- Compute the global scores for the raw data and order (decreasingly) the observations.
- Determine a subset E of units composed of the first p units and replace their raw values with the corresponding edited values.
- Compare the estimates computed using the completely edited data set and the raw data where the subset E is obtained according to step 2.

- Repeat steps 2 and 3 with different values of p until the difference between the two estimates is negligible. Let p^* be the first index such that this condition is fulfilled.
- The threshold t is the value of the GS corresponding to the p^* -th unit.

Remarks:

- The assumption of this approach is that the edited data can be considered as ‘true’ data. This is a limitation because it can be rarely assumed.
- The simulation approach is frequently applied to data of a previous survey occasion to obtain a threshold value to be used for the current survey. It is worthwhile to note that in this case we assume that the error mechanism and the data distribution are the same in the two occasions.
- The method cannot be applied when you deal with the first wave of a survey.

2.4.2 Model based approaches

In this context, some of the main elements of the problem are modelled through a probability distribution: the true data distribution, the error mechanism, the score functions.

The introduction of a model may be useful to give estimates of the error left in data after the revision of the selected units and thus to ease the determination of a threshold for the selection of units to be reviewed.

A first attempt can be found in Lawrence and McKenzie (2000). By denoting with a the threshold value, they assume that the difference between the observed and the predicted value for the non-selected observations follows a uniform distribution in the interval $(-a, a)$, i.e., $U(-a, a)$. The threshold a is determined so that the bias due to not editing a set of units is low if compared to the sampling error.

A conservative solution is $a = \sqrt{\frac{3k}{n}} SE(\hat{Y})$, where $kSE(\hat{Y})$, $k < 1$ is the upper bound for the bias and n is the total number of observations.

The intermittent nature of the error is taken into account in Arbués et al. (2011). The search of a good selective editing strategy is stated as an optimisation problem in which the objective is to minimise the expected workload with the constraint that the expected error of the aggregates computed with the edited data is below a certain constant.

A model based approach is also adopted by Buglielli et al. (2011). They propose to consider (log)- true data y_i^* as realisations from a multivariate Gaussian distribution with mean vector possibly dependent on a set of error-free covariates: $\tilde{y}_i \sim N(\mu_i, \Sigma)$. Errors are supposed to act on a subset of data by inflating the variance, i.e., the covariance matrix of the contaminated data is $\lambda\Sigma$ where λ is a numerical factor greater than one. The intermittent nature of the error is reflected by a Bernoullian random variable with parameter π taking values zero or one depending on whether an error occurs in a unit or not, respectively. This approach naturally leads to a latent class model formulation, where observed data (y) can be viewed as realisation from a mixture of two Gaussian probability distributions associated to contaminated and error-free data:

$$f_Y(y) = (1 - \pi)N(y; \boldsymbol{\mu}, \Sigma) + \pi N(y; \boldsymbol{\mu}, (\boldsymbol{\lambda} + 1)\Sigma).$$

In this context, the parameter π represents the mixing weight of the mixture and can be interpreted as the *a priori* probability of errors in data. The estimated conditional distribution of true data given observed ones is used to build an appropriate score function. More precisely, for a given variable of interest, a relative (local) score function is defined in terms of difference between the observed value and the expectation of the “true” value conditional on the observed one (the prediction). This approach allows to interpret the score function as the expected error, and to relate the threshold for interacting reviewing to the accuracy of the estimates of interest. A global score can be defined in many ways combining the different local score functions. In Buglielli et al. (2011) the global score is defined as the maximum of the single local scores. This ensures that the accuracy of the estimates is kept under control simultaneously for all the variables of interest.

In practice the steps to perform selective editing within this framework are similar to the ones detailed in the simulation approach, with the difference that the predicted value is obtained by using an explicit model, and that the score directly gives an estimate of the error contaminating each observation.

Remarks:

- The introduction of a model for the error mechanism allows to formalise the problem and hence to have a statistical interpretation of the elements characterising selective editing. Furthermore, using a latent class model implies the advantage that no edited data are required, and the bias of the simulation approach due to considering edited data as true data is avoided.
- The main drawback is that the validity of the conclusions depends on the validity of the model assumptions.

2.5 *Dealing with errors remaining in data: a probability sampling approach to selective editing*

Ilves and Laitila (2009) and Ilves (2010) propose a two-step procedure for selective editing. Their proposal is motivated by the fact that the non-selected observations may still be affected by errors resulting in a biased target parameter estimator \hat{T}_Y . To obtain an unbiased estimator a sub-sample is drawn from the unedited observations (below threshold for global scores), follow-up activities with recontacts are carried through and the bias due to remaining errors is estimated.

The estimated bias is used to make the target parameter estimator \hat{T}_Y unbiased. If our target parameter is the total of the population, the bias-corrected estimator is obtained by subtracting the estimated bias from the HT estimator of the total computed on edited (selected by the selective editing procedure) and unedited (non-selected) observations. Formulas for the variance and a variance estimator are derived by using a two-phase sampling approach. The procedure is discussed in general without specifying a particular selective editing technique, but sampling with probabilities proportional to scores seems to be the obvious choice.

3. Design issues

In the following some important elements concerning the design of a selective editing procedure are reported.

- Selective editing can be applied only to numerical variables. This implies that selective editing is mainly applied to business surveys.
- Selective editing is useful when accurate interactive editing can be performed.
- Selective editing can be applied at the early stages of data collection. This kind of application is named *input editing*. The methods used in this context apply to each incoming record individually, classifying each record as critical or non-critical. The advantage of input editing is that time-consuming task procedures as interactive editing and follow-up are started as soon as possible, with positive effects on response burden and the timeliness of the results. The disadvantage is that the parameters needed for the selection of influential errors should be estimated before data are available. This can be performed only when data from previous survey occasions are available (or strong a priori knowledge is disposable), and the assumptions are that the situation is not changed from the previous surveys to the actual one. On the contrary, the approach consisting in applying selective editing when almost all the data are available is named *output editing*. The disadvantage is clearly related to the timeliness of the results because time consuming task as interactive editing or follow-up are moved to a later stage of the process. The advantage is that all the parameters needed for the selection of influential errors are estimated on the data at hand, so they refer to the actual distribution of data with a potential benefit effect on the precision of selection.
- It is advisable to apply selective editing after the process of detection and correction of systematic errors (see “Statistical Data Editing – Main Module”). Actually, also systematic errors can lead to significant bias but they can often be automatically detected and corrected easily and very reliably. It is highly efficient to correct these errors at an early stage.
- The application of selective editing should be limited to the subset composed of the most important target variables.
- Once one observation is selected, all the variables should possibly be revised, not only the ones considered in the score function.
- Sampling weights are important to estimate the impact of errors on the final estimates. When an input editing approach is chosen, initial sampling weights may be used.

4. Available software tools

- SeleMix is an R-package for selective editing based on contamination models (Di Zio and Guarnera, 2011) freely available on the website <http://cran.r-project.org/>.
- Selekt is a set of SAS-macros for selective editing, allowing “traditional” hard and soft edits as well as a nonparametric approach based on quantiles to produce measures of suspicion. Selekt works with one and two-stage samples and several sets of domains in output. (Norberg et al., 2010; Norberg, et al., 2011).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Arbués, I., Revilla, P., and Saldaña, S. (2011), Selective Editing as a Stochastic Optimization Problem. UN/ECE Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9-11 May 2011.
- Buglielli, T., Di Zio, M., Guarnera, U., and Pogelli, F. R. (2011), Selective Editing of Business Survey Data Based on Contamination Models: an Experimental Application. NTTS 2011 New Techniques and Technologies for Statistics, Bruxelles, 22-24 February 2011.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- Di Zio, M. and Guarnera, U. (2011), SeleMix: an R Package for Selective Editing via Contamination Models. *Proceedings of the 2011 International Methodology Symposium, Statistics Canada. November 1-4, 2011, Ottawa, Canada*.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Hedlin, D. (2003), Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics* **19**, 177–199.
- Hedlin, D. (2008), Local and Global Score Functions in Selective Editing. UN/ECE Work Session on Statistical Data Editing, Wien.
- Ilves, M. and Laitila, T. (2009), Probability-Sampling Approach to Editing. *Austrian Journal of Statistics* **38**, 171–182.
- Ilves M. (2010), Probabilistic Approach to Editing. Workshop on Survey Sampling Theory and Methodology Vilnius, Lithuania, August 23-27, 2010.
- Latouche, M. and Berthelot, J. M. (1992), Use of a Score Function To Prioritise and Limit Recontacts in Business Surveys. *Journal of Official Statistics* **8**, 389–400.
- Lawrence, D. and McDavitt, C. (1994), Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics* **10**, 437–447.
- Lawrence, D. and McKenzie, R. (2000), The General Application of Significance Editing. *Journal of Official Statistics* **16**, 243–253.
- Norberg, A. et al. (2010), *A General Methodology for Selective Data Editing*. Statistics Sweden.
- Norberg, A. et al. (2011), *User’s Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing*. Statistics Sweden.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Statistical Data Editing – Automatic Editing
3. Statistical Data Editing – Manual Editing
4. Statistical Data Editing – Macro-Editing
5. Imputation – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

14. Module code

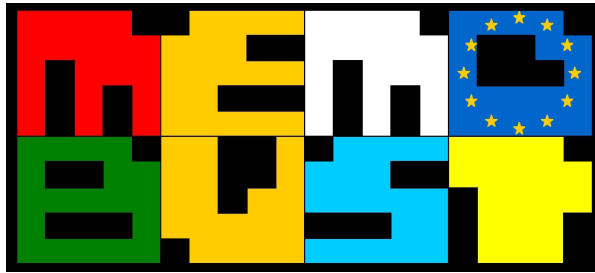
Statistical Data Editing-T-Selective Editing

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	08-02-2012	first version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.2	19-03-2012	second version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.3	06-04-2012	third version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.3.1	04-10-2013	preliminary release		
0.4	15-10-2013	changes according to the EB comments	Di Zio Marco, Guarnera Ugo	Istat (Italy)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:11



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Automatic Editing

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Introduction to automatic editing	3
2.2 Edit rules.....	4
2.3 The error localisation problem	6
2.4 Solving the error localisation problem: the method of Fellegi and Holt	7
2.5 Solving the error localisation problem: other methods	10
3. Preparatory phase	11
4. Examples – not tool specific.....	11
5. Examples – tool specific.....	12
6. Glossary.....	13
7. References	13
Specific section.....	16
Interconnections with other modules.....	18
Administrative section.....	20

General section

1. Summary

The goal of automatic editing is to accurately detect and treat errors and missing values in a data file in a fully automated manner, i.e., without human intervention. Methods for automatic editing have been investigated at statistical institutes since the 1960s (Nordbotten, 1963). In practice, automatic editing usually implies that the data are made consistent with respect to a set of predefined constraints: the so-called *edit rules* or *edits*. The data file is checked record by record. If a record fails one or more edit rules, the method produces a list of fields that can be imputed so that all rules are satisfied.

In this module, we focus on automatic editing based on the (generalised) Fellegi-Holt paradigm. This means that the smallest (weighted) number of fields is determined which will allow the record to be imputed consistently. Designating the fields to be imputed is called error localisation. In practice, error localisation by applying the Fellegi-Holt paradigm often requires dedicated software, due to the computational complexity of the problem.

Although the imputation of new values for erroneous fields is often seen as a part of automatic editing, we do not discuss this here, because the topic of imputation is broad and interesting enough to merit a separate description. We refer to the theme module ‘Imputation’ and its associated method modules for a treatment of imputation in general and various imputation methods.

2. General description of the method

2.1 Introduction to automatic editing

For efficiency reasons, it can be desirable to edit at least part of a data file by means of automatic methods (see “Statistical Data Editing – Main Module”). Assuming that all systematic errors with a known structural cause have already been treated using methods for deductive editing (see the method module “Statistical Data Editing – Deductive Editing”), the task remains to also detect and treat random errors. In the literature on data editing, the problem of identifying the erroneous values in a record containing only random errors is known as the *error localisation problem*. Compared to detecting systematic errors, solving the error localisation problem is usually more difficult and requires complex methodology.

Broadly speaking, there are two approaches to solving the error localisation problem. The first approach uses outlier detection techniques in combination with an implicit or explicit statistical model for the data under consideration. Records corresponding to data points that do not fit the model well are supposed to contain errors, and within such a record, the values that contribute most to the ‘outlyingness’ of that record are identified as erroneous; see, e.g., Little and Smith (1987) and Ghosh-Dastidar and Schafer (2003). This approach appears to be mainly suitable for editing low-dimensional data (data sets containing a small number of variables). Moreover, if there are edit rules that define consistency constraints for the variables in the data set, these cannot be used under this approach. In particular, the edited data will not necessarily satisfy the edit rules. For these reasons, this approach is not ideal for automatic editing in business surveys at statistical offices, where one typically encounters data sets with many variables and many edit rules. In fact, it is seldom used in this context.

In the remainder of this module, we shall focus on the second approach. Under this approach, a set of edit rules is defined for the data set. A record is called *consistent* – and is considered to be error-free – if it satisfies all edit rules. For inconsistent records, the erroneous values are identified by solving a mathematical optimisation problem.

The remainder of this section is organised as follows. Section 2.2 considers edit rules. In Section 2.3, the error localisation problem is formulated as a mathematical optimisation problem. Sections 2.4 and 2.5 describe techniques for solving this optimisation problem.

2.2 Edit rules

Edit rules are introduced in a more general context in “Statistical Data Editing – Main Module”. Here, we focus on aspects of edit rules that are relevant to automatic editing in particular.

A record of data can be represented as a vector of fields or variables: $x = (x_1, x_2, \dots, x_n)$. The set of values that can be taken by variable x_i is called its domain. Examples of variables and domains are *size class* with domain {'small', 'medium', 'large'}, *number of employees* with domain {0,1,2,...}, and *profit* with domain $(-\infty, \infty)$.

Edit rules indicate conditions that should be satisfied¹ by the values of single variables or combinations of variables in a record. For the purpose of automatic editing, all edit rules must be checkable per record, and may therefore not depend on values in fields of other records. However, they may contain parameters based on external sources (for instance, quantiles of univariate distributions in a reference data set that has been edited previously), provided that these parameters are set prior to the start of the editing process.

For automatic editing of numerical data, it is convenient to assume that all edit rules are written as linear relationships such as

$$Turnover \geq 0$$

or

$$Profit + Costs = Turnover.$$

The general form of a linear edit rule for a record (x_1, x_2, \dots, x_n) is as follows:

$$a_{j1}x_1 + \dots + a_{jn}x_n + b_j \geq 0 \tag{1}$$

or

$$a_{j1}x_1 + \dots + a_{jn}x_n + b_j = 0, \tag{2}$$

¹ Edit rules of this type are sometimes called ‘validity rules’. In some applications, edit rules are specified instead in the form of ‘conflict rules’, which means that they indicate conditions that are satisfied by invalid combinations of values. For instance, an edit rule stating that the variable *turnover* should be non-negative can be written either as the validity rule ‘*turnover* ≥ 0 ’ or as the conflict rule ‘*turnover* < 0 ’. Clearly, both formulations are equivalent. The choice of validity or conflict rules should not lead to difficulties, provided that one of the forms is used consistently.

where j numbers the edit rules, a_{ji} are numerical coefficients and b_j are numerical constants. It should be noted that a *ratio edit* – i.e., a bivariate edit rule of the form

$$x_1/x_2 \geq a,$$

where a denotes a numerical constant and x_1 and x_2 are constrained to be non-negative – can also be expressed as a linear edit rule. Namely, the ratio edit can be rewritten as

$$x_1 - ax_2 \geq 0.$$

For categorical data, an edit rule can identify as admissible any combination of values from the domains of the categorical variables. Categorical edit rules are often written in if-then form, for example:

if *Gender* = ‘male’ **then** *Pregnant* = ‘no’.

Finally, mixed data and mixed edit rules, containing both categorical and numerical variables also occur in practice. Mixed edit rules are also often written in if-then form. For example:

if *Size Class* = ‘small’ **then** *Number of Employees* < 10.

For automatic processing, it can be convenient to require that the if-part of a mixed edit only contains categorical variables, while the then-part only contains numerical variables. The above-mentioned example is written in this form. Many types of mixed edits can be rewritten in this simple form, although this may require the introduction of auxiliary variables; see De Waal (2005).

In the remainder of this module, we focus on numerical data and linear edits, because these are most common to business surveys. We refer to De Waal et al. (2011) for a discussion of automatic editing of categorical or mixed data. A numerical variable x_i is said to be *involved* in an edit rule of the form (1) or (2) if it holds that $a_{ji} \neq 0$. Clearly, whether a record fails or satisfies an edit rule only depends on the values of the variables that are involved in that edit rule.

In manual editing, subject-matter specialists often distinguish between *hard* and *soft* edit rules. As mentioned in “Statistical Data Editing – Main Module”, hard edit rules are rules that must hold by definition, while soft edit rules only indicate whether a value, or value combination, is suspicious. A soft edit rule can occasionally be failed by unlikely values that are in fact correct.

In nearly² all methods for automatic editing, no distinction can be made between hard and soft edit rules: all rules are treated as hard edit rules. Thus, in automatic error localisation, all records that fail one or more edit rules are viewed as certainly inconsistent. Hence, formulating edits for the purpose of automatic editing should be done with care (Di Zio et al., 2005). If too many soft edit rules are defined, or soft edit rules that are too strict, there is a danger of *overediting*: the unjustified adaptation of correct values. On the other hand, if too few edit rules are defined, or soft edit rules that are not strict enough, then certain errors might be left in the data after automatic editing.

² In fact, to our best knowledge, all methods for automatic editing that are currently in use at statistical offices do not distinguish between hard and soft edit rules. The method of Freund and Hartley (1967) uses soft edit rules, but it has the important drawback that it cannot handle hard edit rules; hence, it is not recommended to be used in practice. Scholtus (2013) has described a method that incorporates both hard and soft edit rules, but at the time of writing, this method remains to be tested in practice.

2.3 The error localisation problem

For a given record and a collection of edit rules, it is straightforward to verify which values in the record are missing and whether any of the edit rules are failed. However, given that some of the edit rules are failed, determining which values in the record are actually causing the edit failures is much less straightforward. On the one hand, most edit rules involve more than one variable, and on the other hand, most variables are involved in more than one edit rule.

In order to solve the error localisation problem automatically, one has to choose a guiding principle for finding errors. The most commonly used guiding principle for error localisation is the so-called *Fellegi-Holt paradigm*, first formulated by Fellegi and Holt (1976). According to this paradigm, one should minimise the number of observed values that have to be adjusted in order to satisfy all edit rules. This paradigm is often used in a generalised form, for which each variable is given a *reliability weight* $w_i \geq 0$. A high value of w_i indicates that the variable x_i is expected to contain few errors. The generalised Fellegi-Holt paradigm now states that one should search for a subset of the variables E with the following two properties:

- The variables x_i ($i \in E$) can be imputed with values that, together with the observed values of the other variables in the record, satisfy all edit rules.
- Among all subsets that satisfy the first property, E has the smallest value of $\sum_{i \in E} w_i$.

The original Fellegi-Holt paradigm is recovered from this more general form by taking all reliability weights equal, for instance all equal to 1.

A distinctive feature of the (generalised) Fellegi-Holt paradigm is that it does not take the size of the differences between the original and imputed values into account in any way. In fact, the method of Fellegi and Holt only provides a list of variables that can be imputed to satisfy all edit rules, but it does not provide the actual values to impute. These have to be determined in a separate step. This might seem like a drawback, but it actually has the advantage that an appropriate imputation method can be chosen independently of the method used for error localisation. Methods for imputation are discussed in the topic “Imputation”.

Some authors have suggested other guiding principles for error localisation that do look at the size of the adaptations. Casado Valera et al. (1996) proposed to minimise the sum of the squared differences between the observed values and the adjusted values, under the restriction that all edit rules are satisfied by the adjusted values. This leads to a quadratic optimisation problem, which can be solved using standard software. A different formulation of the error localisation problem as a quadratic optimisation problem was proposed by Freund and Hartley (1967).

To illustrate the difference between these principles, we consider a very small example. Suppose that there are two edit rules:

$$\text{Turnover} = \text{Profit} + \text{Costs},$$

$$\text{Turnover} \geq 0,$$

and suppose that we are presented with the following inconsistent record:

$$(\text{Turnover}, \text{Profit}, \text{Costs}) = (-30, 10, 20).$$

Under the Fellegi-Holt paradigm (in its original form, without reliability weights), the optimal solution is to adjust only the value of *Turnover*, because both edits can be satisfied without changing the values of the other variables. After imputation, this certainly yields

$$(\textit{Turnover}, \textit{Profit}, \textit{Costs}) = (30, 10, 20),$$

because the value to impute for *Turnover* is uniquely determined by the edits in this example.

On the other hand, if we minimise the unweighted sum of the squared differences between observed and adjusted values, the optimal solution changes all values:

$$(\textit{Turnover}, \textit{Profit}, \textit{Costs}) = (0, -5, 5).$$

This happens because, under this minimisation criterion, it is optimal to distribute the total adjustment required by the edit rules over as many different variables as possible.

Assuming that errors occur with a low probability and in isolated values, the Fellegi-Holt paradigm appears to be a sensible choice, because it distorts as few of the observed values as possible. Methods that try to distribute the total adjustment over many different variables, such as the quadratic minimisation approach, are less suitable in this context. However, the latter type of method can be useful in the context of micro- or macro-integration, where many small inconsistencies in data from different sources have to be resolved, while preserving patterns that occur in the original data as much as possible. We refer to the topics “Micro-Fusion” (in particular the method module “Micro-Fusion – Reconciling Conflicting Microdata”) and “Macro-Integration” for these subjects.

In order to solve the error localisation problem according to the Fellegi-Holt paradigm, we have to find the smallest subset of the variables that can be imputed so that all edit rules become satisfied. Several methods have been proposed for this. Section 2.4 presents the original method of Fellegi and Holt (1976) for numerical data. Section 2.5 briefly mentions several other methods. These sections contain material that is somewhat more technical than the rest of this module.

2.4 Solving the error localisation problem: the method of Fellegi and Holt

For a given record that fails certain edit rules, we want to determine the smallest subset of the variables that can be imputed so that all edit failures are resolved. A naïve way to solve this problem might proceed as follows: “It is clear that a subset of the variables E can only be a feasible solution to the error localisation problem if every failed edit rule involves at least one variable in E , i.e., if the failed edit rules are ‘covered’ by these variables. Therefore, let us choose the smallest set of variables with this property.” Unfortunately, although ‘covering’ the original failed edits is a necessary condition for a set of variables to be a feasible solution to the error localisation problem, it is not a sufficient condition in general. We will demonstrate this by means of a small example.

Consider the following two numerical edit rules: $x_1 \geq x_2$ and $x_2 \geq x_3$. The unedited record $(x_1, x_2, x_3) = (4, 5, 6)$ fails both edits. Since the variable x_2 is involved in both edit rules – that is to say, the failed edits are ‘covered’ by x_2 –, we might try to obtain consistency with respect to the edit rules by changing only the value of x_2 . This turns out to be impossible, because the imputed value would have to satisfy $4 \geq x_2$ and $x_2 \geq 6$.

Fellegi and Holt (1976) showed that, in order to determine whether a set of variables can be imputed to satisfy all edits simultaneously, it is necessary to derive so-called *implied edits* from the original set of edits. An implied edit is an edit rule that can be derived from the original edit rules by logical reasoning. For numerical data, the number of implied edits that can be derived from even a single original edit is actually infinite; e.g., if $x_1 \geq x_2$ is an edit rule, then so is $\lambda x_1 \geq \lambda x_2$ for any $\lambda > 0$. Fortunately, for the purpose of solving the error localisation problem, it is not necessary to derive all possible implied edits from the original set of edits, but only the so-called *essentially new implied edits* (see below). By adding the essentially new implied edits to the original set of edit rules, one obtains a so-called *complete set of edits*. For a complete set of edit rules, it does hold that any subset of the variables which ‘covers’ all failed edit rules is a feasible solution to the error localisation problem.

In the example above, the complete set of edits consists of the two original edit rules and the (only) essentially new implied edit $x_1 \geq x_3$. The latter edit rule is also failed and it does not involve the variable x_2 , which shows that imputing only x_2 does not solve the error localisation problem. On the other hand, the three failed edits are ‘covered’ by $\{x_1, x_3\}$, and it is easy to see that imputing new values for x_1 and x_3 is indeed a feasible solution to the error localisation problem. In fact, imputing any combination of values with $x_1 \geq 5$ and $5 \geq x_3$ leads to a consistent record in this example.

In general, for a given set of edit rules of the forms (1) and (2), essentially new implied edits are constructed by selecting one of the variables, say x_g , as a so-called *generating variable*. We consider all pairs of edit rules that involve the generating variable, i.e., all pairs (s, t) with $a_{sg} \neq 0$ and $a_{tg} \neq 0$. If one of the edits, say edit s , is an equality, then we may solve this equality for x_g :

$$x_g = \frac{-1}{a_{sg}} (a_{s1}x_1 + \dots + a_{s,g-1}x_{g-1} + a_{s,g+1}x_{g+1} + \dots + a_{sn}x_n + b_s).$$

An implied edit is now obtained from the pair (s, t) by substituting this expression for x_g in edit rule t . This new edit rule is an essentially new implied edit, unless it happens to be identical to an existing edit rule, in which case it is redundant.³

If both edits are inequalities, then we apply a technique called *Fourier-Motzkin elimination* (Williams, 1986; De Waal et al., 2011). First, we check whether the coefficients a_{sg} and a_{tg} have opposite signs, i.e., whether $a_{sg}a_{tg} < 0$. If this is not the case, then this pair does not contribute an essentially new implied edit. Hence, we may assume without loss of generality that $a_{sg} < 0$ and $a_{tg} > 0$. This means that edit rule s can be written as an upper bound on x_g , given the values of the other variables:

$$x_g \leq \frac{-1}{a_{sg}} (a_{s1}x_1 + \dots + a_{s,g-1}x_{g-1} + a_{s,g+1}x_{g+1} + \dots + a_{sn}x_n + b_s).$$

³ To give an example of a redundant edit, suppose that we already have the edit rule ‘ $x \geq 3$ ’ and we derive a new edit rule stating that ‘ $2x \geq 6$ ’. Since the second edit rule is identical to the first one after simplification, it does not provide any new information and is therefore redundant.

Similarly, edit rule t can be written as a lower bound on x_g :

$$x_g \geq \frac{-1}{a_{tg}} (a_{t1}x_1 + \dots + a_{t,g-1}x_{g-1} + a_{t,g+1}x_{g+1} + \dots + a_{tn}x_n + b_t).$$

Combining the two bounds and removing x_g , we obtain the implicit condition

$$\begin{aligned} & \frac{-1}{a_{sg}} (a_{s1}x_1 + \dots + a_{s,g-1}x_{g-1} + a_{s,g+1}x_{g+1} + \dots + a_{sn}x_n + b_s) \\ & \geq \frac{-1}{a_{tg}} (a_{t1}x_1 + \dots + a_{t,g-1}x_{g-1} + a_{t,g+1}x_{g+1} + \dots + a_{tn}x_n + b_t) \end{aligned}$$

which can be written in the general form (1) as

$$a_1^*x_1 + \dots + a_n^*x_n + b^* \geq 0,$$

with $a_i^* = a_{tg}a_{si} - a_{sg}a_{ti}$ ($i=1, \dots, n$) and $b^* = a_{tg}b_s - a_{sg}b_t$. This is an essentially new implied edit that is derived from the pair of inequality edits (s, t) , unless it happens to be redundant (see footnote 3).

It should be noted that, both for equalities and inequalities, the essentially new implied edit generated by this procedure does not involve the generating variable (i.e., the coefficient $a_g^* = 0$). This is in fact the defining property that makes an implied edit ‘essentially new’: it adds information to the existing edit rules by eliminating one of the variables.

According to the method of Fellegi and Holt (1976), a complete set of edits may be constructed by repeatedly applying the above-mentioned procedure of generating essentially new implied edits, using all variables in turn as generating variables, until no more (non-redundant) new edits can be derived. At that point, a complete set of edits has been generated.

Having obtained a complete set of edits, one can solve the error localisation problem for any given record in the following manner:

- Select all edits from the complete set of edits that are failed by the original record.
- Find the smallest (weighted) subset of the variables with the property that each selected (original or implied) edit involves at least one of them.

The first step amounts to evaluating the edits for a given record. The second step entails solving a set-covering problem, which is a well-known mathematical problem for which standard algorithms are available (see, e.g., Nemhauser and Wolsey, 1988). We shall work out a small example with the Fellegi-Holt method in Section 4.

A crucial element of the Fellegi-Holt method is the fact that a complete set of edits is ‘sufficiently large’ to reduce the error localisation problem to a set-covering problem. For a proof of this fact, see Fellegi and Holt (1976). For an explanation of what is meant by ‘sufficiently large’ from the viewpoint of logic, see Boskovitz et al. (2005).

The method discussed in this section works for numerical variables, but an analogous method exists for categorical variables. The only difference lies in the procedure for generating essentially new

implied edits. We refer to Fellegi and Holt (1976) and De Waal et al. (2011) for a description of the Fellegi-Holt method for categorical data.

2.5 *Solving the error localisation problem: other methods*

An important drawback of the method of Fellegi and Holt discussed in Section 2.4 is that the complete set of edits can be extremely large, especially with numerical data. In many practical applications, generating a complete set of edits is simply not technically feasible.⁴ For this reason, other algorithms have been developed that solve the error localisation problem without generating a complete set of edits. We can distinguish several classes of such algorithms.

Algorithms based on vertex generation

It is known from the literature that the optimal solution to the error localisation problem for a given record always corresponds with one of the vertices of an appropriately defined polyhedron; see, e.g., Theorem 3.1 in De Waal et al. (2011). Hence, in principle, the error localisation problem can be solved by generating all vertices of that polyhedron and identifying the optimal one. This approach has been elaborated in several error localisation algorithms. See, among others, Sande (1978), Kovar and Whitridge (1990), Fillion and Schiopu-Kratina (1993), Todaro (1999), and De Waal (2003). Tools for automatic editing that use algorithms based on vertex generation include GEIS (Kovar and Whitridge, 1990), Banff (Banff Support Team, 2008), CherryPi (De Waal, 1996), and AGGIES (Todaro, 1999).

Branch-and-bound algorithm

De Waal and Quere (2003) describe how the error localisation problem may be solved by means of a branch-and-bound algorithm. For a record containing n numerical variables, there are 2^n potential solutions, since each variable is either fixed to its original value or imputed. Basically, the branch-and-bound algorithm systematically considers all potential solutions and checks which of these are feasible. In order to do this, the algorithm generates relevant essentially new implied edits ‘on the fly’, but it does not construct a complete set of edits. Finally, the algorithm selects the feasible solution with the smallest sum of reliability weights. A similar branch-and-bound algorithm can be used for categorical or mixed data. We refer to De Waal and Quere (2003), De Waal (2003), and De Waal et al. (2011) for more details. Tools for automatic editing that use the branch-and-bound algorithm include SLICE (De Waal, 2005) and the R package `editrules` (De Jonge and Van der Loo, 2011).

Algorithms based on cutting planes

With this approach, to solve the error localisation problem for a given record, one starts by finding the minimal subset of the variables that ‘covers’ all original edit rules that are failed. As we have seen above, this solution may be infeasible. In that case, the algorithm generates new constraints, so-called

⁴ One exception occurs when all edit rules are ratio edits: it can be shown that, for a data set with n variables, the complete set of edits contains at most $n(n-1)/2$ non-redundant ratio edits. Thus, for ratio edits, the Fellegi-Holt method is usually feasible; see Winkler and Draper (1997).

cutting planes, and adds these to the original set of edit rules. Next, a minimal covering set of variables is determined for the new problem. Again, this solution may be infeasible, in which case more cutting planes need to be generated. In this iterative manner, the algorithm continues until it finds a feasible solution to the error localisation problem. For more details, we refer to Garfinkel et al. (1988), Ragsdale and McKeown (1996), and De Waal et al. (2011).

Algorithms for mixed integer programming

Finally, it is also possible to formulate the error localisation problem according to the Fellegi-Holt paradigm as a mixed integer programming problem; see, e.g., Riera-Ledesma and Salazar-González (2003). This type of problem can be solved by commercially available solvers.

De Waal and Coutinho (2005) compared the performance of several different algorithms for error localisation. They did not find a strong preference for one particular algorithm. Note that ‘performance’ here refers simply to computational efficiency. All of the above algorithms try to solve the same error localisation problem and hence, in theory, should find the same solution.⁵

3. Preparatory phase

The method discussed in this module is only considered appropriate for identifying random errors. Therefore, it is important to treat systematic errors, such as unit of measurement errors, before applying this method. Methods for detecting and treating systematic errors are discussed in the method module “Statistical Data Editing – Deductive Editing”.

In addition, automatic editing is usually applied in combination with a form of selective editing: the most influential errors are edited manually by subject-matter experts, while the other, non-influential errors are resolved automatically. Selective editing and manual editing are discussed in the theme module “Statistical Data Editing – Selective Editing” and the method module “Statistical Data Editing – Manual Editing”, respectively. We refer to “Statistical Data Editing – Main Module” for a discussion on how to combine different editing methods into one editing process. See also Pannekoek and De Waal (2005) for suggestions on how to set up an automatic editing strategy in practice.

4. Examples – not tool specific

To illustrate the method of Fellegi and Holt discussed in Section 2.4, we work out an example based on Fellegi and Holt (1976). In this example, there are four numerical variables. We do not use different reliability weights. The original set of edit rules consists of two edits:

$$x_1 - x_2 + x_3 + x_4 \geq 0 \tag{3}$$

and

$$-x_1 + 2x_2 - 3x_3 \geq 0. \tag{4}$$

⁵ In practice, the error localisation problem according to the Fellegi-Holt paradigm may have several equivalent optimal solutions, particularly if many variables have the same reliability weight. When this occurs, different implementations of these algorithms may differ in the way they choose between equivalent solutions.

By a repeated application of Fourier-Motzkin elimination, it is possible to derive the following essentially new implied edits from (3) and (4):

$$x_2 - 2x_3 + x_4 \geq 0, \quad (5)$$

$$x_1 - x_3 + 2x_4 \geq 0, \quad (6)$$

and

$$2x_1 - x_2 + 3x_4 \geq 0. \quad (7)$$

It is not possible to generate more essentially new implied edits from (3)–(7), so these five edit rules together constitute a complete set of edits. This means that we can now solve the error localisation problem for any record by solving an appropriate set-covering problem.

Consider the record $(x_1, x_2, x_3, x_4) = (3, 4, 6, 1)$. By checking the edit rules (3)–(7), it is seen that this record fails edits (4), (5), and (6). Thus, in order to solve the error localisation problem, we have to find the minimal subset of variables that ‘covers’ these three edit rules. By inspection, we see that the variable x_3 is involved in edit rules (4), (5), and (6). Thus, in this example, x_3 can be imputed to satisfy all the edit rules. Since $\{x_3\}$ is the only single-variable set with this property, changing the value of x_3 is in fact the optimal solution to the error localisation problem for this record. [Note that the single-variable sets $\{x_1\}$ and $\{x_2\}$ cover the original failed edit (4), but not the implied failed edits (5) and (6).] A consistent record can be obtained by imputing, for instance, the value $x_3 = 1$.

5. Examples – tool specific

The R package `editrules`, which can be downloaded for free at <http://cran.r-project.org>, contains an implementation of the branch-and-bound algorithm of De Waal and Quere (2003). To illustrate the use of `editrules` for automatic editing, we work out the example from Section 4 in R code.⁶

First, we load the package:

```
> library(editrules)
```

Next, we create an object of type “editmatrix” containing the two original edit rules:

```
> E <- editmatrix(c("x1-x2+x3+x4 >= 0", "-x1+2*x2-3*x3 >= 0"))
```

We also have to read in the record that we want to edit as a data frame:

```
> x <- data.frame(x1 = 3, x2 = 4, x3 = 6, x4 = 1)
```

Now, the error localisation problem is solved to optimality by giving the following command:

```
> le <- localizeErrors(E, x)
```

This command runs the branch-and-bound algorithm to solve the error localisation problem and stores the results in a new object called `le`. The results can be inspected by calling attributes of this object.

⁶ Version 2.5 of the `editrules` package was used to run the code in this example.

```
> le$status
  weight degeneracy user system elapsed maxDurationExceeded
1      1           1 0.05      0      0.13                FALSE
```

The attribute `le$status` contains background information on the performance of the algorithm. In this example, an optimal solution has been found with the sum of the reliability weights equal to 1 (as can be seen in the column `weight`). Since we have not specified the reliability weights in this example, R has used the default choice: all weights equal to 1. Other reliability weights can be specified by providing the function `localizeErrors` with an optional argument `weight`. The entry ‘1’ in the column `degeneracy` in `le$status` shows that the optimal solution is unique.

To see which variables have to be changed according to the optimal solution, we inspect the attribute `le$adapt`.

```
> le$adapt
      x1      x2      x3      x4
1 FALSE FALSE TRUE  FALSE
```

This command prints a boolean data frame with the value ‘TRUE’ for variables that have to be changed, and the value ‘FALSE’ for the other variables. In this example, the optimal solution is to change only the value of variable x_3 . This solution is identical to the one found in Section 4 by applying the method of Fellegi and Holt.

We refer to De Jonge and Van der Loo (2011) for more details on the `editrules` package.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Banff Support Team (2008), Functional Description of the Banff System for Edit and Imputation. Technical Report, Statistics Canada.
- Boskovitz, A., Goré, R., and Wong, P. (2005), Data Editing and Logic. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ottawa.
- Casado Valero, C., Del Castillo Cuervo-Arango, F., Mateo Ayerra, J., and De Santos Ballesteros, A. (1996), Quantitative Data Editing: Quadratic Programming Method. Presented at the COMPSTAT 1996 Conference, Barcelona.
- De Jonge, E. and van der Loo, M. (2011), Manipulation of Linear Edits and Error Localization with the Editrules Package. Discussion Paper 201120, Statistics Netherlands, The Hague.
- De Waal, T. (1996), CherryPi: a Computer Program for Automatic Edit and Imputation. Working Paper, UN/ECE Work Session on Statistical Data Editing, Voorburg.
- De Waal, T. (2003), *Processing of Erroneous and Unsafe Data*. PhD Thesis, Erasmus University, Rotterdam.

- De Waal, T. (2005), SLICE 1.5: a Software Framework for Automatic Edit and Imputation. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ottawa.
- De Waal, T. and Coutinho, W. (2005), Automatic Editing for Business Surveys: an Assessment for Selected Algorithms. *International Statistical Review* **73**, 73–102.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- De Waal, T. and Quere, R. (2003), A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics* **19**, 383–402.
- Di Zio, M., Guarnera, U., and Luzi, O. (2005), Improving the Effectiveness of a Probabilistic Editing Strategy for Business Data. Report, ISTAT, Rome.
- Fellegi, I. P. and Holt, D. (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.
- Fillion, J. M. and Schiopu-Kratina, I. (1993), On the Use of Chernikova's Algorithm for Error Localization. Report, Statistics Canada.
- Freund, R. J. and Hartley, H. O. (1967), A Procedure for Automatic Data Editing. *Journal of the American Statistical Association* **62**, 341–352.
- Garfinkel, R. S., Kunnathur, A. S., and Liepins, G. E. (1988), Error Localization for Erroneous Data: Continuous Data, Linear Constraints. *SIAM Journal on Scientific and Statistical Computing* **9**, 922–931.
- Ghosh-Dastidar, B. and Schafer, J. L. (2003), Multiple Edit/Multiple Imputation for Multivariate Continuous Data. *Journal of the American Statistical Association* **98**, 807–817.
- Hoogland, J. and Smit, R. (2008), Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.
- Kovar, J. and Whitridge, P. (1990), Generalized Edit and Imputation System; Overview and Applications. *Revista Brasileira de Estadística* **51**, 85–100.
- Little, R. J. A. and Smith, P. J. (1987), Editing and Imputation of Quantitative Survey Data. *Journal of the American Statistical Association* **82**, 58–68.
- Nemhauser, G. L. and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization*. John Wiley & Sons, New York.
- Nordbotten, S. (1963), Automatic Editing of Individual Statistical Observations. In: *Conference of European Statisticians Statistical Standards and Studies No. 2*, United Nations, New York.
- Pannekoek, J. and de Waal, T. (2005), Automatic Edit and Imputation for Business Surveys: the Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* **21**, 257–286.
- Ragsdale, C. T. and McKeown, P. G. (1996), On Solving the Continuous Data Editing Problem. *Computers & Operations Research* **23**, 263–273.
- Riera-Ledesma, J. and Salazar-González, J. J. (2003), New Algorithms for the Editing and Imputation Problem. Working Paper, UN/ECE Work Session on Statistical Data Editing, Madrid.

- Sande, G. (1978), An Algorithm for the Fields to Impute Problems of Numerical and Coded Data. Technical Report, Statistics Canada.
- Scholtus, S. (2013), Automatic Editing with Hard and Soft Edits. *Survey Methodology* **39**, 59–89.
- Todero, T. A. (1999), Overview and Evaluation of the AGGIES Automated Edit and Imputation System. Working Paper, UN/ECE Work Session on Statistical Data Editing, Rome.
- Williams, H. P. (1986), Fourier's Method of Linear Programming and Its Dual. *The American Mathematical Monthly* **93**, 681–695.
- Winkler, W. E. and Draper, L. A. (1997), The SPEER Edit System. In: *Statistical Data Editing*, Volume 2: *Methods and Techniques*, United Nations, Geneva.

Specific section

8. Purpose of the method

Localising errors in microdata without human intervention

9. Recommended use of the method

1. The method should be used for error localisation in microdata containing only random errors. Any systematic errors that may occur in the original microdata have to be resolved beforehand, using deductive editing methods (see the method module “Statistical Data Editing – Deductive Editing”).
2. If it is known beforehand that certain variables contain more errors than others, then this information should be included in the form of reliability weights (see item 14).
3. The quality of the error localisation strongly depends on the specification of the edit rules. The set of edit rules should be sufficiently powerful to detect the majority of errors, but not so strict that the method results in overedited data.

10. Possible disadvantages of the method

1. In general, it is not possible to construct a set of edit rules that always leads to the correct solution. Thus, the edited data may still contain some errors, although the edited records are consistent with the edit rules. For this reason, automatic editing should not be applied to crucial records, e.g., records belonging to very large businesses. In addition, the quality of automatic editing is lower for records that contain many errors. Both disadvantages can be circumvented by always using automatic editing in combination with a form of selective editing. We refer to “Statistical Data Editing – Main Module” for a discussion on how to incorporate automatic editing in an overall editing strategy.

11. Variants of the method

1. The original method of Fellegi and Holt as described in Section 2.4.
2. Other methods as described in Section 2.5. These methods find the same solution as the original method of Fellegi and Holt, but they use different search algorithms. Examples include:
 - 2.1 Algorithms based on vertex generation;
 - 2.2 Algorithms based on branch-and-bound;
 - 2.3 Algorithms based on cutting planes;
 - 2.4 Algorithms based on (mixed) integer programming.

12. Input data

1. A data set containing unedited microdata.

13. Logical preconditions

1. Missing values
 1. Allowed; they will be considered as erroneously missing, i.e., available for imputation.
2. Erroneous values
 1. Allowed; in fact, the object of this method is to decide which values in a record are erroneous.
 2. It is assumed that the data contain only random errors; systematic errors should be removed beforehand by means of deductive editing.
3. Other quality related preconditions
 1. n/a
4. Other types of preconditions
 1. It is assumed that all edit rules may be interpreted as hard edit rules.

14. Tuning parameters

1. A collection of edit rules for the microdata at hand.
2. A set of reliability weights may be provided for the variables in the data set. By default, all reliability weights are equal to 1.
3. A maximum number of variables to impute may be set to reduce the computational workload. The error localisation problem will not be solved for records that cannot be imputed consistently by changing at most the specified maximum number of variables.

15. Recommended use of the individual variants of the method

1. For variant 1 (the original method of Fellegi and Holt), most of the work lies in the generation of a complete set of edits. Once this complete set is available, the error localisation problem can be solved for any record in a straightforward manner. If the complete set of edits is too large to be generated, this variant of the method cannot be used.
2. For the other variants, the work lies in solving a separate error localisation problem for each individual record. In this case, it is usually necessary to specify a maximum number of variables to impute (see item 14), unless the data set contains few variables (say less than 10).

16. Output data

1. For each record in the microdata, the method attempts to yield a list of variables that can be imputed to obtain a consistent record with respect to the edit rules. For some records, the method may not return such a list, because it could not find a feasible solution to the error localisation problem.

17. Properties of the output data

1. For each record for which the method returns a solution, the variables listed in the solution can be imputed so that the resulting record is consistent with respect to the edit rules. Moreover, they constitute the smallest (weighted) set of variables that has this property.
2. The original values of the variables that are listed for imputation have to be considered as erroneous in all further processing. The natural next step is to impute new values for these variables by means of some imputation method. It should be noted that the imputation step is not a part of the error localisation method itself.
3. For some records, the method may not find a solution. These records have to be processed interactively by subject-matter experts (see the method module “Statistical Data Editing – Manual Editing”).

18. Unit of input data suitable for the method

Incremental processing by record

19. User interaction - not tool specific

1. Ideally, there is no user interaction other than setting parameters and reading in input data at the beginning, and processing output data at the end.

20. Logging indicators

1. The number of records for which the method found/did not find a solution.
2. The computing time per record.

21. Quality indicators of the output data

1. The quality of automatic editing can be assessed in a simulation study. This requires a data set that has been interactively edited by experts to a point where the edited data may be considered error-free. In the simulation study, the original data are edited again using automatic editing. The quality of automatic editing may then be measured in terms of the similarity of the automatically edited data to the interactively edited data.

22. Actual use of the method

1. The method is used at Statistics Netherlands in the production process for structural business statistics. This application uses the tool SLICE, which contains an implementation of the branch-and-bound algorithm of De Waal and Quere (2003). See Hoogland and Smit (2008) for more details.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Main Module
2. Statistical Data Editing – Main Module

3. Statistical Data Editing – Selective Editing
4. Imputation – Main Module
5. Macro-Integration – Main Module

24. Related methods described in other modules

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Statistical Data Editing – Deductive Editing
3. Statistical Data Editing – Manual Editing

25. Mathematical techniques used by the method described in this module

1. Fourier-Motzkin elimination

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1. GEIS
2. Banff
3. CherryPi
4. AGGIES
5. SLICE
6. R package `editrules`

28. Process step performed by the method

Statistical data editing

Administrative section

29. Module code

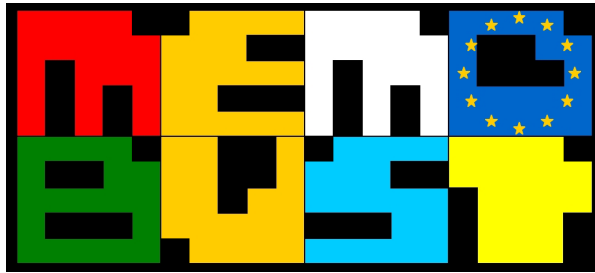
Statistical Data Editing-M-Automatic Editing

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	20-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	20-04-2012	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.2.1	16-07-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.3	04-09-2013	minor improvements	Sander Scholtus	CBS (Netherlands)
0.3.1	09-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:11



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Manual Editing

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Introduction and historical notes	3
2.2 The use of recontacts	4
2.3 Potential problems	5
3. Preparatory phase	6
3.1 The editing staff.....	6
3.2 Editing instructions.....	6
3.3 Error messages	7
3.4 Efficient edit rules for manual editing.....	7
4. Examples – not tool specific.....	9
5. Examples – tool specific.....	10
6. Glossary.....	10
7. References	10
Specific section.....	12
Interconnections with other modules.....	14
Administrative section.....	15

General section

1. Summary

In manual editing, records of microdata are checked for errors and, if necessary, adjusted by a human editor, using expert judgement. Nowadays, the editor is usually supported by a computer program in identifying data items that require closer inspection – in particular combinations of values that are inconsistent or suspicious. Moreover, the computer program enables the editor to change data items interactively, meaning that the automatic checks that identify inconsistent or suspicious values are immediately rerun whenever a value is changed. This modern form of manual editing is often referred to as ‘interactive editing’.

If organised properly, manual/interactive editing is expected to yield high quality data. However, it is also time-consuming and labour-intensive. Therefore, it should only be applied to that part of the data which cannot be edited safely by any other means, i.e., some form of selective editing should be applied (see “Statistical Data Editing – Selective Editing”). Furthermore, it is important to use efficient edit rules and to draw up detailed editing instructions in advance.

2. General description of the method

2.1 Introduction and historical notes

Manual editing is the traditional way to perform data editing. Other data editing methods, in particular automatic editing techniques, did not emerge until the 1960s, and their application has only become widespread from the 1980s onward. Even today, practically all surveys at statistical offices and elsewhere include some form of manual editing. Manual editing is in fact widely viewed as an essential part of any data editing process.

Ideally, a person who performs manual editing – an *editor* – should be an expert who has extensive knowledge of the survey subject, the survey population, and the kind of errors that are likely to occur in the survey data. If necessary, he or she may recontact a respondent to check whether a suspicious value is correct, or to obtain a new value for a data item that was originally missing or incorrect. The editor may compare a survey unit’s data to reference data, such as data on the same unit from a previous survey or from an external register, or data on similar units. Finally, he or she may have access to other sources of information, for instance through internet searches.

In its ideal form, manual editing is expected to yield high quality data. In particular, it should lead to better results than automatic editing. However, it should be clear that the quality of manual editing depends strongly on the competence and training of the available editors. In certain less than ideal situations, the quality of manually edited data need not be significantly higher than that of automatically edited data, and it may even be lower (EDIMBUS, 2007).

Traditionally, manual editing was performed directly on the original paper questionnaires. Later, mainframe computers were used to check the data for inconsistencies and other violations of edit rules. To this end, the information on the questionnaires first had to be keyed in by typists. A list of edit failures identified by the computer was printed out on paper and used by the editors as a guide for making manual adjustments on the original questionnaires. When all questionnaires had been edited, the adjusted data were re-entered into the mainframe computer by typists and the edit checks were run

again, to see the effect of the proposed adjustments on the edit failures. Often, the automated checks revealed that the adjusted data still failed some of the edit rules, and another round of manual editing was required. It was not unusual that five, ten, or even more iterations of automatic checking and manual adjusting were needed before all questionnaires were considered sufficiently edited (Granquist, 1997; Van de Pol, 1995).

The advent of the microcomputer in the 1980s made it possible to integrate automatic checking and manual treatment of errors, thereby improving the data editing process in several ways (Bethlehem, 1987). From now on, the information on the questionnaires had to be keyed in only once.¹ After that, all adjustments could be made by the editors directly on the captured data. This obviously benefited the efficiency and timeliness of the editing process. A second improvement was that the editors could now get immediate feedback on the adjustments they made, because the automatic edit checks could be rerun instantaneously whenever the value of a data item was changed. This made it much easier for them to find adjustments that satisfied the edit rules. In addition, each record/questionnaire could now be edited separately, by one editor, until all violations of edit rules had been either removed or explained. This improved form of manual editing is called *interactive editing*.

Interactive editing requires a survey-processing system that provides the above-mentioned interaction between automated checks and manual adjustments. Well-known examples of survey-processing systems are *Blaise* (see, e.g., Blaise, 2002) and *CSPro* (see, e.g., CSPro, 2008). Pierzchala (1990) discusses general requirements of computer systems for interactive editing.

In today's statistical practice, interactive editing has effectively replaced all older forms of manual editing. Hence, the terms 'manual editing' and 'interactive editing' have become more or less interchangeable. In the remainder of this module, they shall be used as synonyms.

2.2 The use of recontacts

In the previous subsection, possible actions were listed that an editor may take when confronted with a record that requires review. One of these possible actions is recontacting the respondent. At first glance, a recontact may appear to be the natural way of obtaining better values for data items that were reported erroneously during the original field work, as well as items that were originally missing. Actually, depending on the survey, it may not be possible to contact the original respondents. For instance, if an external register is used as a data source and questions are raised about the quality of the incoming data, then the statistical office can usually only contact the supplier of the data set. Direct contact with the individual entities in the register is usually not possible in this case.

However, even when recontacts are possible, this approach can be considered problematic for several reasons. First of all, recontacts clearly increase the burden on respondents, whereas many statistical institutes are trying to reduce the response burden. In addition, recontacts tend to slow down the editing process and can therefore adversely influence the timeliness of statistics. Finally, if one considers that a respondent was not able to give a correct answer in the original survey – supposedly while filling in a meticulously designed questionnaire or talking to a highly qualified interviewer –,

¹ A more recent development is that data often arrive at the statistical office already in digital form, so that no keying is necessary at all. This is true for nearly all registers and for electronic questionnaires. For a discussion of the implications of electronic data collection for the editing process, see the theme module "Questionnaire Design – Editing During Data Collection".

then it is not at all obvious that he/she will give the correct response when talking to an editor. According to EDIMBUS (2007): "...respondents' ability to report should not be overestimated. In fact, if the structure of the questions does not fit their understanding, no amount of badgering will get the 'correct' answers out of them."

Following Granquist (1997), if recontacts are used during interactive editing, their main purpose should be to reveal problems that *cause* respondents to give erroneous answers, rather than merely correcting the individual errors that occurred. When used this way, recontacts can provide important insights into respondents' behaviour – in particular their ability to understand the concepts and definitions used in the survey. They may also reveal differences between what is asked in the survey and what kind of information is readily available in the survey units' accounting systems. These insights may be used as a basis for improvements at the data collection stage in subsequent surveys (see, e.g., Hartwig, 2009; Svensson, 2012).

2.3 Potential problems

There are several potential problems associated with interactive editing. The most important of these are the risks of *overediting* and *creative editing*.

According to Granquist (1995), overediting occurs when "the share of resources and time dedicated to editing is not justified by the resulting improvements in data quality." Manual editing is in fact a very labour-intensive and time-consuming activity, even in its modern, interactive form. Moreover, statistical output is typically affected by all kinds of errors (Bethlehem, 2009), including sampling error, selective unit non-response, coverage errors, measurement errors, etc. Only a subset of these can be treated during data editing: in particular, measurement and processing errors and, to a lesser extent, errors in the survey frame. Therefore, as soon as the data have been edited to a point where the influence of the latter types of errors on the statistical output is negligible compared to other sources of error (e.g., the sampling variance), manual editing should be stopped to prevent overediting. This notion – which was suggested already by Nordbotten (1955) – has received much attention since the 1980s. It has led to the development of methods for selective editing (see the theme module "Statistical Data Editing – Selective Editing") and macro-editing (see the theme module "Statistical Data Editing – Macro-Editing").

Another aspect of overediting is that if the editing process is continued too long, it may actually start to do more harm than good. In general, not all values that appear to be implausible are also incorrect. Hence, replacing all unusual combinations of values by more plausible ones would lead to a data set that does not reflect the natural variability of characteristics in the population. Overediting may therefore adversely influence the quality of the statistical output. An important part of the 'art' of manual editing is understanding which implausible values to adjust and which to leave as they are. This requires expert judgement and, in some cases, a recontact.

A second potential problem is the risk of creative editing: editors inventing their own, often highly subjective, editing procedures. Creative editing often involves complex adjustments of reported data items, done for the sole purpose of making the data consistent with a set of edit rules. Granquist (1995) remarks that creative editing may "hide serious data collection problems and give a false impression of respondents' reporting capacity."

To reduce the risk of overediting and creative editing, it is important to design efficient edit rules and to provide the editors with good editing instructions. These issues are discussed in the next section.

3. Preparatory phase

In this section, several issues will be discussed that are related to the design of manual editing. These are: the desired characteristics of the editing staff (Section 3.1); the use of editing instructions to rationalise the manual editing process (Section 3.2); the design of error messages (Section 3.3); the design of efficient edit rules for manual editing (Section 3.4).

3.1 The editing staff

As mentioned in Section 2.1, the quality of manual editing strongly depends on the competence of the individual editors that are involved. A good editor should have the following characteristics:

- He/she has a large knowledge of the survey subject and of survey methodology. Since most of this knowledge is rather specialised, it has to be acquired through experience and training.
- He/she is communicative and responsive. This is particularly important if recontacts are used. Granquist (1995) remarked that if recontacts are done by telephone, “the editors also become telephone interviewers, needing adequate training and monitoring as in regular telephone interview surveys.”
- He/she is responsible and able to work accurately.
- Preferably, he/she should have an analytical mind, with an interest in problem-solving.

3.2 Editing instructions²

Editing instructions are an important aid in rationalising the manual editing process. They should contain at least the following components:

- A description of the purpose of the survey and the intended statistical output. In addition, the data collection phase and relevant data processing steps prior to editing should be briefly described.
- If relevant, instructions on the order in which the selected records should be treated. If manual editing is used in combination with selective editing (see “Statistical Data Editing – Selective Editing”), then an explanation is needed about the selection criteria and their interpretation. If manual editing is used in combination with macro-editing (see “Statistical Data Editing – Macro-Editing”), then detailed analysis instructions are needed regarding the selection of individual records that need further review.
- An overview of the types of errors that can occur in the data. Common errors in business surveys include classification errors with respect to NACE code or size class (i.e., errors in the survey frame), measurement errors, and processing errors.
- Suggestions about additional sources of information – such as auxiliary registers, sector organisations, and the internet – which should be consulted when following up data that have

² This subsection is to a large extent based on Hoogland et al. (2011).

been flagged by edit rules (see below). For example, many businesses nowadays have websites that contain relevant information for verifying potential NACE code errors.

- For each common type of error, an indication of how the error can be treated. Deterministic correction rules may often be specified for treating systematic errors (see also ‘Deductive Editing’). Clear instructions on this point can prevent the occurrence of creative editing.
- Instructions on how to log the editing actions taken during interactive editing. The survey-processing system should provide a comments field for this. Editors should be encouraged to provide details about the reasons for the adjustments they make. This information can be useful for improving the data collection process as well as the editing process itself.
- Instructions on specific follow-up actions that may be needed for certain types of errors. In particular, in case a NACE code or size class error is detected, it should be clear whether and how this must be communicated to the administrator of the survey frame.

3.3 *Error messages*

As mentioned in the main theme module, an important technique for finding errors in microdata is the inspection of data items that fail *edit rules*. Edit rules (edits for short) describe restrictions that should be satisfied by the data. Edits can be hard (meaning that they have to hold by definition, so that any failure corresponds to an error in the data) or soft (meaning that they are expected to hold for most survey units, but they can sometimes be failed by correct data items).

When edit rules are implemented in a computer system, an error message has to be associated with each edit rule. This message contains the information that the computer system gives to the editor about the unit and variables that are flagged by the edit rule as being (suspected to be) in error. The purpose of the error message is to give sufficient information for a rational follow-up of error flags. It also forms a basis for (process) data about the data collection and production processes.

The content of an error message generally consists of:

- Identifying properties of the flagged unit.
- The name of the flagged variable(s). For the purpose of manual editing, this should be a descriptive name rather than a technical one; e.g., not *TURNOVE100000* but *Total net turnover from domestic sales*.
- The code of the edit rule that was failed.
- A verbal description of the edit rule that was failed or, equivalently, a verbal description of the suspected error.
- If relevant and available, suggestions for auxiliary data that may be consulted in a follow-up of the error flag.

3.4 *Efficient edit rules for manual editing*

Typically, a large part of the work done during manual editing concerns the follow-up of soft edit failures. For this reason, it is important to formulate soft edit rules that are as efficient as possible. Here, an edit rule is considered efficient to the extent that it detects suspected errors that turn out to be

actual errors during manual follow-up, and inefficient to the extent that it detects suspected errors that turn out to be correct. (A measure of efficiency known as the hit rate will be introduced below.)

According to Norberg (2011), most edits that are used in practice consist of three components: an *edit group*, a *test variable*, and an *acceptance region*. The edit group defines the subset of the units to which the edit should be applied. The test variable is a known function of the observed variables that is evaluated by the edit. Finally, the acceptance region describes for which values of the test variable the edit will be satisfied. (Equivalently, one could define a *rejection region* that describes for which values of the test variable the edit will be failed.) Using these components, an edit may be written in one of the general forms

if (*unit* \in *edit group*) **then** (*test variable* \in *acceptance region*)

or

if (*unit* \in *edit group* **and** *test variable* \notin *acceptance region*) **then** *error*.

Both formulations are equivalent. Human editors often find it slightly easier to work with the first formulation (Van de Pol, 1995). In a computer implementation, the second formulation can easily be extended to associate a unique error code and error message to each edit rule.

For a simple example, consider the following edit rule:

if *Size class* = 'small' **then** $0 \leq \text{Number of employees} < 10$.

For this edit, the edit group can be defined as "all units for which *Size class* = 'small'". The test variable is identical to one of the observed variables, *Number of employees*. The acceptance region consists of the interval [0, 10). A computer implementation of this edit could further specify the following actions:

if (*Size class* = 'small' **and** (*Number of employees* < 0 **or** *Number of employees* \geq 10))
then (*error_code_E1* := "failed";
error_message_E1 := "The number of employees does not match the size class.")

The first statement in the then-part assigns the error code "failed" to the current record for this edit (identified here by E1). The second statement gives an error message describing the nature of the current edit failure to the human editor. Of course, the precise implementation of these actions will depend on the survey-processing system.

To give another example, consider the following conditional ratio edit:

if (*Economic activity* = X **and** *Size class* = 'medium')
then $a < \text{Total turnover} / \text{Number of employees} < b$.

Here, the edit group consists of "all medium-sized units with *Economic activity* X", the test variable is defined as the ratio of the observed variables *Total turnover* and *Number of employees*, and the acceptance region is given by the interval (a,b).

Norberg (2011) notes that, for the editing to be efficient, one should choose edit groups that are homogeneous with respect to the test variable. In some cases, the choice of an edit group may be natural (e.g., the first example given above). If this is not the case, suitable edit groups may be derived from an analysis of previously edited data. Norberg (2012) suggests to use classification or regression trees for this. In addition, the acceptance region should reflect the natural variability of the test

variable within the edit group (Norberg, 2012). Again, previously edited data may be analysed (e.g., using box plots) to find suitable acceptance regions. It may be worthwhile to transform a test variable so that its distribution becomes more amenable to summary in the form of an acceptance region (e.g., so that the transformed test variable is approximately normally distributed, or at least symmetrical). Moreover, in repeated surveys, the acceptance regions should be regularly updated.

Outlier detection techniques are often used in the construction of soft edit rules. We refer to EDIMBUS (2007) for a discussion of outlier detection in the context of statistical data editing. Methods that may be used to construct soft edit rules in repeated surveys are discussed in the theme module “Statistical Data Editing – Editing for Longitudinal Data”.

At the design stage, it is useful to assess the efficiency and effectiveness of a proposed set of edits E by means of simulation. This requires historical data that have been fully edited, as well as the original, unedited version of the same data set. Interesting indicators for an edit $e \in E$ include the *failure rate* (the proportion of records in the unedited data that fail edit e) and the *hit rate* (the proportion of edit failures with respect to e in the unedited data that are associated with adjustments in the edited data). Note that for all hard edit rules, the hit rate should be 1. These indicators are local, i.e., defined for one edit at a time. Similar global indicators can be defined for the set of edits E as a whole. It is also interesting to assess to what extent the edits are ‘overlapping’, in the sense that the same error is often detected by multiple edits. Ideally, there should be as little overlap as possible between the edits.

Furthermore, making the assumption that the edited historical data do not contain any errors, one can evaluate the *missed error rate* (the proportion of errors in the original data that were not flagged by any edits in E) and an estimate of the measurement bias due to untreated errors if editing were based on E . See EDIMBUS (2007) and Silva et al. (2008) for formal definitions of these and other indicators.

The Office for National Statistics in the United Kingdom and Southampton University have developed a tool called Snowdon-X which “can be used to understand how current edits are working within the survey and also the impact on quality of any changes to the edit rules” (Skelterbery et al., 2011). Snowdon-X evaluates the indicators mentioned above as well as many other indicators. See Silva et al. (2008) for more details on Snowdon-X.

Note that the failure rate and hit rate of edits can and should be evaluated also during regular production. On the other hand, evaluating the missed error rate requires edited historical data. For repeated surveys, suitable historical data sets are available in theory, if not always in practice (Lindgren, 2012). For a one-off survey, as well as the first cycle of a survey that will be repeated, the situation is different. Often in this case, a small pilot study is conducted beforehand. The data from this study can be used to test the effects of different editing approaches, including experiments with different formulations of edit rules. In addition, experts should be consulted that have had experience with similar surveys in the past.

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bethlehem, J. G. (1987), The Data Editing Research Project of the Netherlands Central Bureau of Statistics. Report 2967-87-M1, Statistics Netherlands, Voorburg.
- Bethlehem, J. G. (2009), *Applied Survey Methods*. Wiley Series in Survey Methodology, John Wiley & Sons, New Jersey.
- Blaise (2002), *Blaise for Windows 4.5 Developer’s Guide*. Statistics Netherlands, Heerlen.
- CSPRO (2008), *CSPRO User’s Guide*, version 4.0. U.S. Census Bureau, Washington, D.C.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Granquist, L. (1995), Improving the Traditional Editing Process. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, 385–401.
- Granquist, L. (1997), The New View on Editing. *International Statistical Review* **65**, 381–387.
- Hartwig, P. (2009), How to Use Edit Staff Debriefings in Questionnaire Design. Paper presented at the 2009 European Establishment Statistics Workshop, Stockholm.
- Hoogland, J., van der Loo, M., Pannekoek, J., and Scholtus, S. (2011), *Data Editing: Detection and Correction of Errors*. Methods Series Theme, Statistics Netherlands, The Hague.
- Lindgren, K. (2012), The Use of Evaluation Data Sets when Implementing Selective Editing. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- Norberg, A. (2011), The Edit. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Norberg, A. (2012), Tree Analysis – A Method for Constructing Edit Groups. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- Nordbotten, S. (1955), Measuring the Error of Editing the Questionnaires in a Census. *Journal of the American Statistical Association* **50**, 364–369.
- Pierzchala, M. (1990), A Review of the State of the Art in Automated Data Editing and Imputation. *Journal of Official Statistics* **6**, 355–377.
- Silva, P. L. N., Bucknall, R., Zong, P., and Al-Hamad, A. (2008), A Generic Tool to Assess Impact of Changing Edit Rules in a Business Survey – An Application to the UK Annual Business Inquiry Part 2. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.

- Skentelbery, R., Finselbach, H., and Dobbins, C. (2011), Improving the Efficiency of Editing for ONS Business Surveys. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Svensson, J. (2012), Editing Staff Debriefings at Statistics Sweden. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- Van de Pol, F. (1995), Data Editing of Business Surveys: an Overview. Report 10718-95-RSM, Statistics Netherlands, Voorburg.

Specific section

8. Purpose of the method

Detecting and treating errors in microdata

9. Recommended use of the method

1. Because of its expensive and time-consuming nature, it is best to apply manual editing only to that part of the data where expert judgement is really needed. In other words, one should always try to use this method as part of a strategy for selective editing or macro-editing (cf. “Statistical Data Editing – Main Module”). This usually means that manual editing is only applied to units that are either very large or complex, or for which the reported data are likely to contain many and/or influential errors.
2. A survey-processing system should be used that allows real-time interaction between manual adjustments and automated checks (i.e., manual editing should be interactive editing)
3. It is important to draw up editing instructions in advance, to guide the decisions made by the editors during manual editing. This lowers the risk of overediting or creative editing. It is also important to design efficient edit rules and informative error messages.

10. Possible disadvantages of the method

1. If recontacts are used as part of manual editing, the method places additional burden on survey units that are recontacted. Recontacts may also affect the timeliness of statistical production.

11. Variants of the method

1. n/a

12. Input data

1. A data set containing unedited microdata.

13. Logical preconditions

1. Missing values
 1. Allowed.
2. Erroneous values
 1. Allowed; in fact, the object of this method is to replace erroneous values with better values.
3. Other quality related preconditions
 1. n/a
4. Other types of preconditions
 1. n/a

14. Tuning parameters

1. A collection of edit rules for the microdata at hand.

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. A data set containing edited microdata.

17. Properties of the output data

1. If manual editing has been performed correctly, the records in the output data set are consistent with all hard edit rules. In addition, all remaining soft edit failures have been explained and accepted by a subject-matter expert.

18. Unit of input data suitable for the method

Incremental processing

19. User interaction - not tool specific

1. As the term ‘interactive editing’ suggests, user interaction is needed throughout. In fact, all changes made to the data during manual/interactive editing are initiated by a human editor.

20. Logging indicators

1. Comments made by the editors to explain the adjustments they made to the data, as well as the soft edit failures that they left in.
2. If recontacts are used: comments made by the editors regarding identified problems that caused respondents to report erroneous values in the original survey.
3. Process indicators for the efficiency and effectiveness of the edit rules used in manual editing include: failure rate, hit rate, missed error rate, estimated measurement bias. See also Section 3.4 of this module, EDIMBUS (2007), and Silva et al. (2008).

21. Quality indicators of the output data

1. It is not straightforward to assess the quality of manually edited data, because in many applications the results of manual editing are actually taken as the standard by which other forms of editing are to be measured. Nordbotten (1955) suggests a way to measure the quality of regular manual editing, i.e., as it occurs in everyday statistical practice. This method takes a random sample of the original data and subjects it to a very refined form of manual editing (under ideal conditions, with near-unlimited resources). The quality of the regular editing process may then be measured in terms of the similarity of the data edited under regular conditions to the data edited under ideal conditions.

22. Actual use of the method

1. Interactive editing is used at Statistics Netherlands in many production processes, including that of the structural business statistics. The survey-processing system Blaise is used as a tool.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Questionnaire Design – Editing During Data Collection
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Statistical Data Editing – Editing for Longitudinal Data

24. Related methods described in other modules

1. Statistical Data Editing – Automatic Editing

25. Mathematical techniques used by the method described in this module

1. n/a

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1. Blaise
2. CSPro

Note: These tools support interactive editing, but – by its very nature – this method relies heavily on human interaction with the tool.

3. Snowdon-X

Note: This tool can be used to evaluate the efficiency of edit rules for manual editing.

28. Process step performed by the method

Statistical data editing

Administrative section

29. Module code

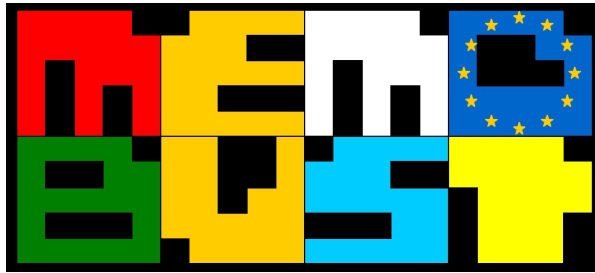
Statistical Data Editing-M-Manual Editing

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-06-2012	first version	Sander Scholtus	CBS (Netherlands)
0.2	01-03-2013	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	12-04-2013	improvements based on second Swedish review	Sander Scholtus	CBS (Netherlands)
0.4	11-11-2013	minor improvements based on final Swedish review	Sander Scholtus	CBS (Netherlands)
0.4.1	26-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:12



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Macro-Editing

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction to macro-editing.....	3
2.2 The aggregate method	4
2.3 The distribution method	6
3. Design issues	8
4. Available software tools.....	8
5. Decision tree of methods	8
6. Glossary.....	8
7. References	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

In most business surveys, it is reasonable to assume that a relatively small number of observations are affected by errors with a significant effect on the estimates to be published (so-called influential errors), while the other observations are either correct or contain only minor errors. For the purpose of statistical data editing, attention should be focused on treating the influential errors. *Macro-editing* (also known as *output editing* or *selection at the macro level*) is a general approach to identify the records in a data set that contain potentially influential errors. It can be used when all the data, or at least a substantial part thereof, have been collected.

Macro-editing has the same purpose as selective editing (see “Statistical Data Editing – Selective Editing”): to increase the efficiency and effectiveness of the data editing process. This is achieved by limiting the costly manual editing to those records for which interactive treatment is likely to have a significant effect on the quality of the estimates. The main difference between these two approaches is that selective editing selects units for manual follow-up on a record-by-record basis, whereas macro-editing selects units by considering all the data at once. It should be noted that in macro-editing all actual adjustments to the data take place at the *micro* level (i.e., for individual units), not the *macro* level. Methods that perform adjustments at the macro level are discussed in the topic “Macro-Integration”.

2. General description

2.1 Introduction to macro-editing

Macro-editing is a general approach to identify potentially influential errors in a data set for manual follow-up. It can be used when all the data, or at least a substantial part thereof, have been collected. In addition, the method is particularly effective when it is applied to data that contain only a limited number of large errors. Given these conditions, macro-editing is typically applied towards the end of a data editing process. At that stage, the errors that one expects to find in the data are either remaining errors that ‘slipped through’ previous editing efforts or errors that were actually introduced during data processing (processing errors). Possible sources of processing errors include automated data handling (e.g., loading the wrong data set, running an application with the wrong set of parameters, a bug in the software) as well as wrong decisions made by editors during manual editing. Macro-editing may succeed in finding these errors by examining the data from a macro rather than a micro level perspective – in other words, looking at the whole data set instead of one record at a time.

Macro-editing proceeds by computing aggregate values from a data set and systematically checking these aggregates for suspicious values and inconsistencies. The following types of checks are typically used:

- Internal consistency checks. In most business surveys, the definitions of the survey variables imply that the aggregated data should satisfy certain logical or mathematical restrictions. For instance, in each stratum, total net turnover (say X) should equal the sum of total net turnover from domestic sales (X_1) and total net turnover from foreign sales (X_2); i.e., it should hold that $X = X_1 + X_2$. In addition, based on subject-matter knowledge the fraction of total net

turnover from domestic sales may be expected to lie between certain bounds; i.e., $a < X_1 / X < b$ for certain constants a and b . These restrictions are the macro-level equivalents of edit rules that were used during micro-editing (see “Statistical Data Editing – Main Module”). Like edit rules, they may be either hard restrictions (identifying erroneous aggregates with certainty, such as the first example given above) or soft restrictions (identifying suspicious aggregates that may occasionally be correct, such as the second example).

- Comparisons with other statistics. It may be possible to compare aggregates to similar estimates from other data sources. If large differences occur, the corresponding aggregates are identified as suspicious. Such comparisons can be useful, if only to promote coherence between different statistical outputs. On the other hand, the comparability of aggregates from different sources is often affected in practice by conceptual and operational differences (e.g., different target populations, differences in variable definitions, different reference periods). It is important to be aware of these differences when they exist.
- Comparisons with previously published statistics. In repeated surveys, one can compare current aggregates to a time series of previously published values. If a sufficiently long time series is available, one may apply time series analysis to identify possible trend discontinuities and hence suspicious aggregates.
- Other quality information about the statistical process so far. For instance, a non-response analysis provides information on aggregates that have a high risk of being biased. If estimates of sampling errors are available, these may also be incorporated in the macro-editing procedure (see Section 2.2).

It should be noted that in macro-editing all actual adjustments to the data take place at the *micro* level, not the *macro* level. Therefore, after one has found suspicious aggregates by any of the above means, the next step is to identify individual units that contribute to these aggregates and may require further editing. The next two subsections describe two generic approaches to do this. The *aggregate method* (Section 2.2) proceeds by ‘drilling down’ from suspicious aggregate values to lower-level aggregates and, eventually, individual units. The *distribution method* (Section 2.3) examines the distribution of the microdata to identify outliers and other suspicious values. In practice, the two methods are often applied together.

2.2 The aggregate method

Given a data set that requires macro-editing, the aggregate method starts by calculating estimates of aggregates at the highest level of publication based on the current data (Granquist, 1994). These provisional publication figures are checked for plausibility and consistency, as discussed in Section 2.1. If an aggregate is identified as suspicious, the next step is to zoom in on the cause of the suspicious value by examining the lower-level aggregates that contribute to the suspicious aggregate. This procedure is sometimes called ‘drilling down’. In this way, macro-editing proceeds until the lowest level of aggregation is reached, i.e., the individual units. Finally, the units that have been identified as the most important contributors to a suspicious provisional publication figure are submitted to manual follow-up (see “Statistical Data Editing – Manual Editing”).

In practice, checking for suspicious aggregates is often implemented by means of score functions, similar to those that are used at the micro level in selective editing (see “Statistical Data Editing – Selective Editing”). In macro-editing, the score function is applied at the aggregate level (e.g., Farwell and Schubert, 2011). In practice, relatively simple score functions are often used, such as:

$$S_j = \frac{\hat{T}_{y_j} - \tilde{T}_{y_j}}{\tilde{T}_{y_j}}, \quad (1)$$

where \hat{T}_{y_j} is the estimated total of variable y_j based on the unedited data, and \tilde{T}_{y_j} is a corresponding anticipated (or predicted) total value. This score function measures the relative deviation from the anticipated value. Possible sources of anticipated values are: estimates from different data sources, such as a register or a different survey, or the value of the same total in a previous survey cycle – possibly corrected for development over time using a time series model (see also Section 2.1).

Comparisons based on ratios of aggregated values are also used, such as:

$$S_{jk} = \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right) / \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}}, \quad (2)$$

using notation similar to (1).

Since macro-editing is applied when all, or nearly all, data are available, there is no need to set a threshold value on the score function in advance. Instead, the aggregates can be put in order of suspicion by sorting on the absolute value of S_j or S_{jk} . In order to prevent the introduction of bias, it is important to treat large positive and large negative deviations from the anticipated values with equal care.

If the estimates are based on a sample of the population, as is often the case in business surveys, a natural amount of variation in the aggregates is expected due to sampling error. From a theoretical point of view, it is good to take this inaccuracy of the estimated aggregates into account in the score function. Thus, instead of (1), one could use

$$S'_j = \frac{\hat{T}_{y_j} - \tilde{T}_{y_j}}{se(\hat{T}_{y_j} - \tilde{T}_{y_j})},$$

and instead of (2), one could use

$$S'_{jk} = \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right) / se \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right),$$

where $se(.)$ indicates the standard error of an estimate. In these alternative score functions, deviations from the anticipated values are only seen as suspicious if they are large compared to the associated sampling error. This refinement is particularly important if there are large differences in accuracy between different aggregates.

For the final step in the aggregate method, the so-called ‘drilling down’ from suspicious aggregates to contributing individual units, the same score functions on the micro level can be used as in selective

editing (see “Statistical Data Editing – Selective Editing”). The main difference is that, again, there is no need to set a threshold value in advance here, because the score function can be computed for all records at the same time. This means that the records can be sorted on their score function value and treated in order of priority.

As an alternative to the aggregate method, one could also consider working directly with the sorted record-level score function values, by manually following up records in descending order of their absolute scores and continuing until all aggregates are deemed sufficiently plausible. This was called the *top-down method*¹ by Granquist (1994).

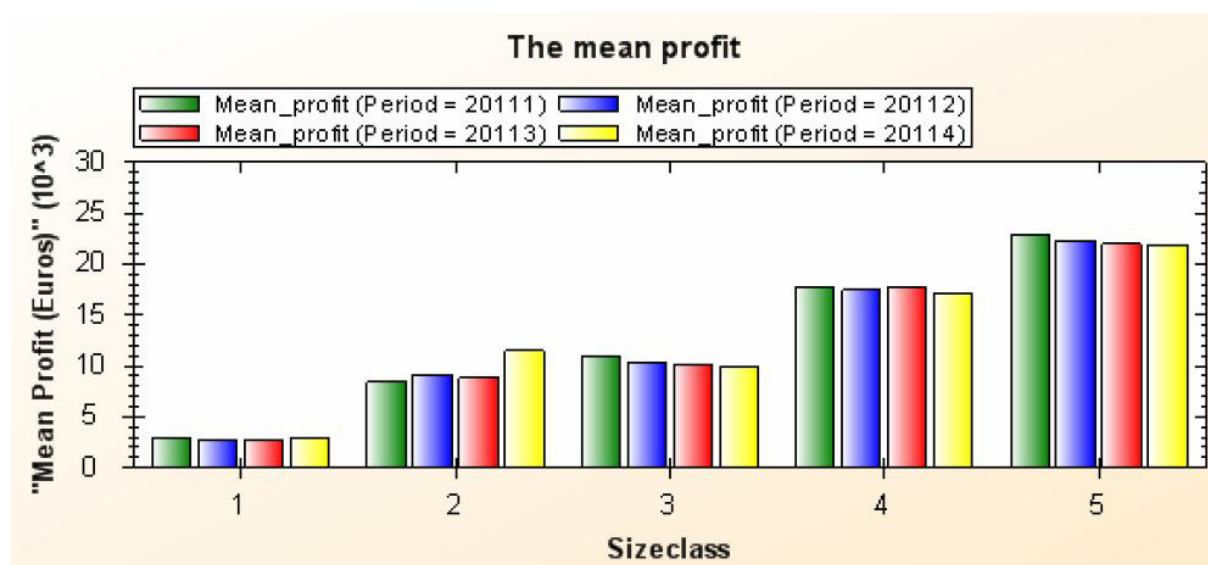


Figure 1. Example of a histogram for macro-editing (taken from Hacking and Ossen, 2012).

In addition to score functions, graphical aids can also be useful for identifying suspicious aggregates. As an example, Figure 1 shows a histogram that compares the mean value of profit across several reference periods and several size classes. It is seen that the mean profit in the last period for size class 2 is unusually high in comparison with previous periods and other size classes. This could be a reason to identify this aggregate as suspicious and drill down to the contributing units.

2.3 The distribution method

Another method for selecting individual units for manual editing, given all or most of the data, is known as the distribution method. This method tries to identify observations that require further treatment by applying techniques for detecting *outliers*, i.e., observations that deviate from the distribution of the bulk of the data. For the purpose of macro-editing, records are then prioritised for manual follow-up by ordering them on some measure of ‘outlyingness’. A discussion of outlier detection techniques in the context of statistical data editing can be found in EDIMBUS (2007).

¹ The name ‘top-down method’ is a potential source of confusion, because it is sometimes used as a synonym for the aggregate method (e.g., De Waal et al., 2011, p. 208). This probably derives from the fact that the aggregate method starts at ‘top level’ aggregates and ‘drills down’ to lower-level aggregates.

Theoretically speaking, there exists some overlap between this approach and the above approach based on score functions, because many common criteria for detecting outliers can be expressed as score functions; see, e.g., De Waal et al. (2011).

Graphical displays can also be useful for detecting observations that deviate from the distribution of the bulk of the data. Common examples include box plots, scatterplots, and other techniques from Exploratory Data Analysis (Tukey, 1977). Figure 2 gives an example of a scatterplot that could be used in this context. A graphical analysis can be particularly effective if the software allows an editor to interact with a display. In the plot of Figure 2, whenever a user moves his mouse to one of the points, information about the relevant unit is automatically displayed. This can be taken one step further by letting a user access a record for further editing by simply clicking on the point that represents the record in the graphical display. See, e.g., Bienias et al. (1997) and Weir et al. (1997) for examples of applications of graphical macro-editing. For some more recent innovations, see Tennekes et al. (2012).

In practice, the distribution method is often applied in conjunction with the aggregate method. Thus, the macro-analysis starts by identifying suspicious aggregates at the highest level and ‘drills down’ to suspicious aggregates at a lower level. Subsequently, the distribution method is applied to identify the records that are likely to contribute most to the total error in the identified low-level aggregates.

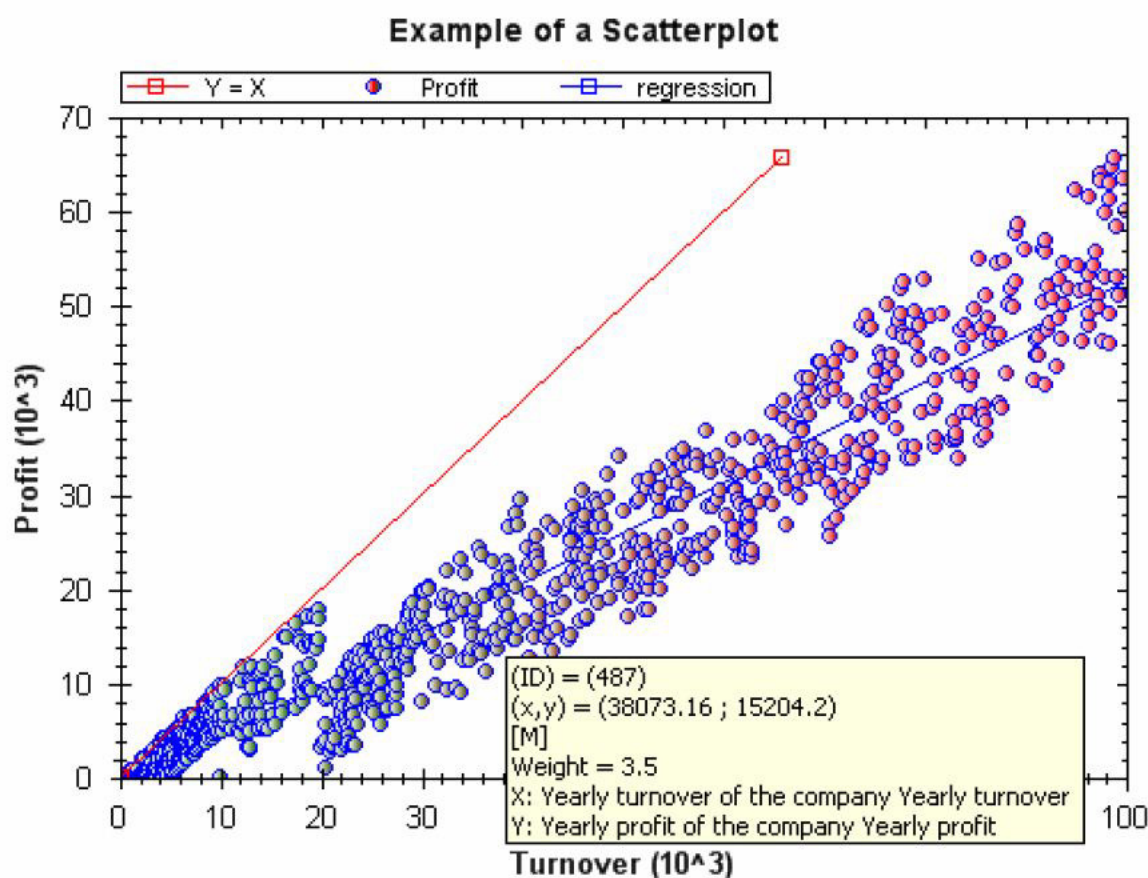


Figure 2. Example of a scatterplot for macro-editing (taken from Hacking and Ossen, 2012).

3. Design issues

4. Available software tools

Many statistical offices have developed macro-editing tools. Quite often, several such tools exist within one office, each one dedicated to a particular survey.

Statistics Netherlands has developed a generic macro-editing tool called *MacroView*; see Ossen et al. (2011) and Hacking and Ossen (2012). It is currently used for macro-editing in the production processes of the Dutch structural business statistics and the Dutch short-term statistics, as well as several smaller statistical processes. It is currently not made available to other statistical offices.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bienias, J. L., Lassman, D. M., Scheleur, S.A., and Hogan, H. (1997), Improving Outlier Detection in Two Establishment Surveys. In: *Statistical Data Editing, Volume 2: Methods and Techniques*, United Nations, Geneva, 76–83.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Farwell, K. and Schubert, P. (2011), A Macro Significance Editing Framework to Detect and Prioritise Anomalous Estimates. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Granquist, L. (1994), Macro-Editing – a Review of Some Methods for Rationalizing the Editing of Survey Data. In: *Statistical Data Editing, Volume 1: Methods and Techniques*, United Nations, Geneva, 111–126.
- Hacking, W. and Ossen, S. (2012), User Manual MacroView. Report PMH-20121125-WHCG, Statistics Netherlands, Heerlen.
- Ossen, S., Hacking, W., Meijers, R., and Kruiskamp, P. (2011), MacroView: a generic software package for developing macro-editing tools. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Tennekes, M., de Jonge, E., and Daas, P. (2012), Innovative Visual Tools for Data Editing. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.

Tukey, J. W. (1977), *Exploratory Data Analysis*. Addison-Wesley, London.

Weir, P., Emery, R., and Walker, J. (1997), The Graphical Editing Analysis Query System. In:
Statistical Data Editing, Volume 2: Methods and Techniques, United Nations, Geneva, 96–104.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Statistical Data Editing – Selective Editing
3. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

1. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

1. n/a

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

12. Tools explicitly referred to in this module

1. MacroView

13. Process steps explicitly referred to in this module

1. Statistical Data Editing

Administrative section

14. Module code

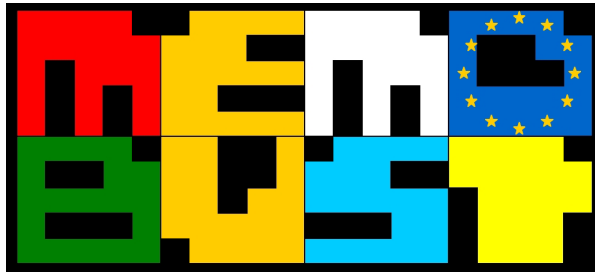
Statistical Data Editing-T-Macro-Editing

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	04-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.2	18-04-2013	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	19-07-2013	minor improvement based on second Swedish review	Sander Scholtus	CBS (Netherlands)
0.3.1	09-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:12



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing Administrative Data

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Statistical data editing of administrative data.....	4
2.2 Types of errors in administrative data	6
2.3 Data editing methods for administrative data.....	8
2.4 Information about data quality	11
3. Design issues	13
4. Available software tools.....	14
5. Decision tree of methods	14
6. Glossary.....	14
7. References	15
Interconnections with other modules.....	16
Administrative section.....	17

General section

1. Summary

The use of administrative data as a source for producing statistical information is becoming more and more important in Official Statistics. Several methodological aspects are still to be investigated. This module focuses on the editing and imputation phase of a statistical production process based on administrative data. The paper analyses how much the differences between survey and administrative data affect concepts and methods of traditional editing and imputation (E&I), a phase of the production of statistics that nowadays has reached a high level of maturity in the context of survey data. This analysis enables the researcher to better understand how and to which extent traditional E&I procedures can be used, and how to design the E&I phase when statistics are mainly based on administrative data.

2. General description

The use of external information in statistical production processes is increasing its importance in the National Statistical Institutes (NSIs).

External information generally refers to *secondary data*, i.e., data not collected directly by the user. An interesting discussion on the use of this kind of data can be found in Nordbotten (2012). In this paper, the focus is on administrative data, which is a subset of secondary data. They have the characteristic of being collected for non-statistical purposes and at the moment they are the mostly used external source of information in NSIs.

Administrative data are collected for administrative purposes, e.g., to administer, regulate or tax activities of businesses or individuals. Although not yet fully explored from a methodological point of view, the field of the statistical use of administrative data can be considered in an advanced state for a number of critical issues like accessibility, confidentiality and risk of misuse.

The usefulness of administrative data depends on their concepts, definitions and coverage (and the extent to which these factors stay constant), the quality with which the data are reported and processed, and the timeliness of their availability. These factors can vary widely depending on the administrative source and the type of information (Statistics Canada, 2010).

It is worthwhile to remark that, although this definition could be applied to survey data, in the context of administrative data it assumes a particular importance since most of the elements considered in the statement are not under the control of the NSIs, while on the contrary for survey data NSIs can, at least in principle, design opportunely all or most of them.

The main advantages deriving from the statistical use of administrative data include: the reduction of costs (in the long term) and of respondent burden, deriving from the reduction of information needs from direct surveys; the improvement of timeliness and accuracy of statistical outputs; the increased potentials for more detailed spatial-demographic and longitudinal analysis.

Main drawbacks are connected to the initial costs due to gain access to the new sources, matching classifications, harmonising concepts and definitions with respect to the target units and the statistics of interests, and assessing quality. Concerning the latter aspect, it is worthwhile noting that the quality of data collection, data capture, coding and data validation are under the control of the administrative

program and may focus on aspects that could be not relevant for the NSI's purposes. In general, these validation activities cannot be considered sufficient to ensure the statistical usability of the data, and extensive additional data editing activities need to be performed before incorporating external data into statistical processes. Methods and tools are to be developed to this aim taking into account the peculiarities of administrative data. In addition, the use of an administrative source generally implies the need of other sources (including surveys) to compensate for non-covered units/variables, thus editing strategies for multi-source data should be developed.

The impact of using administrative data in statistical production processes depends also on their supposed use. Two different scenarios can be distinguished:

- 1) administrative data support surveys: they are used to maintain frames, to improve the efficiency of sample surveys (calibration), to provide information which might be used to assist the E&I process, as an information source that might be used for quality assurance (for instance to compare results);
- 2) administrative data serve as a source for providing the statistical output required, in this case they can be used as a primary source or by integrating them with survey data.

In this paper the focus is on the use of administrative data under scenario 2.

The paper is structured as follows. In Section 2.1 the main objectives of data editing for survey data are discussed in the framework of administrative data. Section 2.2 is dedicated to the illustration of error characteristics in administrative data. The application of traditional methods used E&I is discussed in Section 2.3. How to provide information about data quality is illustrated in Section 2.4. General ideas about the design of E&I of administrative data are proposed in Section 3.

2.1 Statistical data editing of administrative data

The main objectives of statistical data editing are reported in the following list (cf. "Statistical Data Editing – Main Module"):

- OB1 To identify possible sources of errors so that the statistical process can be improved in the future;
- OB2 To provide information about the quality of the data collected and published;
- OB3 To detect and correct influential errors in the collected data;
- OB4 To provide complete and consistent data.

When discussing E&I for administrative data, the main question is how much the concepts developed so far for E&I of a single statistical survey (see EDIMBUS, 2007) can be translated into the administrative data framework. The question is translated in two main questions: 1) whether the above mentioned objectives are still valid, and 2) whether error characteristics and methods usually adopted for detection and treatment are the same. To give an answer to those questions, differences between administrative and survey data should be highlighted.

Two important distinctive characteristics are:

- i. the process of gathering information is not generally under the control of the entity (for instance the NSI) that will provide the final figures,
- ii. information is gathered for other purposes.

Other important differences are that:

- iii. generally the sizes of the data bases concerning administrative data are much larger than those concerning survey data,
- iv. administrative data are frequently used in a statistical production process where data sources are combined and integrated. The integration of data sources becomes a specific trait of the use of administrative data since, as they are gathered for other purposes, they generally do not observe all the variables of interest, and most of the times they refer to a population covering a part of the target population. In those cases, integration between administrative sources and surveys is required to fill the gaps.

Those peculiarities influence the objectives of statistical data editing procedures, a short discussion about interactions between main objectives of E&I and peculiarities of administrative data follows.

Objective OB1

The identification of source of errors becomes in this context particularly important. In fact, one of the main problems is that the definition of collected variables is not designed for the survey purposes, and even after a process of harmonisation, some differences may still remain. The process of editing can help to reveal unexpected differences and to find whether there is a systematic nature of the error suggesting that the definitions are still not completely harmonised. Unfortunately, the improvement of the statistical process is limited by the fact that the process is not completely under the control of the NSI. Most of the times it is not easy or even impossible to return to the administrative entity collecting data and to make the agency change the definition of the variables, the data collection and so on.

Objective OB2

As for E&I of survey data, the data quality assessment in terms of input and output data is a key aspect also for statistics based on administrative data. The fact that two separate entities influence the data and the data production process, i.e., data holder and statistics provider (NSI), implies that two different points of view can be used for quality evaluation: a data perspective and a perspective oriented to the production of statistics. The first one is useful to provide information to the data holder to improve data quality for other data collection occasions, while the second one is important to measure the quality of the statistics provided inside and outside the NSI.

Objective OB3

The generally large dimension of databases has an impact on the detection and correction of influential data (which especially characterise quantitative variables), since for their treatment an expensive data editing procedure based mainly on re-contacting units is generally adopted. On the other hand, the use of multiple data sources may lead to have multiple observed values for a single observation, this information can be used to improve the selective editing procedure in terms of both identification of influential errors and value correction when an influential observation is selected. The same considerations hold when longitudinal information is available on units covered by administrative sources. These aspects will be later discussed in the subsection on editing methods.

Objective OB4

In case of integration of several data sources, the data consistency becomes an essential aspect, because the integration will increase the possible conflicts into the available information. However, as

previously stated, the presence of multiple observations is an important aspect that can improve the E&I procedures, although at this time not many methods are developed to exploit as much as possible this richness of information. This issue will be discussed in the subsection on editing methods.

In the end, we can state that the general setting designed by the objectives of E&I of survey data remains still valid for administrative data. On the other hand, it is important to be aware of the impact of peculiarities of administrative data giving a different perspective to the objectives, those peculiarities will have an impact in the design and use of methods for E&I of administrative data

2.2 *Types of errors in administrative data*

As previously discussed, also in case of administrative data, one of the most important objectives of statistical data editing is to deal with errors, for this reason is important to discuss the characteristics of errors affecting administrative data. Before starting with the description of errors is useful to clarify a question: are administrative data affected by errors? It is difficult to imagine that data relating, for instance, to tax declaration can be affected by errors. It is nowadays accepted the idea that administrative data can be affected by errors (Groen, 2012), in fact also for this type of source errors may arise in many phases of the data production process, e.g., at the data transmission phase between data holder and NSI. Furthermore, there are also less controlled administrative data sources where the information is not so immediately sensible to make the data holder perform a check. A discussion about errors can be found later in this section.

Summarising, as well as survey data, administrative data are normally affected by different types of error: in the most recent literature, it is actually accepted that the non-sampling errors that normally emerge in surveys may also occur in registers (Bakker, 2011; Zhang, 2012). We start from the assumption that all the errors dealt with at the E&I phase in case of a single source survey are potentially present in a single administrative data source, hence the discussion is focused on the new additional aspects characterising errors in administrative data, with special attention to the case of statistics produced by integrating different data sources.

The E&I procedures are mainly designed to deal with measurement errors and missing values, the latter concerning usually item non-response. These sets of errors are analysed in the following.

Measurement errors are defined as differences between the recorded values of variables and the corresponding real values (*intended measure* of the variable). They mainly arise because of the fact that administrative sources are the result of processes which, being designed for purposes other than statistical, may use different concepts and/or definitions than those required for the specific statistical purposes. Important differences between the sources of measurement errors in survey data and in administrative data derive from the fact that the measurement process is very different in the two situations. In surveys using questionnaires, measurement errors derive from a cognitive process (comprehension of the question, retrieval of the information, judgment and estimation, reporting the answer) which also acts in case of administrative data but is not the most important one. A most important role in this case is played by administrative and legislation rules and accounting principles (Wallgren and Wallgren, 2007, p. 180). Typical measurement errors in administrative data are errors in accounting routines, or misunderstanding due to legally complicated questions, or errors deriving from the misspecification of rules used for deriving statistical variables from administrative variables. Furthermore, as some variables recorded for administrative purposes are more important than others,

their accuracy is expected to be superior, as it can be assumed that enterprises answer to less important questions with lower precision. It is worth mentioning that the cognitive process also acts in case of administrative data: measurement errors may derive from the fact that respondents may provide different data to the different government agencies depending on their specific purpose, they may understand administrative concepts and definitions incorrectly (thus introducing errors by deviating from definitions, e.g., including wrong elements in the reported variables), or they can make unintentional errors in providing information.

Among measurement errors, also in case of administrative data variable values may contain ***systematic errors*** (cf. “Statistical Data Editing – Main Module”), which in this case can be due, for example, to a misinterpretation of record descriptions, originated by changes in the record descriptions and/or variable names in the administrative data bases.

An important source of errors for statistics based on multi source administrative data is the process of data integration itself. When the statistical population is created, objects are adjoined and linked, variables are imported from different sources and derived variables are created. The most relevant types of errors associated to the integration process are *coverage errors*, *identification errors*, *consistency errors*, *aggregation errors*, *missing values* (Zhang, 2012; Wallgren and Wallgren, 2007, p. 177). While coverage errors are not usually treated through E&I, the others are dealt with by or have an impact on the E&I process, for this reason they are described in the following.

Identification errors. They may be originated by errors in identifying variables used to match the different sources. As a consequence, identification errors may give rise to doublets, mismatches (e.g., false hits), item and total non-response, data inconsistencies (as variables may be referred to not properly matched objects). Identification errors may also generate outliers, and influential errors.

Consistency errors. They may also originate from the integration of variables from many sources. This type of error is especially increased when using multi source data, on the contrary with a single statistical survey, the use of a unique questionnaire ensures a better consistency in the data. Consistency errors can be caused by errors in units and errors in variables. They may also have a longitudinal origin, e.g., due to identifying variables either in error or changing over time for a same unit, splits/fusions of a unit over time.

Incoherent variable values giving rise to consistency errors in microdata may occur in the situation where the integrated administrative sources are overlapping regarding (a subset of) variables.

Inconsistencies with information from other sources and outliers can be originated from modifications of the variables’ definitions adopted in a source (e.g., resulting from legislative changes), and from the fact that units may change their structural characteristics (e.g., fusions or splits). Outliers can also be determined by taxation measures that produce anomalous changes in variables values over time, and by integration errors (e.g., different units are linked in administrative sources). Outliers can either correspond or not to influential errors, depending on their impact on the target estimates.

Aggregation errors. They may occur when data from different administrative sources with different types of units are integrated in order to derive statistical variables (Wallgren and Wallgren, 2007), e.g., enterprise labour cost deriving from fiscal archives on enterprise employees. Aggregation errors may originate internal inconsistencies among variables referring to the same unit, outliers and longitudinal inconsistencies.

Missing values. As for statistical surveys, also in case of administrative data, missing values may correspond to two types of non-response : *unit non-response* (all the information for a statistical unit is unavailable) and *item non-response* (incompleteness of information, for some units, on topics which are of interest for statistical purposes). In case of administrative data, unit non-response corresponds to under-coverage, for example, when the integrated administrative sources relate to sub-populations which do not cover the overall target population. Item non-responses typically derive from the fact that the content of administrative sources is defined on the basis of administrative requirements, thus not all topics of interest may be covered by the administrative data. Possible sources of item non-response can arise for other different reasons: variable values can be missing for certain objects due to flaws of a source; mismatches at the integration phase due to missing objects in a source, giving rise to missing values for all the variables which are imported from that source; reported values which are “cancelled” as recognised invalid at the editing stage; values which fail to be reported, or are reported with a delay. Item non-response can also be associated to the fact that the content of a source is subject to modifications, resulting from legislative changes, like the drop-out of some information from the administrative forms; in a longitudinal perspective, non-responses can also appear as missing information on target variables for units considered over time: this can be due again to modifications of the units (fusions/splits, other structural changes) or to changes in legislation. Finally, as administrative sources may refer to either a point in time (i.e., they describe the units set at that point in time), or to a calendar year (in this case they contain all units that have existed at any point during the year), item non-responses may rise when sources with different time characteristics are integrated.

2.3 *Data editing methods for administrative data*

In this section we focus the attention on methods which can be used to detect and treat measurement errors and item non-response, that are in fact the errors dealt with by an E&I procedure..

Several classifications for the data editing techniques are available; we follow the one proposed in “Statistical Data Editing – Main Module”. The techniques can be classified as:

1. Deductive editing.
2. Selective editing.
3. Automatic editing.
4. Interactive editing.
5. Macro-editing.

The order follows the strategy that is generally adopted in an E&I process for a statistical survey (cf. “Statistical Data Editing – Main Module”).

In this section we discuss the impact of the peculiarities of administrative data on the features of each data editing technique.

Deductive editing is the phase where methods for detecting and treating errors with a structural cause that occurs frequently in responding units (systematic errors) are used (see “Statistical Data Editing – Deductive Editing”). In administrative data, especially when more sources are used, deductive editing has an important role in the production process. Variables collected in the administrative sources may have similar definitions but they may have structural gaps given to the convenience of declaring some

information in an item rather than in another one, for instance, declaring something either in a cost or in an investment item. The first step in an E&I process should be to look for systematic errors in the observed values, also in the case the definition of variables is almost the same with respect to the corresponding statistical target variable. Hence, deductive editing is substantially the same as the one carried out in a classical data editing process, in fact the detection of systematic errors implies the involvement of subject matter experts, and the error treatment, that is usually completely automated, is not affected by the large dimension of administrative databases.

The aim of *selective editing* is indeed the optimisation of the process of selection of units to be deeply revised (in most cases, re-contacted) by restricting the editing only to those affected by an important error, and this naturally stresses the importance of selective editing in this context where data sets have usually a large dimension. On the other hand the use of selective editing is actually limited by resources' constraints because even a small percentage of units to be analysed may be too large in a large data set. A further constraint for selective editing on administrative data derives from the difficulty of re-contacting units for this kind of data. This limitation is alleviated when multi-source data are used, in this case the availability of different values for the same observation is an important aspect that can help the statistician in understanding where the error is located and to recover a likely value. The previous considerations mainly illustrate the problems in applying selective editing to administrative data. However, some further remarks concerning positive aspects of selective editing with administrative data are worthwhile to be mentioned. In selective editing, observations are prioritised according to a score function measuring the impact on the target estimates of the expected error in the unit. The error is frequently measured by comparing the observed value with a suitable prediction. In the context of administrative data, there is frequently the possibility of using longitudinal data, and this can improve the efficiency of selective editing as better predictions can be obtained. Finally, it is worthwhile to note another specific difference characterising the application of selective editing in administrative data with respect to the survey data. In a survey, the error is generally weighted with sampling weights. Since the prioritisation of an observation should be based on the impact of the error on the estimates, the final sampling weights should be taken into account in this process. In practice, this can be rarely performed, as final weights are generally computed once the editing step is completed, so an approximation is generally used by considering initial sampling weights. In the case of administrative data this problem is naturally overcome because sampling weights are not an issue for these kinds of data and a more precise estimation of the impact of errors on estimates can be obtained.

Automatic editing refers to all E&I procedures that detect and treat errors and missing values in a data file in a fully automated manner, i.e., without human intervention (see "Statistical Data Editing – Automatic Editing"). In the last years, most of the methods for automatic editing are based on the Fellegi-Holt paradigm, which means that the smallest number of fields should be changed to a unit to be imputed consistently. The algorithms are based on edits that represent rules/constraints characterising the relationships among variables.

In principle, if the focus is just on one data source, we are in the same situation as the one we would have in an E&I process of statistical survey data. However, as already remarked, most of the times different data sources are integrated, and in this case some additional problems may arise. A first issue to take into account is whether the data sources should be treated simultaneously as a unique data set after the integration process. This could be an interesting option, because the amount of information

would increase, and an improvement in the E&I procedure is expected. In this case, edits simultaneously involving variables of the different data sources should be considered. A special but not infrequent case is when the same (at least in principle) variable is observed in the different data sources. For the sake of simplicity, let us suppose that there are only two data sets with the same variable. According to the Fellegi-Holt approach, we are assuming that with a high probability at least one of the two variables in turn is not affected by error. In the case that this assumption is not reliable, a different approach should be followed, for instance, a prediction conditionally on the observed values of the two variables can be obtained. Techniques developed to this aim are described in the module “Micro-Fusion – Reconciling Conflicting Microdata”.

Concerning *interactive editing* for administrative data, the most relevant aspect is that, as already remarked, it is frequently not possible to re-contact the observed units, so one of the main advantages motivating interactive editing declines. However, interactive editing can be considered effective in order to understand error sources and possibly resolve errors in the short term, while in the long term it can contribute to the increase of the subject-matter expertise for the staff working on administrative data, increasing their knowledge of the characteristics and the contents of administrative data and gaining understanding of how the data can be used in a more suitable way (Wallgren and Wallgren, 2007).

Macro-editing aims at looking for anomalous aggregates. The anomalies are identified based on the comparison of aggregates with some reference values that, for instance, may be obtained by previous published figures. Once anomalous aggregates are selected, a drill-down procedure is applied in order to find the units that mostly contribute to this behaviour (see “Statistical Data Editing – Macro-Editing”). This editing approach requires the computation of the final aggregates (e.g., domain estimates), and for this reason, in the usual E&I procedure it is generally performed at the end of the E&I process. In this context, one generally works on complete data sets, in fact administrative data are gathered for other purposes and they are usually provided to the NSIs at the end of their collection. This implies that in this context macro-editing methods can be used at the beginning of an E&I procedure in order to look for important errors.

Macro-editing can be a useful tool to reveal whether some important errors due to an incomparability of the sources in some estimation domain are still present in data. For instance, it can happen that the definition of a variable is the same in two data sources. Nevertheless, for a specific economic sector some particular businesses could not provide the complete amount of the value in one source because of fiscal benefits typically allowed only for that segment of units. Macro-editing can be useful to isolate those critical situations that the subject matter expert may study and interpret in order to fix the problem wherever it is possible. Macro-editing can also reveal errors due to data linking or to the incomplete delivery of some sources, as anomalous aggregates may result from not enough covered domains from one time period to the subsequent one.

As already mentioned, administrative data are subject to partial non-response as well. **Imputation** (see the topic “Imputation”) can be used to manage missing values in order to obtain a completed data set on which the usual statistical analysis can be applied. The methods usually adopted are based on the missing at random (MAR) assumption that is, roughly speaking, the probability of non-response on a given variable depends on the observed values and not on the unobserved ones of the variable itself. For instance, missing values in administrative data can be due to lack of timeliness, and it is generally

supposed that businesses answering in due time have the same behaviour as the not observed ones. Actually this situation could hide the presence of a problem in the business, and in this case the estimates could be biased because the observed and non-observed populations are actually different. A similar concept applies in the case of an integrated use of administrative data. It can happen that each administrative source covers only some specific part of the target population. Imputation can be used to complete the missing values, again under the assumption that the population not covered has the same behaviour of the observed one.

Finally, since the production process of administrative data is generally beyond the control of NSIs, a continuous assessment of the data quality should be planned. Edit rules and macro-editing based approaches could be used to this aim. An anomalous rate of edit failure and/or anomalous variation of statistical aggregates in two consecutive times could alert data producer that some important changes could have been introduced in the administrative data production process, which could be related to a change in the data collection, to a change in the legislation that impacts on the definition of measured variables, consequences of a different fiscal policy, and so on.

2.4 Information about data quality

One of the main goal of E&I is to provide information about the quality of the data collected and published.

Quality of statistical output has several dimensions, they are thoroughly discussed in Eurostat (2011) for the European Statistics Code of practice, Eurostat (2009) for a handbook (soon to be revised) on reporting quality of statistical data according to the European output quality components, and the handbook module “Quality Aspects – Quality of Statistics”.

In this section it is important to refer to the quality dimensions in the context of administrative data in order to describe on which of them the E&I is a useful tool for providing information. In the BLUE-ETS (2011) document, the quality dimensions of administrative sources and the related indicators are discussed. In that document the focus is on the quality dimensions of the administrative data sources in the input phase of a statistical production process, this point of view is adopted in this paper as well. As far as the quality dimension of the statistical output based on administrative data is concerned, we assume that at the end of the E&I process data are statistically transformed, and hence the general considerations made for statistical output based on survey data are still valid. This is a simplistic position, that is also motivated by the fact that at this time this issue is still under discussion, and further studies are needed in this context. For the use of E&I procedures as a useful tool for providing information on quality of statistical data, the reader may refer to EDIMBUS (2007).

A first interesting remark relates to the point of view chosen to look at the quality aspects. It reflects the peculiarity of statistics based on administrative data where generally two different main actors are involved: the data holder and the statistics provider (NSI). Two main points of view are introduced: a data archive perspective and a perspective oriented to the production of statistics. In the first one, the quality is independent of the specific statistical use of the administrative data that is supposed to be done, while in the second one the quality is related to the statistical use of the data planned at the NSI. Both these aspects are important for E&I, in fact the first one has to be assessed in order to foster data holder to improve the quality of the data, while the second one is related to the quality of published data.

In the BLUE-ETS document, the following quality dimensions are defined:

1. *Technical checks*, that is the technical usability of the file and data in the file.
2. *Accuracy*, that is the extent to which data are correct, reliable, and certified.
3. *Completeness*, that is the degree to which a data source includes data describing the corresponding set of real-world objects and variables.
4. *Time-related dimension*, in which timeliness, punctuality, and overall time lag applied to the delivery of the input data are taken into account.
5. *Integrability*, that is the extent to which the data source is capable of undergoing integration or of being integrated.

The *technical check* dimension is mainly related to IT aspects, e.g., data accessibility, correct conversion of the data, data complies with the metadata-definition. These aspects are not related to an E&I procedure as it is defined in “Statistical Data Editing – Main Module”.

E&I has certainly impact on *accuracy*, and it naturally provides information about some dimension indicators described in BLUE-ETS (2011) related to this aspect. Some of the dimension indicators for accuracy proposed in BLUE-ETS are supposed to measure:

- *Measurement error*: deviation of actual data value from ideal error-free measurement;
- *Inconsistent values*: extent of inconsistent combinations of variable values;
- *Dubious values*: presence of (or combinations of) implausible values for variables.

Those elements are treated and analysed during an E&I procedure, and indicators measuring them are developed and generally automatically provided by the usual procedures (see EDIMBUS, 2007).

E&I may be useful to gather information also for other quality dimensions, that apparently are less naturally related.

Completeness is a concept referred to units and variables, and for the latter the quality dimension indicators proposed in BLUE-ETS (2010) are: the amount of missing values and the amount of imputed values. As previously stated, the treatment of missing data (imputation) is one of the main activities carried out in an E&I process; hence, indicators on those aspects are easily obtained in this context.

As far as the *time related dimension* is concerned, a proposed indicator focuses on the stability of variables. To this aim, the comparison in different times of indicators generally provided by E&I may be useful: for instance, an anomalous variation of the failure rates of some edits may hide some changes in the administrative data production process or in the source contents, or in the use of a different definition for a variable, or in a different data collection mode. Also the comparison of the amount of imputed values and missing data can reveal some changes in the data source which have to be taken into account in order to avoid biasing effects on statistical results.

A summary of the editing undertaken and the results of the checks should be sent to the database owner to make him aware of the problems possibly existing in the data set, in order to reduce them as much as possible in the future and improve the overall quality of the data. As a consequence, managing and improving co-operation with administrative bodies plays a central role in this context:

NSIs need to increase co-operation and to determine appropriate incentives in order to improve the overall communication and interaction with data owners, to get them to set up better editing practices and conform to statistical classifications and definitions, and to provide feedback to the NSI in the data verification process (Shlomo and Luzi, 2004).

3. Design issues

In the design of an E&I process for administrative data the first important issue to take into account is whether the target statistics are based only on a single administrative source or on the use of multiple integrated administrative sources. Moreover, editing strategies must take into account the trade-off between the potential gain in accuracy deriving from the availability of detailed and extensive information, and the additional costs needed for validating it.

When only one source is used, as discussed in the previous sections, we are in a similar situation to that of E&I of a single survey, even if we remind that peculiarities of administrative data should be taken into account because of their impact in the E&I methods. The reference flow-chart introduced in the module “Statistical Data Editing – Main Module” can be applied to this case.

When more sources are integrated, different scenarios can be depicted.

A first scenario may consist of the following macro-phases:

1. check separately each single administrative source;
2. integrate the edited data sources;
3. edit the integrated sources in order to assess the consistency among variables’ values obtained from the different sources.

This is actually the flow-chart reported in Wallgren and Wallgren (2007, p. 101).

The drawback of this way of proceeding is that it is resource demanding since many different E&I procedures must be set and applied, and it is well known that the E&I is one of the most expensive parts of the statistical production process. Moreover, not all the amount of information is used at the same time, for instance, for the imputation of a variable in a data source it could be useful to exploit variables observed in the other data sources. Let us imagine the case when two data sources are integrated and in one source the income is observed, while in the other one information on consumption is gathered. The imputation of the two variables separately would disregard the strong relationship existing between them. An advantage of this way of proceeding is that certain typologies of errors (e.g., systematic errors like unity measure errors, balance errors, errors due to incomplete delivery of data for some administrative objects) can be removed from each single source before the integration phase, thus reducing the amount of consistency errors on the linked data deriving from these situations; longitudinal information could be used at this stage.

An alternative scenario corresponding to the opposite solution is:

1. integrate the sources;
2. apply an E&I procedure to the integrated data set.

In this case, less resources would be demanded since only one data verification process is required, but the complexity of such a process would increase. Furthermore, as the integrated data set is not

generally composed of all the variables observed in the different administrative sources, in this case some relations linking variables in each data source could be disregarded.

A third scenario is a compromise of the previous ones:

1. apply a 'light' E&I procedure to each single administrative source;
2. integrate the edited data sources;
3. edit the integrated data sources.

The question is when an E&I procedure can be defined as light. The idea is that the time and effort spent in editing sources should be minimised while maintaining an acceptable level of quality of the data sources. This general idea resembles what is done in selective editing, where the effort is focused on the most important errors having a high impact on the target aggregates. This situation is slightly different because there is no requirement on a sufficient level of quality of aggregates for each single data source, but the level of quality is required at micro level: in effect, the use of each single source will be in a micro perspective given that the integration process is generally performed at this level. A proposal could be that of applying only corrections of systematic errors in the first editing step.

It is clear that a general flow-chart is not available; however, at least three scenarios have been designed. The main point is to see the E&I process as a unique process possibly composed of two steps. The choice of the most appropriate strategy should be based on the trade-off between the expected quality of the final aggregates and the resources which are actually available to obtain the required level of quality. Concerning the latter, an element which can be considered as relevant to increase the effectiveness of editing and correction activities is the availability of subject-matter experts, who are familiar with the administrative systems that have generated the data and their specific contents, and who are in good relations with the data providers.

Finally, independently on the chosen scenario, indicators providing information about input and output data quality should be part of the E&I process. Moreover, since the process of gathering information is out of control of NSIs, it is important to establish a system of indicators alerting about some possible changes in the data production process of the data holder, in order to avoid important and non-measurable errors in the published statistics.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

- Bakker, B. F. M. (2011), Micro-Integration: State of the art. In: *Report WP1: State-of-the-art on Statistical Methodologies for Data Integration*, ESSNET on Data Integration, available at <http://www.cros-portal.eu/content/wp1-state-art>.
- BLUE-ETS Project (2011), *Deliverable 4.2: Report on methods preferred for the quality indicators of administrative data sources*. Available at <http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf>.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Eurostat (2009), *ESS Handbook for Quality Reports*. Eurostat Methodologies and Working papers.
- Eurostat (2011), *European Statistics Code of Practice*. For the national and community statistical authorities. Adopted by the European Statistical System Committee 28th September 2011 (revised version).
- Groen, J. A. (2012), Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics* **28**, 173–198.
- Nordbotten, S. (2010), The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries. In: Carlson, Nyquist, and Villani (eds.), *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, 205–225. Available at officialstatistics.wordpress.com.
- Shlomo, N. and Luzzi, O. (2004), Editing by Respondents and Data Suppliers. In: *Federal Committee on Statistical Methodology, Statistical Policy Working Paper 38: Summary Report on the FCSM-GSS Workshop on Web-based Data Collection, April 2004*, 75–90.
- Statistics Canada (2010), *Survey Methods and Practices*. Catalogue no. 12-587-X. <http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf>.
- Wallgren, A. and Wallgren, B. (2007), *Register-based statistics – Administrative data for statistical purposes*. John Wiley and Sons, Chichester.
- Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* **66**, 41–63.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Main Module
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Imputation – Main Module
6. Weighting and Estimation – Estimation with Administrative Data
7. Quality Aspects – Quality of Statistics

9. Methods explicitly referred to in this module

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Statistical Data Editing – Deductive Editing
3. Statistical Data Editing – Automatic Editing
4. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

14. Module code

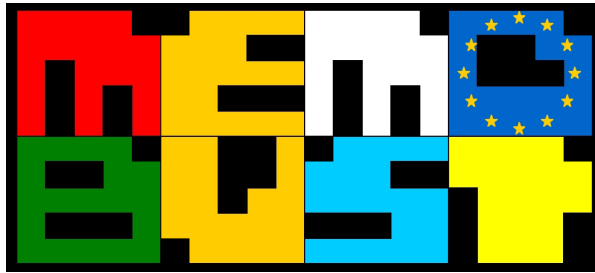
Statistical Data Editing-T-Administrative Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	13-03-2013	first version	M. Di Zio, O. Luzi	Istat
0.2	17-06-2013	introduction of a new section concerning quality indicators	M. Di Zio, O. Luzi	Istat
0.3	07-08-2013	minor revisions	M. Di Zio, O. Luzi	Istat
0.3.1	04-10-2013	preliminary release		
0.4	20-12-2013	revision based on EB comments	M. Di Zio, O. Luzi	Istat
0.4.1	09-01-2014	revision based on EB comments	M. Di Zio, O. Luzi	Istat
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:13



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing for Longitudinal Data

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Longitudinal data.....	3
2.2 Introduction to editing for longitudinal data.....	4
2.3 Editing scheme in a longitudinal context	4
2.4 Type of edits	5
2.5 Methods for longitudinal data	6
2.6 The case of categorical data	8
3. Design issues	9
4. Available software tools.....	9
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	11
Administrative section.....	12

General section

1. Summary

We refer to longitudinal data as repeated observations of the same variables on the same units over multiple time periods. They can be collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each unit from historical records. The process of Editing and Imputation can exploit the longitudinal characteristic of the data as auxiliary information, useful at both the editing and the imputation stages. This theme describes the editing process applied to longitudinal data, that could be performed for all aforementioned types of data, with special focus on Short Term Statistics context.

2. General description

2.1 Longitudinal data

Another term for longitudinal data is panel data. This definition focuses on the particular sample, which units are selected to be observed several times with some degree of regularity. The occurrence of those observations can be once along several years (every four years or biannual) or once a year (annually) or several times during the same year (quarterly or even monthly). Panel data are mostly used to describe patterns of change within and between the statistical units under observation, in other cases to highlight and to identify differences and changes over time of a specific parameter of the population under study. In general, for each unit $i = 1, \dots, n$ there are $t = 1, \dots, T$ different measurements, one for each wave of interview. The period t can be a month, a quarter or a year; the first two cases drive to infra-annual longitudinal data. As a consequence, given the period t , a vector of cross-sectional observations is available, while as regards the i -th observation a vector of longitudinal data is available and a strong correlation is expected among its values. According to the type of required estimates, different types of panel are considered, so it can always follow the same units or rotate some of them after a period (rotating panel). The different design will create different type of longitudinal data set.

In the context of business statistics, longitudinal data can be used both in structural and in short-term analysis. The difference between Structural Business Statistics (SBS) and Short Term Statistics (STS) actually depends on the combination of the survey occurrence and the type of final target parameter; see also the modules “General Observations – Different Types of Surveys” and “Repeated Surveys – Repeated Surveys”. In the SBS context, totals, means, levels are usually the object of the estimates; in the STS the main objective is usually to publish regular series of statistics on changes of totals for specific domains. These are frequently published in the form of index numbers, whose main purpose is to measure net changes between two periods. In these cases the rationale for a panel design is to improve the precision of estimates, because the minor variance of estimates is assured by the presence of historical correlation between data referred to the same units over the period in which the observations take place; see also the topics “Sample Selection” and “Weighting and Estimation”. On the other hand, also from an operational point of view, the use of a panel for an infra-annual survey can yield important cost savings. Indeed, to interview the same units is often less expensive than starting afresh, at each wave, the contacts on new units.

2.2 *Introduction to editing for longitudinal data*

In general, two main aspects are crucial in an editing process framework:

- 1) the rule to identify an acceptance region for a test variable;
- 2) the technique used to change a value detected as wrong during the process.

In a longitudinal context, these aspects have to be fitted to the specific target parameter, which is often given by the estimation of the change of a population parameter (mostly the mean) concerning a quantitative not-negative variable y . It is strongly recommended to use the available historical information of the observation units for two main reasons:

- 1) a strong correlation is expected among different measurements of the same variable on the same units, thus any detecting rule can rely on relevant information about the unit profile and can result in being more efficient;
- 2) since most of the time the target parameter is the change of a main parameter along time, any observed change between sequential periods on the observations can be used as a precious source of information with regards the final estimation.

In general, the editing process in a longitudinal context must take into account the characteristics of the change under investigation and the timeliness constraints. The control rules can be defined taking into account comparisons between values of the same variable on the same unit at different times, i.e., the two values y_t and y_{t-k} , where t is a month or a quarter, $t-k$ is a previous period and k varies according to the variable features and/or to the type of change under observation. Additional specifications are generally required, they are briefly described in the following.

2.3 *Editing scheme in a longitudinal context*

When the editing process is set on longitudinal data, there are some issues which assume a strategic meaning:

- 1) Longitudinal and cross-sectional checks can be carried out at the same time; this is because longitudinal surveys keep a statistical relevance for cross-sectional analyses as well. For instance, a certain variable x may have a direct connection with the target variable y and, as a consequence, a specific cross-sectional check is needed. In this case, a troublesome decision concerns the priority level among the cross-sectional and the longitudinal checks, even though the last ones should come first. Thus, it is important to coordinate them in order to avoid the risk to oversize the overall number of checks as well as the amount of changes carried out on the original micro-database (Granquist and Kovar, 1997). On the other hand only cross-sectional checks may be applicable in case of “new” units, for which no past data are available.
- 2) Given the target parameter and the characteristics of the variable under investigation, at each reference time t there is the need to specify which are the previous periods to be considered in the editing process. For example, for monthly data the periods $t-1$ and $t+1$ or $t-12$ and $t+12$, most of the times because of the presence of significant seasonal components.
- 3) Economic units may change their demographic features over time (such as change of their ownership, location, economic activities carried out, number of local units, employment and so on) as a result of events of different nature (i.e., mergers or splits). Statistical units interested by

these changes could lose their “longitudinal” identity and their data cannot be compared in a longitudinal data analysis process. As a consequence suspected changes may come up, which are not the results of real mistakes, but they are due to structural changes of the unit economic profile along time. In a longitudinal survey context – in particular, in a short-term survey framework – it is often difficult: a) to identify cases when there are anomalous increases or decreases due to demographic changes and not to real measurement errors (lack of updated information even from the business register); b) to apply a proper amendment to microdata able to overcome the non-comparability of data over time.

- 4) In a short-term survey framework, the required timeliness for the elaboration of the indicators becomes a hard constraint for the editing strategy, as it strongly reduces the available time to check all the microdata. It is a good solution to identify a sub-set of “critical” units, for which a deeper analysis can guarantee the required quality. This approach is generally defined as *selective editing*, which presumes the definition of a *score function* to rank the observations according to their impact on the target estimates; see the module “Statistical Data Editing – Selective Editing”. Several score functions are proposed in literature, the difference among them is mainly given by the way to measure the impact on the final estimates, that anyway usually depends on: i) the given sampling weights; ii) the size of the possible error; iii) the longitudinal behaviour of each respondent.

2.4 Type of edits

The error detection process usually consists of a set of integrated error detection methods dealing each with a specific type of error (EDIMBUS, 2007), which results are flags pointing to missing, erroneous or suspicious values. Error detection is often based on the use of edit rules, that are restrictions to the values of one or more data items that correspond to missing, invalid or inconsistent values potentially in error (cf. “Statistical Data Editing – Main Module”). In a longitudinal context, the coherence of individual historical data is the basic rationale to analyse the data, because the units are believed to be strongly characterised by their own longitudinal profile. According to this point of view, the data of each unit at the occasion t can be checked by comparison with other values observed on the same unit at other times, i.e., belonging to its profile, with regards to an expected value or range.

In the following, the typology of edits is described according the needs and the features of a longitudinal context:

- Consistency checks: their purpose is to detect whether the value of two or more variables on the same unit are in contradiction, hence, whether the values of two or more data items do not satisfy some predefined expected relationship. In this regard, comparisons with other sources which produce comparable microdata are included. Data items can refer also to measurement on the same unit in different periods, it is important that this reference data has been previously checked for errors¹. The reference data used and the way in which the comparison takes place depend on the target parameter.

¹ If the past value y_{t-k} refers to the previous year, past data can be supposed to have been fully checked on the basis of information available from sources external to the survey, so that normally suspect ratios y_t/y_{t-k} lead to change the actual value y_t (but not the past value). However, this rule is not rigid and past data may be changed as well (that is the case of wrong reporting by some units which can review past values even one year later).

- Balance edits: often the value of a variable at time t can be obtained by the sum of the values in the previous period and the registered flow in the reference period for that variable; e.g., the number of persons employed at the end of month $t-1$, plus the number of persons who started working between months $t-1$ and t , minus the number of persons who stopped working between months $t-1$ and t , must be equal to the number of persons employed at the beginning of month t .
- Check for unity measure errors: some errors are due to misunderstandings about the measure according to which a variable x is collected, e.g., thousand instead of billion and so on. In these cases, there is a thousand-error if one of the following relations is verified:

$$\text{abs}(x_t) > h \cdot [\text{abs}(x_{t-k})] \quad \text{for some } k \in \{1, \dots, P\} \quad (1a)$$

$$h \cdot [\text{abs}(x_t)] < \text{abs}(x_{t-k}) \quad \text{for some } k \in \{1, \dots, P\} \quad (1b)$$

where $x_{t-k} > 0$, $\text{abs}(x)$ is the absolute value of the variable x and h is a constant to be chosen properly by the expert.

- Ratio edits. These edit rules are bivariate restrictions taking the general form $a \leq x / y \leq b$, where x and y are numerical variables and a and b are constants. In a longitudinal context, the comparison is based on the two measurements y_t and y_{t-k} , k will vary according to case under study (type of data, characteristics of the variable, etc.).
- A further type of edit is related to a specific feature of longitudinal surveys, because it is possible to ask twice for the same data, with reference to the same variable for the same period. Normally, it happens when a certain value is asked in two consecutive waves at times $t-1$ and t . Let $y_{it(t-1)}$ be the value of the variable y on the unit i asked in the wave t even though referred to the $t-1$ period, then a frequent longitudinal check is given by:

$$y_{it(t)} = y_{it(t-1)} \quad (2)$$

This option may help both to check for the quality of supplied longitudinal information and to take under control changes of some accounting figures inside the unit; it is also very useful to achieve longitudinal data from units characterised by wave non response, e.g., those units which may be non-respondent in $t-1$ and respondent in t , or vice-versa. This solution has to be defined accurately, in order to be worth without increasing the statistical burden on the respondent units.

2.5 *Methods for longitudinal data*

In a longitudinal context, one of the most relevant test variables is the “individual trend” or “individual change”, defined as:

$$c_{it} = y_{it} / y_{it-k} \quad (3)$$

As a consequence most data controls are based on the study of (3) and on rules to check whether the individual trend is too large or too low. The main issue is to define a criterion to decide whether a given level satisfies or not the acceptance rules. The unit trend information can be used in different ways, a couple of them is shortly resumed as follows.

2.5.1 *The Hidioglou-Berthelot method for detecting outliers*

The empirical distribution of all the individual trends can supply useful information for the editing process, by comparing each c_{it} with some main indicators of such distribution. In this regards, the

Hidioglou-Berthelot method (Hidioglou and Berthelot, 1986) proposes a way to establish an acceptance interval for c_{it} , based on a function of its interquartile, in order to detect outliers.

Firstly, for each occasion t the median of all the c_{it} is elaborated, defined as $q_{0.5}(c_t)$. Afterwards, a transformation is applied to every c_{it} , to ensure more symmetry of the distribution tails:

$$s_{it} = \begin{cases} 1 - q_{0.5}(c_t)/c_{it}, & \text{if } 0 < c_{it} < q_{0.5}(c_t) \\ q_{0.5}(c_t)/c_{it} - 1, & \text{if } c_{it} \geq q_{0.5}(c_t) \end{cases} \quad (4)$$

Let also define:

$$E_{it} = s_{it} \cdot \{\max(y_{it}, y_{it-1})\}^U \quad (5)$$

which is the “effect” concerning unit i at time t ; it is based on the “individual trend” component s_{it} defined by (4) and the “size” component due to the y -levels of the same unit. The parameter $U \in [0, 1]$ is a tuning parameter which should balance the magnitude of the size component with respect to the individual trend. Then, given the first and the third quartile, $q_{0.25}(E_t)$ and $q_{0.75}(E_t)$, the following values are defined:

$$D_1 = \max \{q_{0.5}(E_t) - q_{0.25}(E_t), A \cdot q_{0.5}(E_t)\} \quad (6)$$

$$D_3 = \max \{q_{0.75}(E_t) - q_{0.5}(E_t), A \cdot q_{0.5}(E_t)\} \quad (7)$$

where the constant A is chosen to avoid difficulties which can arise when the differences $q_{0.5}(E_t) - q_{0.25}(E_t)$, and $q_{0.75}(E_t) - q_{0.5}(E_t)$ are small (generally it is set to 0.05).

Hence, the acceptance region is defined as follows:

$$(q_{0.5}(E_t) - A \cdot D_1, q_{0.5}(E_t) + A \cdot D_3) \quad (8)$$

and each observation y_{it} which falls out of such interval is considered to be an outlier.

It is worthwhile to underline how the identification of anomalous ratios c_{it} due to errors (not necessarily outlier observations) may be carried out according to an analogous methodological scheme.

2.5.2 Score functions ranking

In case a selective editing scheme has to be defined, the basic rationale is the evaluation of the impact of the change of each unit on the overall trend, considering its size and its sampling weight. This kind of analysis can be carried out ranking the units on the basis of a score function, which takes into account the above mentioned dimensions. Thus, a simple score function to be applied to each unit depends on the three dimensions:

$$\text{Score} = (\text{longitudinal trend}) \times (\text{sampling weight}) \times (\text{size}).$$

In the following, a score function is described that takes these elements into account, for which a transformation of the individual trend c_{it} is defined in order to take into account different options of needs. A preliminary transformation is made to assign high priority to units characterised by either a very high or a very low change:

$$d_{ij} = \max(c_{it}, 1/c_{it}) = \max(y_{it}/y_{it-k}, y_{it-k}/y_{it}) \quad (9)$$

New units, for which no historical data are available, will be assigned $c_{it}=1$.

Then, the following conversion will be used to define the final score function:

$$r_{it} = |k_{1i}d_{it} - k_{2i}|$$

where k_{1i} and k_{2i} can be chosen according to any needs expressed by the given survey, a typical choice is to put both k_{1i} and k_{2i} equal to 1.

Thus, the score function for a generic unit i and a given time t can be built up as follows:

$$\Phi_{it} = r_{it}^{\alpha} w_{it}^{\beta} z_{it}^{\gamma} \quad (10)$$

where w is the sampling weight and z is a “size” variable (for instance, turnover, production, number of persons employed). Parameters α , β and γ should be used in order to balance the relative importance of each score component on the final score Φ . Normally it is recommended to use parameter values chosen from the interval [0,1] (Gismondi and Carone, 2008). After the calculation of the score (10) for each unit, scores can be ordered in a non-decreasing ranking: the units occupying the “first positions” in the ranking will be detected as influent suspicious units, to be checked with priority or even re-contacted. Some techniques for assessing the number of influent units have been proposed by McKenzie (2003), Philips (2003), Chen and Xie (2004).

2.6 The case of categorical data

There are particular kinds of business longitudinal surveys for which categorical variables play a fundamental role. That may happen when the main goal:

- a) is still the evaluation of the change of a quantitative variable, but a preliminary step consists in the assessment of the presence (or absence) of a certain phenomenon (binary variable: 1=present, 0=absent);
- b) consists in the evaluation of a set of opinions and their developments over time (qualitative variables).

An example of the kind a) is the survey on job vacancies. The main goal is the estimation of the number of job vacancies at the end of each quarter, but a preliminary step consists in assessing if an enterprise is searching for new personnel or not. There are the following possibilities:

- The firm declares an amount of job vacancies higher than zero, that implies the firm is searching for new staff. In this case no problem is encountered.
- The firm declares zero job vacancies. This value may be right, but it may be wrong as well, for instance, because the firm is not able to correctly count the number of job vacancies (and prefers to declare zero in order to tackle the question quickly). A signal in favor of a potential error may be given by a simple ex post longitudinal check: the comparison between the number of persons employed at times $(t+1)$ and (t) . If the former amount is higher than the latter, it is not possible that the number of job vacancies declared at time (t) was zero.
- The firm does not declare anything. Also in this case, longitudinal checks may be useful for making proper changes, but they may not be enough and the binary variable presence/absence of job vacancies will be object of estimation (for instance, using a logistic model where the explicative variables are often given by past responses provided by the same unit) or will be asked again to the firm (when it will be possible, according to budget and time constraints).

An example of the kind b) is given by tendency surveys. Tendency surveys concern enterprises and consumers and are aimed at asking a series of qualitative questions related to economic situation, household budget, purchases planning, employment, prices, etc. Questions ask for opinions concerning the development of each issue with respect to a previous period. Normally response modalities are: i) strong increase, ii) increase, iii) no change, iv) decrease, v) strong decrease. Macro figures are calculated as weighted differences between optimistic opinions i)+ii) and the pessimistic ones iv)+v). In tendency surveys main quality checks do not refer explicitly to past longitudinal data. This may be due to the use of rotated samples and/or to the weak correlation between responses provided by the same unit in two consecutive survey waves. The basic control is that for each unit and each question one and only one response must be provided.

3. Design issues

The design of the editing and imputation process should be part of the design of the whole survey process. In the frame of editing and imputation procedures three main logical phases are usually carried out, based on the following actions:

1. Identification and elimination of errors that are evident and easy to treat with sufficient reliability, that can involve both interactive and automatic methods;
2. Selection and treatment of influential errors through a careful inspection of influential observations; automatic treatment of the remaining non influential errors, through a selective editing procedure;
3. Check of the final output looking for influential errors that have been undetected in the previous phases or introduced by the procedure itself, that involves macro-editing procedures.

In a longitudinal context, the identification and the calculation of a set of indicators based on macrodata may be based on ratios between the same macrodata related to two different periods, where macrodata of the previous period are supposed to be good (already validated at previous occasions). If the macro indicator falls inside an acceptance range, then no other controls are needed, otherwise it is necessary to go back to microdata and to run again all or a part of controls already activated in the previous micro-editing phase a). Usually, acceptations intervals for macro indicators are determined according to subjective choices by survey experts.

Finally, in the last phase, provisional publication figures are elaborated and analysed using historical data or external sources. If the aggregate figures are implausible, the individual records are examined in order to check for further outliers or error affecting influential records; in these cases data can be modified if necessary. The errors detected at this stage may have been not individuated in the earlier phases of the editing process, or may have been introduced by the process itself. Anyway, also every treatment of these kinds of errors is always made at micro level. If the provisional figures are plausible, the detection of errors and their treatment process is concluded.

The edited file is used in the subsequent statistical process for aggregation purposes, for the estimation of totals and for further analyses.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Chen, S. and Xie, H. (2004), Collection Follow Up Score Function and Response Bias. *Proceedings of the SSC Annual Meeting – Survey Methods Section*, Statistics Canada, 69–76.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.

http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Gismondi, R. and Carone, A. (2008), Statistical criteria to manage non-respondents’ intensive follow up in surveys repeated along time. *Rivista di Statistica Ufficiale*, 1/2008, 5–29.

Granquist, L. and Kovar, J. (1997), Editing of Survey Data: How Much Is Enough? In: L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, 415–435.

Hidirolou, M. A. and Berthelot, J. M. (1986), Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology* **12**, 73–83.

McKenzie, R. (2003), *A Framework for Priority Contact of Non Respondents*. Available at: www.oecd.org/dataoecd.

Philips, R. (2003), The Theory and Application of the Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Proceedings of the SSC Annual Meeting – Survey Methods Section*, Statistics Canada, 121–126.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Different Types of Surveys
2. Repeated Surveys – Repeated Surveys
3. Sample Selection – Main Module
4. Statistical Data Editing – Main Module
5. Statistical Data Editing – Selective Editing
6. Weighting and Estimation – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 2.5 Design statistical processing methodology
2. 5.3 Review, validate and edit

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Data validation

Administrative section

14. Module code

Statistical Data Editing-T-Longitudinal Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-02-2013	first version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.2	30-05-2013	second version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.3	20-08-2013	third version (accepted corrections)	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4	15-11-2013	fourth version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4.1	20-12-2013	preliminary release		
0.4.2	08-01-2014	final release	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:13

Method: Data imputation using statistical roulette

0. General information

0.1 Module name

Method: Data imputation using statistical roulette

0.2 Module type

Method

0.3 Module code

Method-statistical roulette

0.4 Version history

Version	Date	Description of changes	Author	NSI
1.0	6-6-2011	First version	Andrzej Młodak	GUS (PL)

Template version used	1.0 d.d. 25-3-2011
Print date	10-8-2011 11:19

Contents

General description – Method:	3
1. Summary	3
2. General description	3
3. Examples – not tool specific	3
4. Examples – tool specific	9
5. Glossary	9
6. Literature	10
Specific description – Method:	11
A.1 Purpose of the method	11
A.2 Recommended use of the method	11
A.3 Possible disadvantages of the method	11
A.4 Variants of the method	11
A.5 Input data sets	11
A.6 Logical preconditions	11
A.7 Tuning parameters	12
A.8 Recommended use of the individual variants of the method	12
A.9 Output data sets	12
A.10 Properties of the output data sets	12
A.11 Unit of processing	12
A.12 User interaction - not tool specific	12
A.13 Logging indicators	12
A.14 Quality indicators of the output data	12
A.15 Actual use of the method	12
A.16 Relationship with other modules	12

General description – Method:

1. Summary

The module presents an original method of data imputation. It combines in some sense the random hot deck and distance hot deck methods and consists of several steps, i.e. clustering of donors into internally homogeneous subsets, definition of their representatives, optimal choice of the representative for any recipient and assignment of a relevant implant to this recipient from a donor selected from a cluster indicated by the optimal representative. The last task is realized using the mechanism of the so called ‘statistical roulette’, i.e. the probabilistic model accounting for the empirical distribution of the variable to be imputed within each cluster of donors. Some examples are also provided to show method where the method can effectively be used. Most of the content of this module is based on the original manuscript by A. Młodak (2010).

2. General description

Data shortage for an essential part of observed units still poses a serious problem. This is due to two main reasons. Firstly, a big part of units remains outside the actual random samples. Secondly, during directly conducted surveys non-sampling errors often occur. They are the result of various circumstances, e.g. refusal to answer, absence of a respondent competent to answer the survey, incomplete or imprecise answers (caused, for instance, by problems with remembering facts), etc. These inconveniences are well known to experienced researchers. Therefore, there is a need to develop effective methods to complete the missing data on the basis of collected information, which could improve also the quality of small area estimation. This technique is called *data imputation* and the values introduced to the database in this manner are said to be *statistical implants*.

Work in this field has resulted, among others, in the development of a new method of imputation which will be presented in this paper. It is in line with the general theory of imputation (cf. G. Kalton and D. Kasprzyk (1982, 1986), D. B. Rubin (1987)) and uses some special properties of practically available data. According to the typology proposed by M. D’Orazio *et al.* (2006) the new method can be regarded as a micro approach, i.e. it is aimed at receiving a synthetic and complete data set. It is universal in the sense that it can be applied to any similar statistical survey and data coming from various sources and it is useful both in the context of *mass* imputation (using implants from a relatively small sample) and in the supplementary completion of missing data (*order* imputation). Its main advantage is its robustness to outliers (in terms of clustering and choice of representatives), sensitivity to all partial deviations between a recipient and a representative and the high computational efficiency. It is worth underlining that an object is regarded as an outlier if its distance from other objects in the analyzed collection is significantly bigger than mutual distances between them.

We assume that there are some common variables, which are available for all analyzed records and distinguish two disjoint subsets of them. The first set contains records for which full information is available (called *donors*). It is worth noting, however, that data for the non-shared variables may be collected from other sources than those for the shared variables. The second subset contains records where no data for non-shared variables are available, called *recipients*. Some well known and obvious imputation methods, e.g. the regression, could then be significantly biased. The proposed algorithm

consists of several steps. Firstly, the set of records, is divided into internally homogeneous and disjoint subsets according to the common variables using a special type of cluster analysis. Secondly, for each cluster, its representative is determined. This can be done by means of the Weber median or its ‘actual’ equivalent (i.e. a donor for which the sum of its distances from remaining records within a given cluster is minimal). For each record, where some data will be imputed (i.e. *recipient*), the best representative (and simultaneously, cluster) is found by minimizing the distance between a recipient and representatives computed as a maximum of partial distances between particular observations for the common variables. This current approach is much more general. Finally, the implants for imputed variables are determined using the mechanism of the ‘statistical roulette’ proposed by Professor Bogdan Stefanowicz. It is based on the model resembling the roulette wheel, where sizes of areas are proportional to respective realizations of the empirical distribution of a given variable. Because the mechanism was constructed primarily for categorical variables, possibilities of its adaptation to interval or ratio variables will also be considered. It should be underlined that only the final stage of the proposed procedure is based on the cited paper – the most important remaining part using taxonomical methods is an original contribution of the author of this study. Without these solutions, the algorithm is not effective. Therefore this paper doesn’t describe only an application.

The method can be described as follows. Let \mathbb{N} be the set of natural numbers, $m, n \in \mathbb{N}$ and $U = \{1, 2, \dots, n\}$ be a population of size n described by m statistical variables X_1, X_2, \dots, X_m . These variables can be observed using various measurement scales: nominal, ordinal, interval or ratio. In censuses the prevailing part of collected variables are nominal or ordinal (e.g.. country of birth is usually coded at the nominal scale and age groups – at the ordinal scale). Assume that for a subset $D \subset U$ of n_1 ($n_1 \in \mathbb{N}, n_1 < n$) units, all collected variables are available and in the case of remaining $n_2 = n - n_1$ units belonging to the set $B = U \setminus D$ the information is incomplete, i.e. there exists a subset of m_1 ($m_1 \in \mathbb{N}, m_1 < m$) variables, for which no information on records belonging to B is available. For simplification, without loss of generality, one can assume that the missing values concern $m_2 = m - m_1$ variables $X_{m_1+1}, X_{m_1+2}, \dots, X_m$. Then D is called a set of *donors*, and B – a set of *recipients*.

In practice, imputation in the above situation may be perceived in two ways. On the one hand, it can be performed to fill the existing gaps resulting mainly from systematic errors appearing only in some records and then it is said to be *order imputation*. This is the case when the number of recipients (n_2) is substantially lower than the total population size (n). Our long-time experience in conducting various surveys allows us to estimate that, on average, the percentage of nonresponse records amounts to about 20%. On the other hand, the m_2 variables $X_{m_1+1}, X_{m_1+2}, \dots, X_m$ may be collected only in a sample survey, and therefore $n_2 \ll n$. Moreover, because it is assumed that the recipients were not sampled, the source of data for X_1, X_2, \dots, X_{m_1} must be quite different altogether (e.g. administrative database). This is the second situation, leading to the necessity of *mass imputation*. Taking these circumstances into account, we propose an original method consisting of several stages. Now let us present the introductory steps, i.e. clustering of donors.

The first step of the algorithm consists in clustering a set of donors into internally homogeneous and mutually heterogeneous disjoint clusters. They will reflect particular groups of information. The current methodology of cluster analysis (cf. e.g. B. S. Everitt *et al.* (2011)) proposes two possibilities of choosing the basis of clustering – it can start either from the direct data matrix, possibly normalized, or from the matrix of distances between particular objects constructed on the basis of these data. Due to

the fact that (as we have already mentioned) the observed variables can be often nominal or categorical, where performing arithmetic operations (such as averaging) has practically no methodological sense, the second option seems to be much better.

In this case it is very important to have a method of computing such a distance to account for the character of the variable. The Gower's distance formula seems to be an effective solution to this problem. Assuming that the record $i \in D$ can be represented as $\gamma_i = (x_{i1}, x_{i2}, \dots, x_{im})$, this distance is defined to be

$$d_G(\gamma_i, \gamma_k) = 1 - \delta_G(\gamma_i, \gamma_k), \quad (1)$$

where $\delta_G(\gamma_i, \gamma_k) = \sum_{j=1}^m w_j \rho_{ikj} / \sum_{j=1}^m w_j$, and w_j is the weight associated with the variable X_j , and ρ_{ikj} denotes the Gower's probability measure established as:

– if X_j is nominal, then

$$\rho_{ikj} = \begin{cases} 1 & \text{if } x_{ij} = x_{kj}, \\ 0 & \text{if } x_{ij} \neq x_{kj}, \end{cases}$$

– if X_j is ordinal, interval or ratio, then

$$\rho_{ikj} = 1 - |x_{ij} - x_{kj}|$$

for every $j = 1, 2, \dots, m$ and $i, k \in D$, $i \neq k$.

This way, each variable is treated according to its performance. Moreover, the obtained measure reflects a practical sense of the distance and can be easily interpreted. However, no special weighting will be introduced; that is, we assume that all weights are equal to 1 (i.e. $w_j = 1$ for every $j = 1, 2, \dots, m$).

Like the method of distance measure, the clustering procedure is also very important. We adopt here the 'elastic beta' algorithm proposed by G. N. Lance and W. T. Williams (1967), which is a recursive hierarchical method based on the Wrocław taxonomy (also called *single linkage*) (cf. K. Florek *et al.* (1951)), where the distance between clusters at the level u is defined by the distance of clusters at the level $u-1$, $u = 2, 3, \dots, n_1$. Denote by $d_{P_u}(P_{uh}, P_{ug})$ the distance between clusters P_{uh} i P_{ug} at the level u . In this special situation our procedure starts from a set of trivial one-element clusters (i.e. at the level $u=1$, where each record is regarded as an independent set and the distance between these clusters is computed using the formula (1)), and next at each level $u = 2, 3, \dots, n_1$ such clusters P_{ug} i P_{uk} are merged, which minimizes the distance expressed by the formula:

$$d_{P_{u+1}}(P_{[u+1]h}, P_{ug} \cup P_{uk}) = (d_{P_u}(P_{uh}, P_{ug}) + d_{P_u}(P_{uh}, P_{uk})) \cdot \frac{1-b}{2} + b \cdot d_{P_u}(P_{ug}, P_{uk})$$

for every $h, g, k = 1, 2, \dots, p_u$, $h \neq g, k$, where p_u is the number of clusters at the level u , and $u = 1, 2, 3, \dots, n_1$. The parameter b can be determined in various ways, more often $b := -0,25$. G. Milligan (1989) suggests using a lower coefficient like $b := -0,5$ if there are outliers among the analyzed data. It enables us to increase the robustness of the algorithm to merging clusters containing such observations. If outliers are theoretically possible, but one cannot expect that their number is relatively large, we recommend setting the value of $b := -0,3$. One advantage of the 'elastic beta' method that has

practical significance is that – unlike many others – at no stage does it use any arithmetic operations that are not allowed at some measurement scales.

The last step of the clustering consists in the definition of the threshold of clustering and, by the same token, the indication of its interruption point. That is, we should establish the value q of the distance of clusters joined at each stage such that hierarchical aggregation of the classes is continued while $d_u^* \leq q$, where d_u^* is the (optimal) distance of clusters merged at the step $u = 1, 2, 3, \dots, n_1$; then the aggregation is stopped and the obtained division is regarded to be final. This threshold ensures optimal (i.e. the most homogeneous and heterogeneous) clustering of quantile classes. To avoid the unfavorable impact of outliers, we assume

$$q = \text{med}(\mathbf{d}^*) + 2,5 \cdot \text{mad}(\mathbf{d}^*), \quad (2)$$

where $\mathbf{d}^* = (d_1^*, d_2^*, \dots, d_{n_1}^*)$ is the vector of minimum of distances at successive steps of clustering, $\text{med}(\mathbf{d}^*)$ – its median, and $\text{mad}(\mathbf{d}^*) = \text{med}_{u=1,2,\dots,n_1}(|d_u^* - \text{med}(\mathbf{d}^*)|)$ – is its median absolute deviation. The choice of the threshold (2) is justified also by its tendency to reduce the number of clusters, which is important from the point of view of the computational efficiency of the algorithm. The constant 2.5 is called the *threshold value of robustness* (cf. P. J. Rousseeuw and A. M. Leroy (1987)).

The set of donors divided into internally homogeneous and mutually heterogeneous clusters is an effective source of information necessary for data imputation for recipients. This efficiency consists in that instead of analyzing the whole (often very broad) set of donors, it is sufficient to select only a group which is closest to a given recipient. It significantly shortens the duration and computation cost.

To meet this condition, it is recommended that for each group of donors its ‘representative’ is chosen, i.e. a ‘true’ or artificial record, which seems to be the most ‘typical’ of such a group. This can effectively be obtained by means of the Weber median of a cluster, i.e. the multivariate generalization of the classical median notion (cf. A. Młodak (2006)). This is a vector which minimizes the sum of Euclidean distances from given points representing the analyzed objects. More formally, we look for a vector $\gamma_P = (\gamma_{P1}, \gamma_{P2}, \dots, \gamma_{Pm})$ such that

$$\sum_{i \in P} \left(\sum_{j=1}^m (x_{ij} - \gamma_{Pj})^2 \right)^{1/2} = \min_{\theta \in \mathbb{R}^m} \sum_{i \in P} \left(\sum_{j=1}^m (x_{ij} - \theta_j)^2 \right)^{1/2},$$

where $P \subseteq D$ is a given cluster of donors.

The advantage of this choice is that the Weber median lies to some extent ‘in the middle’ of the donors and simultaneously is robust to the existence of outliers. However, sometimes its direct application is impossible from the practical point of view. This situation may occur if some variables are nominal or ordinal and then two following problems appear:

- the construction of the Weber median is based on the Euclidean distance, which owing to the nominal or ordinal character of some variables is inappropriate,
- the automatically generated vector is rather artificial from the practical point of view and therefore its coordinates are usually measured at the ratio scale, whereas in the analyzed situation we expect to have information reflecting the character of particular variables (e.g. if a variable can take only values 0 or 1, the Weber median could assume the coordinate of 0.5).

Taking the above into account, finding a solution closest to the Weber median and simultaneously effective in the case of existing categorical variables will involve determining in each group of donors $P \subset D$ a record, for which the sum of Gower's distances from the remaining records belonging to this set is the smallest, i.e. such a vector γ_P , that $\sum_{i \in P} d_G(\gamma_i, \gamma_P) = \min_{k \in P} \sum_{i \in P} d_G(\gamma_i, \gamma_k)$. This vector will be regarded as a *representative* of the cluster P .

Next, for every recipient we find a representative which is closest to it with respect to the available variables. That is, we determine a representative for which the maximum of its partial Gower's distances (in terms of the known variables) from a given record-recipient is the smallest. More formally, for any $i \in B$ we find a set $P^* \subset D$ belonging to the given cluster structure such that $\tilde{d}_G(\gamma_i, \gamma_{P^*}) = \min_P \tilde{d}_G(\gamma_P, \gamma_i)$, where $\tilde{d}_G(\gamma_P, \gamma_i) = \max_{j=1,2,\dots,m_2} w_j(1 - \rho_{Pij})$. The assumptions and symbols adopted here are the same as in the formula (1). Of course, one can use the traditional Gower's distance in this context, but as previous experience in applying this method shows, the results were so divergent that it was necessary to apply an additional criterion specific to the currently analyzed data. It is possible because some partial deviations are compensated for by other ones, of a reverse nature. This approach is much more general and stable.

The optimum representative for a given recipient indicates the best cluster of donors where the most effective implant should be looked for. Having it at our disposal, we can supplement the missing data with information derived from members of such a group. It is conducted by a random hot desk type mechanism, called the *statistical roulette*, proposed by professor Bogdan Stefanowicz where the random choice of implants is based on empirical approximation of distribution of imputed variables within a given cluster of donors. Its design for the record $i \in B$ can be presented as follows:

1) construction of the 'roulette wheel':

- we assume that the hypothetical 'wheel' of the roulette is of the length 1,
- let us suppose that the variable X_j to be imputed for recipients can take r possible values $a_{j1}, a_{j2}, \dots, a_{jr}$; the perimeter of the wheel is divided into r ($r \in \mathbb{N}$) segments – each of them is associated with one of possible values of X_j . The length of each segment (t_{js}) is fixed to be the frequency of the observation a_{js} in the empirical distribution of X_j restricted to the members of the group P_i optimal for the recipient i , i.e. $t_{js} = \frac{f_{jP_i}(a_{js})}{|P_i|}$, where $f_{jP_i}(a_{js})$ denotes the number of observations of a_{js} for X_j within the group P_i , and $|P_i|$ is its cardinality. We have, of course, $0 \leq t_{js} \leq 1$ for $s = 1, 2, \dots, r$ and $\sum_{s=1}^r t_{js} = 1$,
- the start point of the roulette is established at 0, and next the beginning of the s -th segment, q_{js} , is determined. It is done in the following way: $q_{j1} = 0$, $q_{js} = \sum_{z=1}^{s-1} t_{jz}$, $s = 2, 3, \dots, r$;

2) activation of the 'wheel':

- from a set of random numbers we choose a random number λ belonging to the interval $[0, 1]$. From the computer science point of view, the simplest method of doing this consists in running the random number generator from the uniform distribution of $[0, 1]$. Then the number λ reflects the distance from the point 0 on the perimeter of the wheel and simulta-

neously indicates uniquely a segment of this ‘wheel’, from which the implant will be chosen.

- let then $s \in \{1, 2, \dots, r\}$ be such that $q_{js} \leq \lambda \leq q_{j(s+1)}$; if $\lambda < (q_{j(s+1)} - q_{js})/2$, then we take the implant $x_{ij} := a_{js}$, and otherwise we put $x_{ij} := a_{j(s+1)}$.

This operation of construction and activation of the ‘roulette’ is repeated for every $j = m_1 + 1, m_1 + 2, \dots, m$ and every $i \in B$.

As one can see, the above mechanism is based on the assumption that the imputed variables are nominal or ordinal. In practice, however, a situation may occur, when some variables are interval or ratio. Therefore we should consider also a modification of this algorithm aimed at satisfying this expectation. A good approach in this context is to divide the within-cluster span of X_j (i.e. the interval $[\min_{l \in P_i} x_{lj}, \max_{l \in P_i} x_{lj}]$) into a large number $r \in \mathbb{N}$ (e.g. $r = 1000$) of disjoint intervals $[c_0, c_1], [c_1, c_2], \dots, [c_{r-1}, c_r]$, where $\min_{l \in P_i} x_{lj} = c_0 < c_1 < c_2 < \dots < c_{r-1} < c_r = \max_{l \in P_i} x_{lj}$ (it is convenient to assume additionally that $c_{s-1} - c_{s-2} = c_s - c_{s-1}$ for every $s = 2, 3, \dots, r$) and association of each observation with the identifier of interval which it belongs to. For instance, if $x_{lj} \in [c_{s-1}, c_s]$, then we put $a_{js} := s$, $s \in \{1, 2, \dots, r\}$. The value obtained by the roulette procedure indicates the interval which the implant should be taken from and, with the centre of this interval selected as the implant, i.e. if $x_{ij} := a_{js} = s$, then the final implant, \hat{x}_{ij} , is of the form $\hat{x}_{ij} := (c_{s-1} + c_s)/2$, $s \in \{1, 2, \dots, r\}$, for $j = m_1 + 1, m_1 + 2, \dots, m$ and for every $i \in B$.

The particular steps (distance weighting, clustering threshold, etc.) should be adjusted depending on the purpose of research, type of available data and properties of the analysis model. However, one can consider in this context also some alternative methods of clustering such as quantile grid (cf. A, Młodak (2011)).

3. Example – not tool specific

The method of the “statistical roulette” seems to be new and therefore it has so far been tested in a simulation study using census data (this is why it was primarily constructed to support the 2011 National Population and Housing Census in Poland). To verify the efficiency of the introduced methodological solutions, a simulation experiment was carried out using data collected during the National Population and Housing Census conducted in 2002. One of the most important questions and challenges to appear in the course of the experiment concerned the choice of the spatial level of data aggregation. The performed trials showed that filtering potential donors from the main file and constructing the distance matrix for them in the case of a larger database leads to computational difficulties. The most serious problem concerns the processing of the distance matrix. In the case of higher level units (i.e. NUTS 2, NUTS 3 and NUTS 4), which in Poland include at least about 100,000 records, a lot of memory space is required (the distance matrix should have about $100,000 \times 100,000$ elements), which is often difficult, although the computer tools used in the study (SAS Enterprise Guide 4.1., next 4.2.) are much more efficient than others. To enable fast computation, it was decided that the current analysis will be conducted at the level of gmina (NUTS 5 unit), where it could be completed relatively quickly. Therefore, the proposed algorithm was observed to be more effective than many

other commonly available (e.g. traditional hot desk or relevant procedures written in the R software environment).

For purposes of the simulation, a gmina (NUTS 5) unit with a total of 9630 records was chosen. Using the simple random sampling without replacement and with equal probabilities of sampling, three samples of size 5%, 10%, 20% were drawn. They were regarded further to be the sets of recipients and the remaining records were treated as donors. This setup was an example of the order imputation model. In the case of testing mass imputation these assumptions were reversed, i.e. the samples constituted a set of donors and the other records were regarded as recipients. The set of variables collected in the described database contains data on general categories of the population (sex, age date of birth, place of residence, etc.), civil status, education, information on disability, country of birth, citizenship, nationality and language used. It was assumed that these data will be potentially available, because they are contained in administrative registers. Others (such as education or actual disability) can be obtained only by personal interview and therefore (due to the possible occurrence of some of the above mentioned errors) will probably have to be imputed. During the simulation the relevant data were removed from recipient records.

Using data sets prepared in this way the whole procedure described in chapters 1 and 2 was performed. Although the numbers of donor clusters obtained as a result seem to be rather large (e.g. out of 9147 records from a 5% sample, which formed a set of donors, 1114 clusters were created) but they were the smallest quantities which could be obtained in an endogenous way, i.e. when the threshold of clustering was established as a statistic of the elements of a distance matrix. The best result in this context was obtained using approach (2).

To assess the quality of the final results, a comparative analysis of the imputed data for the ‘missing’ variables with their true values existing in the primary database was conducted. This study was carried out in 2 stages:

- 1) for every recipient record the distance between its version with true data for imputed variables and its option with the implants was computed; the averaged Gower’s distance was used, i.e. the formula (1) was applied, where $\delta_G(\gamma_i, \gamma_{ir}) = \sum_{j=m_1+1}^m \rho_{ijr}$ and γ_i is then the vector with ‘true’ data for the record i and γ_{ir} – the vector containing relevant imputed data for this record, $i \in B$,
- 2) for imputed and ‘true’ variables the summary statistics (i.e. total values for the whole population) were determined, and next differences between these structures were compared using the Student’s t-test, verifying the hypothesis on their non-significance. To improve the efficiency of our analysis, sign and Wilcoxon’s signed rank tests were used as well.

The experiment showed that the method is very effective and quick in terms of computation. The precision of imputation was also satisfactory. For more details see A. Młodak (2010),

4. Examples – tool specific

It is easy to see that this method is universal. In other words, it can be effectively applied in business statistics. To do it, firstly we must investigate which data on business statistics are available in administrative sources. One can expect that they will concern history, ownership and employment in the

entity (business registers) and revenue subject to taxation as well as tax payments (tax registers). Moreover, if an entity conducts international trade, then some additional information could be available in the INTRASTAT system.

Other information, such as structure of fixed assets or capital relations as well as accidents at work or strikes will usually require imputation. This procedure will be necessary mainly in the case of small enterprises which are usually investigated in sample surveys and therefore mass effective data imputation for units not drawn to the sample is required.

5. Glossary

Term	Definition
NUTS	Classification of Territorial Units for Statistical Purposes, a hierarchical system for dividing up the economic territory of the EU for the purpose of the collection, development and harmonization of EU regional statistics, socio-economic analyses of the regions and framing of EU regional policies. See http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction
R	R package; a free software environment for statistical computing and graphics consisting of several thousand modules enabling professional and highly specialized statistical research and analysis, See http://www.r-project.org/
SAS	Former Statistical Analysis System, a complex software supporting statistical surveys, data analysis and dissemination of results (including business analysis and business intelligence) See http://www.sas.com
INTRASTAT	The statistical system covering trade between EU Member States and based on data obtained from INTRASTAT declarations see e.g. http://www.stat.gov.pl/gus/5840_574_ENG_HTML.htm

6. Literature

Everitt B. S., Landau S., Leese M., Stahl D. (2011), *Cluster Analysis*, 5th Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester. UK.

Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951), *Sur la Liaison et la Division des Points d'un Ensemble Fini*, Colloquium Mathematicae vol. 2, pp. 282 – 285.

Kalton, G. and Kasprzyk, D. (1986), *The Treatment of Missing Survey Data*, Survey Methodology, vol. 12, pp. 1 – 16.

Kalton G. and Kasprzyk D. (1982), *Imputing For Missing Survey Responses*, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 22 – 31 (http://www.amstat.org/sections/srms/proceedings/papers/1982_004.pdf).

Lance, G. N. and Williams, W. T. (1967), *A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems*, Computer Journal, vol. 9, pp. 373 – 380.

Młodak A. (2011), *Classification of Multivariate Objects Using Interval Quantile Classes*, Journal of Classification, vol. 28 (to appear).

Młodak A. (2010), *A Method of Data Imputation Using the Statistical Roulette*, manuscript, submitted for publication.

Młodak A. (2006), *Multilateral normalizations of diagnostic features*, Statistics in Transition, vol. 7, pp. 1125–1139.

D’Orazio, M., Di Zio, M. and Scanu, M. (2006), *Statistical Matching. Theory and Practice*, John Wiley & Sons, New York.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Specific description – Method:

A.1 Purpose of the method

The method using cluster analysis and the ‘statistical roulette’ is aimed at efficient imputation of missing data for some records when other data are available for all records in the analyzed database. The imputed data are taken from those records for which they are complete.

A.2 Recommended use of the method

1. The method is especially useful for variables of various type and measures at various measurement scales (nominal, ordinal, interval or ratio).
2. Due to the possibility of adjustment of several steps to a given situation the algorithm seems to be flexible and offers users a wide scope of choice with respect to the submethods and parameters adequate to given data and purposes.
3. It is computationally efficient for larger sets of donors and recipients.

A.3 Possible disadvantages of the method

1. Sensitivity to outliers in some cases of possible clustering thresholds.
2. Threat of an ineffective choice of weights in the distance formula or threshold of clustering.

A.4 Variants of the method

1. Various alternative to the Gower’s distance formula can be used.
2. Various weighting in the Gower’s distance formula can be used depending on a situation
3. The method can be applied using various cluster algorithms.
4. Various thresholds used to optimize collection of cluster can be applied.

A.5 Input data sets

1. Sets of donors – from which data will be imputed.
2. Set of recipients – where data will be imputed.

A.6 Logical preconditions

A.6.1 Missing values

1. Allowed, but the larger their number, the lower the imputation precision.

A.6.2 Erroneous values

1. Rather not allowed, they negatively affect the results of imputation; however, if they are identified, they can be removed and the respective records are then tested as recipients, taking the remark A.6.1. into account

A.6.3 Other preconditions

1. Proper choice of clustering methods.

A.7 Tuning parameters

1. A calibration of weights used in the Gower's distance may be desirable in special situations.

A.8 Recommended use of the individual variants of the method

1. The effective parameters to be used were proposed in the description. Other solutions should be developed in relevant case studies.

A.9 Output data sets

1. Records with imputed data

A.10 Properties of the output data sets

1. The output set can be a good basis for estimation of required data for small areas at various territorial levels.

A.11 Unit of processing

Processing groups of units

A.12 User interaction - not tool specific

1. Not applicable

A.13 Logging indicators

1. Variables available in administrative data sources and obtained in sample surveys (the latter only for sampled records)

A.14 Quality indicators of the output data

1. The precision seems to be satisfactory

A.15 Actual use of the method

1. Data imputation in the censuses

A.16 Relationship with other modules

A.16.1 Themes that refer explicitly to this module

1. Weighting

2. Quality assessment
3. Estimation
4. Sampling

A.16.2 Related methods described in other modules

1. Calibration
2. Cluster analysis

A.16.3 Mathematical techniques used by the method described in this module

1. Gower's distance
2. 'Elastic beta' single linkage algorithm
3. Choice of representatives of clusters
4. Usage of uniform distribution

A.16.4 GSBPM phases where the method described in this module is used

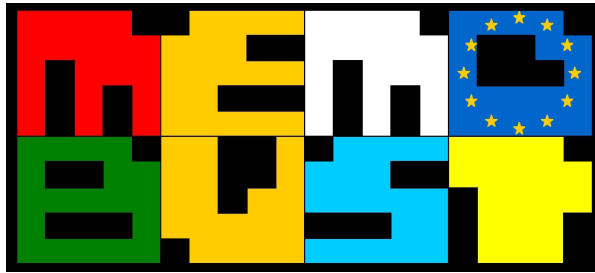
1. 5.4 Impute

A.16.5 Tools that implement the method described in this module

1. Now there is also original algorithm written un SAS Enterprise Guide by the Author of this module

A.16.6 The Process step performed by the method

Data imputation and calculation of aggregates.



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Deductive Imputation

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Simple imputation rules.....	3
2.2 The use of equality restrictions.....	3
2.3 The use of non-negativity constraints.....	5
3. Preparatory phase	6
4. Examples – not tool specific.....	6
4.1 Example: deductive imputation with equality restrictions	6
4.2 Example: deductive imputation with equality and non-negativity restrictions	8
5. Examples – tool specific.....	9
6. Glossary.....	10
7. References	10
Specific section.....	11
Interconnections with other modules.....	12
Administrative section.....	14

General section

1. Summary

In general, imputations are predictions for the missing values, based on an explicit or implicit model. In some cases, however, imputations can also be derived directly from the values that were observed in the same record, using derivation rules that do not contain any parameters to be estimated, such as is the case in models.

For instance: suppose that businesses are asked in a survey to report their total turnover (T), turnover from the main activity (T_1), and turnover from side-line activities (T_2). If the value of one of these variables is missing, and if it may be assumed that the two observed values are correct, then the missing value can be calculated using the rule: $T_1 + T_2 = T$.

The above imputation rule is an example of *deductive* or *logical imputation*. In this imputation method, one identifies cases where it is possible, based on logical or mathematical relationships between the variables, to unambiguously derive the value of one or more missing variables from the values that were observed, under the assumption that the observed values are correct. For the missing variables for which this is possible, the uniquely derived value is the deductive imputation. The assumption that all observed values are correct requires that all erroneous values in the original data have been removed in a previous process step.

2. General description of the method

2.1 Simple imputation rules

Many deductive imputations can be performed using simple rules in ‘if-then’ form, for example:

if (*total labour costs* = ‘missing’ **and** *employees on the payroll* = 0)
then *total labour costs* := 0.

These rules are compiled by subject-matter experts. They can be applied with many different types of software.

In the remainder of this section, we discuss two methods that generate deductive imputations automatically based on restrictions that must be satisfied by the data. These methods work only for numerical data. A similar method for categorical data is given by De Waal et al. (2011, Section 9.2.4), but we do not discuss this method here, because business surveys usually involve numerical data.

2.2 The use of equality restrictions

A particularly rich source for deductive imputations is formed by the extensive systems of equations that should hold for Structural Business Statistics. A typical survey may involve around 100 variables with 30 equality restrictions. Most of these equality restrictions have the general form

$$\text{Total} = \text{Subtotal}_1 + \text{Subtotal}_2 + \dots + \text{Subtotal}_s. \quad (1)$$

If, in such a case, one of the subtotals or the total is missing, it is immediately clear with which value the missing variable should be imputed: there is a single equation with a single unknown, so a unique solution exists.

More generally, we may encounter several variables with missing values that are involved in several inter-related equality restrictions. This means we have a system of equations with multiple unknowns, for which it is not immediately clear whether the values of some missing variables are uniquely determined by this system, and, if so, what these unique values would be. However, this problem may be solved using techniques from linear algebra. Below we describe a method that automatically generates the deductive imputations from a given system of equations. This description is based on Pannekoek (2006).

Suppose that a record consists of p variables and that q linear equality restrictions apply to these p variables. The restrictions may be represented in the form

$$\mathbf{R}\mathbf{y} = \mathbf{b}, \quad (2)$$

where \mathbf{y} is a vector of length p with the variables, \mathbf{b} is a vector of length q with constant terms that appear in the restrictions, and \mathbf{R} is a $q \times p$ matrix in which each row represents one restriction and each column represents one variable. For example, consider a business survey where the operating income block consists of the following five variables:

Net turnover from main activity	y_1
Net turnover from other activities	y_2
Total net turnover	y_3
Total other operating income	y_4
Total operating income	y_5

Two restrictions apply to these variables: $y_1 + y_2 = y_3$ and $y_3 + y_4 = y_5$. These restrictions can be formulated as a system of equations in the form (2) with $\mathbf{b} = \mathbf{0}$ and

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}.$$

If the vector with variables \mathbf{y} consists of p_o observed values and p_m missing values, then, after a permutation of elements, this vector can be partitioned as $\mathbf{y} = (\mathbf{y}'_o, \mathbf{y}'_m)'$, in which \mathbf{y}_o is a vector of length p_o with the observed values and \mathbf{y}_m is a vector of length p_m with the missing values. If we partition \mathbf{R} accordingly, we can write:

$$\begin{bmatrix} \mathbf{R}_o & \mathbf{R}_m \end{bmatrix} \begin{bmatrix} \mathbf{y}_o \\ \mathbf{y}_m \end{bmatrix} = \mathbf{b},$$

so that, say,

$$\mathbf{R}_m \mathbf{y}_m = \mathbf{b} - \mathbf{R}_o \mathbf{y}_o = \mathbf{a}. \quad (3)$$

Note that \mathbf{a} can be computed using only the observed values in the record. Thus, expression (3) is a system of linear equations that involves only the missing variables \mathbf{y}_m . The intention of deductive imputation is to derive as many missing values as possible from this system.

For a system of linear equations, one usually distinguishes between three cases:

- I) There are no solutions (the system is inconsistent);
- II) There is exactly one solution;

III) There are an infinite number of solutions.

For system (3) – assuming that the original restrictions in (2) do not contradict each other –, Case I can only occur if there are errors in the observed values. We assume here that all errors have been detected previously and replaced by missing values. Moreover, if this has been done using error localisation methodology as described in the module ‘Automatic Editing’, then it is certain that the missing values can be imputed in such a way that the restrictions are satisfied. Thus, under these assumptions, Case I cannot occur.

Case II occurs if \mathbf{R}_m is a matrix with rank equal to the number of missing values p_m . In the special case that \mathbf{R}_m is square, the unique \mathbf{y}_m that satisfies the restrictions is given by

$$\tilde{\mathbf{y}}_m = \mathbf{R}_m^{-1} \mathbf{a},$$

where \mathbf{R}_m^{-1} denotes the inverse matrix of \mathbf{R}_m . If \mathbf{R}_m is not square, we can still obtain a unique solution in this form after a suitable transformation of \mathbf{R}_m and \mathbf{a} to remove any linear dependent rows. Thus in Case II, all missing variables can be imputed deductively, since all missing values are uniquely determined by the system of equations and the observed values. This is an ideal situation.

In general, however, we will encounter Case III: there are an infinite number of solutions for \mathbf{y}_m . In this last case, it is still possible that some elements of \mathbf{y}_m have the same values in all possible solutions. These elements can be deductively imputed.

The general solution for \mathbf{y}_m to system (3) is given by (see, e.g., Rao, 1973, or Harville, 1997):

$$\tilde{\mathbf{y}}_m = \mathbf{R}_m^- \mathbf{a} + (\mathbf{R}_m^- \mathbf{R}_m - \mathbf{I}) \mathbf{z} = \mathbf{d} + \mathbf{Cz}, \quad (4)$$

where \mathbf{R}_m^- is a so-called generalised inverse of \mathbf{R}_m (i.e., a $p_m \times q$ matrix such that $\mathbf{R}_m \mathbf{R}_m^- \mathbf{R}_m = \mathbf{R}_m$), \mathbf{I} is the $p_m \times p_m$ identity matrix, and \mathbf{z} is an arbitrary vector of length p_m . Because \mathbf{z} can be chosen arbitrarily, expression (4) generates an infinite number of solutions for \mathbf{y}_m , except in the event that \mathbf{C} is a matrix of zeros, which can only occur in the above-mentioned Case II. However, if the matrix $\mathbf{C} = \mathbf{R}_m^- \mathbf{R}_m - \mathbf{I}$ contains rows with only zeros, then the corresponding elements of $\tilde{\mathbf{y}}_m$ are the same for all possible solutions, i.e., for each arbitrary choice of \mathbf{z} . These elements can thus be deductively imputed with the corresponding values of $\mathbf{d} = \mathbf{R}_m^- \mathbf{a}$. A straightforward procedure for computing a generalised inverse of any matrix is given by Greville (1959).

This method is illustrated by means of an example in Section 4.1.

2.3 The use of non-negativity constraints

Another possibility to perform deductive imputation is to use the fact that many variables have to be non-negative. Suppose, for example, that for the variables in restriction (1), only the value of Total and the values of Subtotal_1 and Subtotal_2 are observed, and suppose that these observed values satisfy:

$$\text{Total} = \text{Subtotal}_1 + \text{Subtotal}_2.$$

Clearly, the sum of the missing variables (Subtotal_3, ..., Subtotal_s) must be zero in this case. If the missing variables are not allowed to be negative, then this means that they can all be deductively imputed with zero.

To find these types of solutions in general, we again consider the system of equations $\mathbf{R}_m \mathbf{y}_m = \mathbf{a}$ found in (3). Suppose that there is an element a_j of \mathbf{a} that is equal to zero. For the corresponding row of \mathbf{R}_m , denoted by $\mathbf{r}'_{m,j}$, it must then hold that $\mathbf{r}'_{m,j} \mathbf{y}_m = 0$. Now, if, for all elements of \mathbf{y}_m that have non-zero coefficients in $\mathbf{r}'_{m,j}$, it is true that

- i) these elements y_{mi} must all be non-negative,
- ii) the non-zero coefficients in $\mathbf{r}'_{m,j}$ are either all negative or all positive,

then it is deduced that these elements of \mathbf{y}_m are all equal to zero.

The deductive imputations derived in this way for the missing values \mathbf{y}_m are therefore given by:

$$\tilde{y}_{mi} = 0, \text{ if } a_j = 0 \text{ and conditions i and ii are satisfied.}$$

This method is illustrated by means of an example in Section 4.2.

3. Preparatory phase

4. Examples – not tool specific

4.1 Example: deductive imputation with equality restrictions

To illustrate the method described in Section 2.2, we consider a fictitious survey with eleven variables that should satisfy five equality restrictions:

$$\left\{ \begin{array}{l} y_1 + y_2 = y_3 \\ y_2 = y_4 \\ y_5 + y_6 + y_7 = y_8 \\ y_3 + y_8 = y_9 \\ y_9 - y_{10} = y_{11} \end{array} \right.$$

This system of equations can be written in the form (2) with $\mathbf{b} = \mathbf{0}$ and

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{bmatrix}.$$

Suppose that we want to use deductive imputation to treat as many missing values as possible in the following incomplete record (where ‘-’ indicates a missing value):

$$\begin{array}{cccccccccccc}
y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 & y_9 & y_{10} & y_{11} \\
154 & - & 166 & - & - & - & - & 25 & - & 204 & -
\end{array}$$

Making the appropriate partitions of \mathbf{R} and \mathbf{y} into observed and missing components, we compute

$$\mathbf{a} = -\mathbf{R}_o \mathbf{y}_o = - \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} 154 \\ 166 \\ 25 \\ 204 \end{bmatrix} = \begin{bmatrix} 12 \\ 0 \\ 25 \\ -191 \\ 204 \end{bmatrix}$$

and thus obtain the following system $\mathbf{R}_m \mathbf{y}_m = \mathbf{a}$:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} y_2 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_9 \\ y_{11} \end{bmatrix} = \begin{bmatrix} 12 \\ 0 \\ 25 \\ -191 \\ 204 \end{bmatrix}.$$

The following matrix \mathbf{R}_m^- satisfies $\mathbf{R}_m \mathbf{R}_m^- \mathbf{R}_m = \mathbf{R}_m$ and hence is a generalised inverse of \mathbf{R}_m :

$$\mathbf{R}_m^- = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix}.$$

Using this matrix in expression (4), we finally obtain:

$$\tilde{\mathbf{y}}_m = \mathbf{d} + \mathbf{Cz} = \begin{bmatrix} 12 \\ 12 \\ 25 \\ 0 \\ 0 \\ 191 \\ -13 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \end{bmatrix} = \begin{bmatrix} 12 \\ 12 \\ 25 + z_4 + z_5 \\ -z_4 \\ -z_5 \\ 191 \\ -13 \end{bmatrix}.$$

By inspection, it is seen that the first, second, sixth, and seventh rows of \mathbf{C} contain only zeros. This shows that we may deductively impute y_2 , y_4 , y_9 and y_{11} with the corresponding elements of \mathbf{d} . In this manner, we obtain the following partially imputed record:

$$\begin{array}{cccccccccccc}
y_1 & \tilde{y}_2 & y_3 & \tilde{y}_4 & y_5 & y_6 & y_7 & y_8 & \tilde{y}_9 & y_{10} & \tilde{y}_{11} \\
154 & 12 & 166 & 12 & - & - & - & 25 & 191 & 204 & -13
\end{array}$$

The remaining missing values in this example could not be imputed deductively. Imputations for these values have to be estimated by a non-deductive method. It should be noted that the accuracy of these estimated imputations may benefit from the fact that we have used deductive imputation, because more non-missing auxiliary values are now available.

4.2 Example: deductive imputation with equality and non-negativity restrictions

To illustrate the method described in Section 2.3, we consider the same set of restrictions as in the previous example, but with a different incomplete record:

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}
154	–	166	–	25	–	–	25	–	204	–

The only difference between this record and the record from Section 4.1 is that the value of y_5 is now also observed. In addition, all variables except y_{11} are now assumed to be non-negative.

Again partitioning \mathbf{R} and \mathbf{y} into observed and missing components, we obtain this time

$$\mathbf{a} = -\mathbf{R}_o \mathbf{y}_o = - \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} 154 \\ 166 \\ 25 \\ 25 \\ 204 \end{bmatrix} = \begin{bmatrix} 12 \\ 0 \\ 0 \\ -191 \\ 204 \end{bmatrix}$$

and hence the following system $\mathbf{R}_m \mathbf{y}_m = \mathbf{a}$:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} y_2 \\ y_4 \\ y_6 \\ y_7 \\ y_9 \\ y_{11} \end{bmatrix} = \begin{bmatrix} 12 \\ 0 \\ 0 \\ -191 \\ 204 \end{bmatrix}.$$

We note that the third row of this system states the following equation: $y_6 + y_7 = 0$. This equation has all the properties that we mentioned in Section 2.3: the right-hand-side equals zero, all coefficients have the same sign, and all variables involved have to be non-negative. Thus, we may deductively impute the values $\tilde{y}_6 = \tilde{y}_7 = 0$. The second row of the above system also represents an equation with right-hand-side equal to zero: $y_2 - y_4 = 0$. However, this equation contains both a positive and a negative coefficient, so it cannot be used to impute zeros in a deductive manner.

Since there are now two additional variables with non-missing values, we may update the partitions of \mathbf{R} and \mathbf{y} into observed and missing components. Using the method from Section 2.2 in the same way as before, we finally obtain the following, completely imputed record:

y_1	\tilde{y}_2	y_3	\tilde{y}_4	y_5	\tilde{y}_6	\tilde{y}_7	y_8	\tilde{y}_9	y_{10}	\tilde{y}_{11}
154	12	166	12	25	0	0	25	191	204	-13

5. Examples – tool specific

The R package `deducorrect`, which can be downloaded for free at <http://cran.r-project.org>, contains an implementation of the deductive imputation methods from Sections 2.2 and 2.3. To illustrate the use of `deducorrect`, we work out the two examples from Section 4 in R code.

First, we load the package:

```
> library(deducorrect)
```

Next, we create an object of type “editmatrix” containing the system of restrictions:

```
> E <- editmatrix( c("y1 + y2 == y3",
+                   "y2 == y4",
+                   "y5 + y6 + y7 == y8",
+                   "y3 + y8 == y9",
+                   "y9 - y10 == y11",
+                   "y1 >= 0", "y2 >= 0", "y3 >= 0",
+                   "y4 >= 0", "y5 >= 0", "y6 >= 0",
+                   "y7 >= 0", "y8 >= 0", "y9 >= 0",
+                   "y10 >= 0") )
```

We also have to read in the two records that we want to treat as a data frame:

```
> y <- data.frame( y1 = c(154, 154),
+                  y2 = c(NA, NA),
+                  y3 = c(166, 166),
+                  y4 = c(NA, NA),
+                  y5 = c(NA, 25),
+                  y6 = c(NA, NA),
+                  y7 = c(NA, NA),
+                  y8 = c(25, 25),
+                  y9 = c(NA, NA),
+                  y10 = c(204, 204),
+                  y11 = c(NA, NA) )
```

This produces the following data frame with two rows:

```
> y
   y1 y2  y3 y4 y5 y6 y7 y8 y9 y10 y11
1 154 NA 166 NA NA NA NA 25 NA 204  NA
2 154 NA 166 NA 25 NA NA 25 NA 204  NA
```

Deductive imputation may now be applied to these records by calling the function ‘`deduImpute`’ provided by the package:

```
> d <- deduImpute(E, y)
```

This command creates a list (named ‘d’ here) which contains the results of deductive imputation. We first check the status of each record:

```
> d$status
```



```

      status imputations
1   partial          4
2 corrected          6

```

This shows that the first record was partially imputed (with four imputations), while the second record was completely imputed (with six imputations). The imputed data itself is also stored in the list:

```

> d$corrected
      y1 y2  y3 y4 y5 y6 y7 y8  y9 y10 y11
1  154 12 166 12 NA NA NA 25 191 204 -13
2  154 12 166 12 25  0  0 25 191 204 -13

```

We refer to Van der Loo and De Jonge (2011) for more details on the `deducorrect` package.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- Greville, T. N. E. (1959), The Pseudoinverse of a Rectangular or Singular Matrix and Its Application to the Solution of Systems of Linear Equations. *SIAM Review* **1**, 38–43.
- Harville, D. A. (1997), *Matrix Algebra from a Statistician’s Perspective*. Springer-Verlag, New York.
- Pannekoek, J. (2006), Regression Imputation with Linear Equality Constraints on the Variables. Working Paper, UN/ECE Work Session on Statistical Data Editing, Bonn.
- Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, second edition. John Wiley & Sons, New York.
- Van der Loo, M. and de Jonge, E. (2011), Deductive Imputation with the `deducorrect` Package. Discussion Paper 201126, Statistics Netherlands, The Hague.

Specific section

8. Purpose of the method

Imputing missing values in microdata on logical grounds

9. Recommended use of the method

1. Deductive imputation is most effective when it is applied at the very beginning of the imputation process, after the removal of erroneous values, but before other forms of imputation have been used. In this way, other imputation methods have more non-missing auxiliary variables available, e.g., to estimate model parameters.

10. Possible disadvantages of the method

1. The method should be used, in principle, only for imputing values that can be derived with certainty from the observed values. In all other cases, it is usually better to use non-deductive methods, such as model-based imputation (see “Imputation – Model-Based Imputation”) or donor imputation (see “Imputation – Donor Imputation”).

11. Variants of the method

1. Deductive imputation by means of if-then rules specified by subject-matter specialists.
2. Automatic deductive imputation based on equality and non-negativity restrictions.

12. Input data

1. A data set containing microdata with missing values.

13. Logical preconditions

1. Missing values
 1. Allowed; in fact, the object of this method is to impute some of them.
2. Erroneous values
 1. Not allowed. Erroneous values have to be removed from the data in a previous step. They may be replaced by missing values.
3. Other quality related preconditions
 1. n/a
4. Other types of preconditions
 1. n/a

14. Tuning parameters

1. If relevant, a collection of restrictions (linear equations and – optionally – non-negativity constraints) for the microdata.

15. Recommended use of the individual variants of the method

1. Deductive imputation by means of if-then rules requires that subject-matter specialists design a collection of if-then rules beforehand.
2. Automatic deductive imputation is only possible if the data are restricted by equations and (optionally) non-negativity constraints. If such restrictions exist, then this variant is highly recommended.
3. Automatic deductive imputation based on equality and non-negativity restrictions requires software that can handle matrix computations. Not all survey-processing systems contain this type of functionality.
4. The two variants may be used in combination. In that case, it is recommended to start with automatic deductive imputation based on restrictions.

16. Output data

1. A data set containing partially imputed microdata, which is an updated version of the first input data set.

17. Properties of the output data

1. In the output data, all missing values in the input data have been imputed that could be derived on logical grounds from the observed values in the input data.
2. Typically, the output data still contain some missing values that have to be imputed by other methods.

18. Unit of input data suitable for the method

Incremental processing by record

19. User interaction - not tool specific

1. User interaction is not needed during an execution of deductive imputation.

20. Logging indicators

1. A list of (the number of) imputations per record, for future analyses.

21. Quality indicators of the output data

1. The fraction of missing values that have been imputed by the method.

22. Actual use of the method

1. ?

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Imputation – Main Module

2. Imputation – Model-Based Imputation
3. Imputation – Donor Imputation

24. Related methods described in other modules

1. n/a

25. Mathematical techniques used by the method described in this module

1. (Generalised) matrix inversion

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.4: Impute

27. Tools that implement the method described in this module

1. R package `deducorrect`

28. Process step performed by the method

Imputation, i.e., determining and filling in new values for occurrences of missing or discarded values in a data file

Administrative section

29. Module code

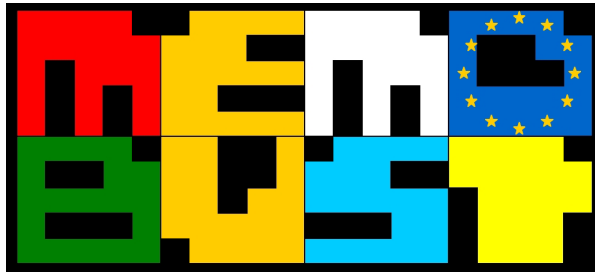
Imputation-M-Deductive Imputation

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	23-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	29-03-2011	improvements based on Norwegian review	Sander Scholtus	CBS (Netherlands)
0.2.1	04-03-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.2.2	21-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:15



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Model-Based Imputation

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction to model-based imputation.....	3
2.2 Mean imputation.....	3
2.3 Ratio imputation	4
2.4 Regression imputation	5
2.5 Practical issues	7
2.6 Multivariate methods.....	8
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

The objective in model-based imputation is to find a predictive model for each target variable in the data set that contains missing values. The model is fitted on the observed data and subsequently used to generate imputations for the missing values. Several commonly-used imputation methods are special cases of model-based imputation; this includes mean imputation, ratio imputation, and regression imputation.

2. General description¹

2.1 Introduction to model-based imputation

The objective in model-based imputation is to find a predictive model for each target variable in the data set that contains missing values. The model is fitted on the observed data and subsequently used to generate imputations for the missing values. Many practical applications use a separate model for each variable in the data set. Some multivariate extensions will be briefly discussed in Section 2.6. Before that, we will discuss *mean imputation* (Section 2.2), *ratio imputation* (Section 2.3), and *regression imputation* (Section 2.4). Section 2.5 treats certain practical issues related to the application of these methods.

2.2 Mean imputation

In mean imputation, each missing value is replaced by the observed mean of all item respondents. That is, if y_i denotes the score of the i^{th} unit on the target variable, then each missing value is imputed by

$$\tilde{y}_i = \bar{y}_{obs} = \frac{\sum_{k \in obs} y_k}{n_{obs}}, \quad (1)$$

with obs denoting the set of n_{obs} item respondents for variable y .

Obviously, mean imputation leads to a peak in the distribution of y , because the same value is imputed for all item non-respondents. On the micro level, the quality of the imputations produced by this method is generally low. The method is potentially suitable if the intended output is limited to estimates of population means and totals. In general, mean imputation is not suitable for estimating dispersion measures such as the standard deviation, frequency distributions, or correlations between target variables, because these can all be distorted by imputing observed means. The main advantage of this method is its simplicity.

It is possible to apply mean imputation within imputation classes, i.e., groups that are more or less homogeneous with respect to the target variable. In this case, formula (1) is replaced by

$$\tilde{y}_{hi} = \bar{y}_{h:obs} = \frac{\sum_{k \in h \cap obs} y_{hk}}{n_{h:obs}},$$

¹ This section is to a large extent based on Chapters 3, 4, and 5 of Israëls et al. (2011).

where y_{hi} is the score of the i^{th} unit in imputation class h and $n_{h:obs}$ is the number of item respondents for variable y in h . This extension is sometimes referred to as ‘group mean imputation’. In the context of business surveys, domain estimates by economic activity and size class are often part of the intended output. In that case, it is natural to define imputation classes based on these classifying variables, which are in fact known to correlate strongly with many economic target variables. Compared to using overall mean imputation, the use of group mean imputation should significantly improve the quality of the domain estimates and, usually, also the population estimates.

In general, group mean imputation produces a set of smaller peaks in the distribution of y (one for each imputation class). If the imputation classes are very effective in discriminating among the units, so that the variation of y between classes is much larger than the variation within classes, then this method can also be used to reasonably estimate dispersion measures. This is true because only the variation of y within classes is disregarded under this method.

2.3 Ratio imputation

For ratio imputation, we assume that there is a single auxiliary variable x that is always observed (or previously imputed) and that is more or less proportional to the target variable y . First, the unknown ratio between y and x , say R , is estimated from the units with both y and x observed:

$$\hat{R} = \sum_{k \in obs} y_k / \sum_{k \in obs} x_k .$$

Subsequently, the missing y_i are imputed by applying this ratio to the observed x_i :

$$\tilde{y}_i = \hat{R} x_i = \frac{\sum_{k \in obs} y_k}{\sum_{k \in obs} x_k} x_i . \quad (2)$$

Thus, the imputed values are obtained by assuming that the proportion that was estimated from the respondents holds exactly for the item non-respondents.

As an illustration, suppose that y denotes *turnover* and x denotes *number of employees*. Then the ratio R represents the average turnover per employee. According to (2), multiplying the observed *number of employees* for the i^{th} unit by the estimated average turnover per employee yields an estimate of *turnover* for the i^{th} unit, and this estimate is used as an imputation.

A common application of ratio imputation occurs in repeated surveys, where the value of y measured at an earlier time (say $t-1$, with t denoting the current time) is used as auxiliary information. In this case, we can write $y = y^t$ and $x = y^{t-1}$. The imputation is then given by

$$\tilde{y}_i^t = \hat{R} y_i^{t-1} ,$$

with \hat{R} the estimated development of the target variable between $t-1$ and t . We refer to the module “Imputation – Imputation for Longitudinal Data” for more details on imputation in this context.

As with mean imputation, ratio imputation can also be applied within imputation classes. In this case, a separate ratio R_h is estimated for each imputation class and used in formula (2). This may be called

‘group ratio imputation’. In general, this extension is useful if the relationship between x and y differs strongly, or at least significantly, between the imputation classes. It should be noted that ratios of groups are usually more homogeneous than group means. Regarding domain estimates in business surveys, the same remarks apply here as for group mean imputation.

2.4 Regression imputation

Regression imputation generalises mean and ratio imputation by assuming a regression model for the prediction of y given a set of auxiliary variables x_1, \dots, x_q . In many cases, a standard linear regression model is used:

$$y = \alpha + \beta_1 x_1 + \dots + \beta_q x_q + \varepsilon, \quad (3)$$

with $\alpha, \beta_1, \dots, \beta_q$ unknown parameters and ε a disturbance term, where it is assumed that the disturbances for all units are drawn independently from the same normal distribution with mean 0 and variance σ^2 .

The parameters in model (3) are estimated – usually through ordinary least squares – from the records for which both y and the auxiliary variables are observed. This results in a prediction for y given the auxiliary variables:

$$\hat{y} = a + b_1 x_1 + \dots + b_q x_q, \quad (4)$$

with a, b_1, \dots, b_q denoting the least squares estimates of $\alpha, \beta_1, \dots, \beta_q$. Assuming that the auxiliary variables are always observed, this predicted value can be computed for both item respondents and item non-respondents on y .

There are now two generic ways to obtain an imputation \tilde{y}_i from the regression model: without a disturbance term or with a disturbance term. In the first case, the predicted value from (4) is substituted directly for the missing value:

$$\tilde{y}_i = \hat{y}_i = a + b_1 x_{1i} + \dots + b_q x_{qi}. \quad (5a)$$

This results in a deterministic imputation. In the second case, we add a disturbance to the predicted value, i.e., we impute:

$$\tilde{y}_i = \hat{y}_i + e_i = a + b_1 x_{1i} + \dots + b_q x_{qi} + e_i. \quad (5b)$$

The disturbance e_i can be a random draw from the normal distribution with mean 0 and variance σ^2 , to be in line with the posited regression model (3). (Actually, σ^2 is unknown in practice and is often estimated by the residual error of the fitted model.) Alternatively, a donor can be selected from the item respondents (either at random or according to some deterministic criterion; see the module “Imputation – Donor Imputation”) and the residual of the donor with respect to the model prediction, say $e_d = y_d - \hat{y}_d$, can be substituted for e_i . In both cases, the disturbance is obtained using the regression model. Adding a disturbance results in a stochastic imputation, unless one uses a donor that is selected in a deterministic way. We refer to “Imputation – Main Module” for a discussion of the

differences between imputing with and without a disturbance term and between deterministic and stochastic imputation.

It should be noted that mean imputation can be seen as a special case of regression imputation, namely in the absence of auxiliary variables. In this case, model (3) reduces to

$$y = \alpha + \varepsilon ,$$

and the least squares estimate α is just the observed mean \bar{y}_{obs} , so that formula (5a) is identical to (1). Similarly, ratio imputation can be seen as a special case of regression imputation with one auxiliary variable and with the constant term fixed to 0. In this case, model (3) reduces to

$$y = \beta x + \varepsilon .$$

Under the alternative assumption that the variance of the disturbances equals $\sigma^2 x$ rather than σ^2 , the weighted least squares estimate for β is just the observed ratio \hat{R} , and formula (5a) is identical to (2). Note that there also exist stochastic versions of mean and ratio imputation; these are obtained by taking formula (5b) instead of (5a) in the above special cases.

In practice, the standard linear regression model may not always be appropriate. More generally, a non-linear regression model could be used, i.e.,

$$y = f(\beta_1 x_1 + \dots + \beta_q x_q)$$

for some non-linear function $f(\cdot)$. The disturbance term ε can be added to this model, or it can be implicitly contained therein.

In the case of a binary target variable with scores 0 and 1, a logistic regression model is often used:

$$\log \frac{p}{1-p} = \alpha + \beta_1 x_1 + \dots + \beta_q x_q + \varepsilon ,$$

where p denotes the probability that y takes the score of 1, given the auxiliary variables. As before, the data of the item respondents can be used to estimate the model parameters (e.g., using maximum likelihood). Next, for each unit with y_i missing, the probability that $y_i = 1$ is estimated according to

$$\hat{p}_i = \frac{\exp(a + b_1 x_{1i} + \dots + b_q x_{qi})}{1 + \exp(a + b_1 x_{1i} + \dots + b_q x_{qi})} \in (0,1) .$$

Having estimated these probabilities, imputed values may be obtained either by directly imputing $\tilde{y}_i = \hat{p}_i$ (this yields a deterministic imputation) or by randomly drawing $\tilde{y}_i = 1$ with probability \hat{p}_i and $\tilde{y}_i = 0$ with probability $1 - \hat{p}_i$ (this yields a stochastic imputation).

Note that if we impute $\tilde{y}_i = \hat{p}_i$ in the above case, the individual imputations are not valid scores (i.e., they are not equal to 0 or 1). More generally, regression imputation can produce imputations outside the domain of values that are theoretically possible for the target variable. For instance, an imputed number of employees may be non-integer, an imputed turnover may be negative, etc. Typically, this is not a problem for the estimation of population means, totals and many other statistics, but it may be problematic in applications where the microdata themselves are part of the output. If valid individual imputations are desired, then it may be better to turn to donor imputation (see “Imputation – Donor

Imputation”). See also the module “Imputation – Imputation under Edit Constraints” for the more general problem of imposing (multivariate) restrictions on the imputed values.

2.5 *Practical issues*

The regression model (3) is defined for a quantitative target variable and quantitative auxiliary variables. Categorical auxiliary variables, such as *NACE code* or *size class*, can be included in this model by defining appropriate dummy variables. In particular, group mean imputation is obtained as a special case of regression imputation by including only a dummy variable for each imputation class. For categorical target variables, other models should be used, such as (a multinomial extension of) logistic regression.

It is important to assess the quality of imputations. A direct comparison between the imputed values and the actual values is usually impossible, since the actual values are unknown. In some cases, it may be possible to obtain an impression of the quality of imputation through external validation, by comparing the imputed data to data from another source, either for the individual imputed values or at an aggregate level. Usually, however, there are conceptual differences between the various sources (different variable definitions, different target populations, etc.) so that opportunities for these types of validation are limited.

An indirect measure of the quality of a model-based imputation is provided by various indicators of model fit. For linear regression analysis with the least squares estimator, the fraction of explained variance R^2 can be used to quantify the strength of the model among the item respondents. In this way, different imputation models can be compared with one another; note that gains in R^2 for larger models should be set off against increases in degrees of freedom. For more general models, the likelihood can be used as an indicator, or a measure derived from the likelihood such as AIC or BIC. See Draper and Smith (1998) – or any other introductory book on regression analysis – for a more comprehensive discussion of model selection and ways to assess model fit. A limitation of using the model fit to assess imputation quality is that, in theory, it is possible for model *A* to have a better fit than model *B* among the item respondents, while model *B* provides better predictions than model *A* among the item non-respondents.

Another possibility to obtain an impression of the quality of different imputation methods in a particular context is to perform a simulation experiment with either the actual data set or historical data. In such an experiment, observed values are temporarily suppressed and new values are imputed for these left-out values. To the extent that the imputed values are similar to or – for categorical variables – even equal to the original values, an imputation method appears to be useful for a particular application. By defining a suitable distance function between the imputed and observed values – or, often more aptly, between target estimates based on these values –, it is possible to compare different imputation methods/models and choose the most appropriate one. This can be seen as an application of cross-validation. We refer to Schulte Nordholt (1998) and Pannekoek and De Waal (2005) for examples of such experiments. A good introduction into the design and use of simulation studies is given by Haziza (2006).

2.6 Multivariate methods

In the previous subsections, we have treated model-based imputation methods that impute a data set on a variable-by-variable basis. There also exist model-based methods that take a multivariate approach to imputation. Although these multivariate methods are more complex to use, they do have some theoretical advantages (De Waal et al., 2011, pp. 277-279). If y is imputed by a single-variable method, then typically the relationships between y and all other variables in the data set will be distorted *except* for those variables that were included as auxiliary variables in the imputation model for y . Thus, if the intended output includes correlations between target variables or other statistics of a multivariate nature, it is important to take this into account in the choice of the imputation model. Multivariate imputation methods provide a natural way to preserve correlations between target variables. Another advantage of multivariate methods is that there exist techniques that estimate a multivariate model by making use of all the available observed data (see below). As discussed above, for single-variable methods, the model has to be fitted using only the units with all predictors and the target variable observed.

2.6.1 Multivariate regression imputation

Using matrix-vector notation, a straightforward extension of the standard linear regression model (3) to the case of multiple target variables is given by:

$$\mathbf{y} = \boldsymbol{\mu}_y + \mathbf{B}_{y,x}(\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\varepsilon}, \quad (6)$$

where, for simplicity, we make the assumption that each target variable in \mathbf{y} is modeled using the same vector of auxiliary variables \mathbf{x} . In the absence of missing data, the matrix of regression coefficients $\mathbf{B}_{y,x}$ could be estimated from the data using least squares:

$$\hat{\mathbf{B}}_{y,x} = \mathbf{S}_{y,x} \mathbf{S}_{x,x}^{-1},$$

with $\mathbf{S}_{y,x}$ the matrix of observed covariances between the target variables and the auxiliary variables, and $\mathbf{S}_{x,x}$ the observed covariance matrix of the auxiliary variables. In addition, $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ could be estimated by their observed means: $\hat{\boldsymbol{\mu}}_y = \bar{\mathbf{y}}$ and $\hat{\boldsymbol{\mu}}_x = \bar{\mathbf{x}}$.

In the presence of missing data, the above estimates cannot be computed, but one could base analogous estimates only on those units for which all relevant variables are observed. However, this approach has two important drawbacks. Firstly, in particular for larger models, the number of fully observed units may be very small and the resulting estimates may be unreliable. Secondly, and perhaps more importantly, the fully observed units may form a selective subset of all units. As a result, using the fitted model to impute the item non-respondents may produce a bias in the statistical output.

A more satisfactory solution may be provided by maximum likelihood estimation with incomplete data. Under certain assumptions on the mechanism that causes the missing values, the so-called *Expectation-Maximisation (EM) algorithm* provides valid estimates of the parameters in model (6). This approach uses all the available information in the observed data to estimate these parameters, including the units with partially observed records. The interested reader is referred to De Waal et al. (2011, Ch. 8) for a brief introduction and Little and Rubin (2002) for more details.

Having obtained estimates of the unknown parameters in model (6), imputations for the missing values in a record \mathbf{y}_i may be obtained as before from the observed vector \mathbf{x}_i . That is, a deterministic imputation is obtained directly from the predicted value,

$$\tilde{\mathbf{y}}_i = \hat{\mathbf{y}}_i = \hat{\boldsymbol{\mu}}_y + \hat{\mathbf{B}}_{y,x}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x),$$

and a stochastic imputation is obtained by adding a random disturbance to this prediction:

$$\tilde{\mathbf{y}}_i = \hat{\mathbf{y}}_i + \mathbf{e}_i = \hat{\boldsymbol{\mu}}_y + \hat{\mathbf{B}}_{y,x}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x) + \mathbf{e}_i.$$

A common choice is to draw \mathbf{e}_i from a multivariate normal distribution with mean vector zero and the covariance matrix of the residuals of the regression of \mathbf{y} on \mathbf{x} (cf. De Waal et al., 2011).

2.6.2 Sequential regression imputation

In practice, applying multivariate model-based imputation as described in the previous subsection can be complicated, particularly if the data set contains a large number of variables of different types (continuous, semi-continuous, binary, etc.). It is difficult, if not impossible, to find an explicit joint model that is appropriate for such data. Van Buuren et al. (1999) and Raghunathan et al. (2001) proposed a different method, known as *sequential regression imputation* or *multivariate imputation by chained equations*. Under this approach, one models the distribution of each target variable separately, conditional on the values of the other variables. This yields a set of single-variable regression models, which have to be estimated in an iterative manner. To do this, the following procedure can be used:

1. Initialise the procedure by imputing each missing value in the original data set by a simple method (e.g., mean imputation).
2. For each variable in turn:
 - a. Estimate the parameters of the conditional regression model using all records in the current data set for which this variable was originally observed.
 - b. Use the estimated conditional model to impute the originally missing values for this variable. This updates the current data set for the next iteration.
3. Repeat Step 2 until ‘convergence’.

Note that in Step 2a, the conditional regression model is estimated using the most recent imputed version of each independent variable. In Step 3, ‘convergence’ may be assessed in terms of stability across iterations of the estimated regression parameters or the imputed values. The imputations from the final iteration are to be used in subsequent processing.

As noted above, the main practical advantage of the sequential regression approach lies in the flexibility provided by the use of separate, conditional regression models. It should be noted that this approach is theoretically justified only if the conditional models imply a proper joint model for the data. (The conditional models have to be ‘compatible’.) Otherwise, the iterative estimation procedure will not converge to a stable solution. Although this assumption usually cannot be verified beforehand, experiences so far suggest that it does not pose a problem in most practical applications (Tempelman, 2007).

Sequential regression is often applied in the context of multiple imputation. (A short discussion of multiple imputation is provided in “Imputation – Main Module”.) In fact, it is straightforward to repeat the above iterative procedure to generate multiple imputed data sets. Note that stochastic imputation should be used to make this procedure meaningful.

A good practical introduction into the sequential regression approach to imputation is provided by Azur et al. (2011). Applications in the context of business survey data are described by Tempelman (2007) and Drechsler (2009).

3. Design issues

4. Available software tools

Mean and ratio imputation can be implemented using almost any statistical software. Regression imputation with common types of models (e.g., linear regression, logistic regression) is provided as a standard feature in tools such as SPSS, SAS, and Stata. It is also straightforward to implement in R. Specialised packages are available for sequential regression imputation, such as IVEware (in SAS), and `mice` and `mi` (in R).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011), Multiple Imputation by Chained Equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research* **20**, 40–49.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, 3rd edition. John Wiley & Sons, New York.
- Drechsler, J. (2009), Far from Normal – Multiple Imputation of Missing Values in a German Establishment Survey. Working Paper, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.
- Haziza, D. (2006), Simulation Studies in the Presence of Nonresponse and Imputation. *The Imputation Bulletin* **6**, 7–19.
- Israëls, A., Kuijvenhoven, L., van der Laan, J., Pannekoek, J., and Schulte Nordholt, E. (2011), *Imputation*. Methods Series Theme, Statistics Netherlands, The Hague.

- Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, second edition. John Wiley & Sons, New York.
- Pannekoek, J. and de Waal, T. (2005), Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* **21**, 257–286.
- Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* **27**, 85–95.
- Schulte Nordholt, E. (1998), Imputation: methods, simulation experiments and practical examples. *International Statistical Review* **66**, 157–180.
- Tempelman, D. C. G. (2007), *Imputation of Restricted Data*. PhD Thesis, University of Groningen.
- Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999), Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine* **18**, 681–694.

Interconnections with other modules

8. Related themes described in other modules

1. Imputation – Main Module
2. Imputation – Donor Imputation
3. Imputation – Imputation for Longitudinal Data
4. Imputation – Imputation under Edit Constraints

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

1. Least squares estimation
2. Maximum likelihood estimation
3. EM algorithm

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.4: Impute

12. Tools explicitly referred to in this module

1. SPSS
2. SAS
3. Stata
4. R

13. Process steps explicitly referred to in this module

1. Imputation, i.e., determining and filling in new values for occurrences of missing or discarded values in a data file

Administrative section

14. Module code

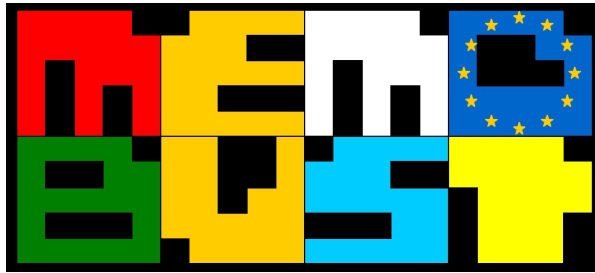
Imputation-T-Model-Based Imputation

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.1.1	27-03-2013	minor changes	Sander Scholtus	CBS (Netherlands)
0.2	11-07-2013	improvements based on Norwegian and Swedish reviews	Sander Scholtus	CBS (Netherlands)
0.2.1	21-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:16



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Donor Imputation

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction to donor imputation.....	3
2.2 Random and sequential hot deck imputation.....	4
2.3 Nearest-neighbour imputation	4
2.4 Predictive mean matching	6
2.5 Practical issues	6
3. Design issues	7
4. Available software tools.....	7
5. Decision tree of methods	7
6. Glossary.....	7
7. References	7
Interconnections with other modules.....	9
Administrative section.....	10

General section

1. Summary

The objective in donor imputation is to fill in the missing values for a given unit by copying observed values of another unit, the donor. Typically, the donor is chosen in such a way that it resembles the imputed unit as much as possible on one or more background characteristics. The rationale behind this is that if the two units match (exactly or approximately) on a number of relevant auxiliary variables, it is likely that their scores on the target variable will also be similar.

2. General description¹

2.1 Introduction to donor imputation

The objective in donor imputation is to fill in the missing values for a given unit (the *recipient*) by copying the corresponding observed values of another unit (the *donor*). The term *hot deck* donor imputation applies when the donor comes from the same data set as the recipient. In the context of business statistics, this is the most commonly encountered form of donor imputation. If the donor is taken from another data set, this is known as *cold deck* donor imputation. Most applications of cold deck imputation use data that were collected at a previous point in time. Often, the donor record is then simply an earlier observation of the recipient unit itself. This type of donor imputation is only valid for variables that can be considered more or less constant between observation times; its applicability in the context of business statistics is therefore limited. In the remainder of this module, we shall focus on hot deck imputation.

Letting y_i denote the score of the i^{th} unit on the target variable y and using the index d for a donor, we can write the generic formula for hot deck donor imputation as:

$$\tilde{y}_i = y_d. \quad (1)$$

Typically, one searches for a donor that resembles the recipient as much as possible on one or more auxiliary variables. There exist different ways to select a donor, leading to different variants of hot deck imputation. In this module, we shall describe *random* and *sequential hot deck imputation* (Section 2.2), *nearest-neighbour imputation* (Section 2.3), and *predictive mean matching* (Section 2.4). Some practical issues are discussed in Section 2.5.

In formula (1) and in the description below, we focus on imputing one target variable at a time. In practice, one often encounters records with several missing values. In that case, the standard approach is to impute all missing values in a record from the same donor. This helps to preserve the multivariate relations between the imputed variables. In fact, an important practical advantage of donor imputation compared to model-based imputation is that it can be extended to multivariate imputation in this natural way.

¹ This section is to a large extent based on Chapter 6 of Israëls et al. (2011).

2.2 *Random and sequential hot deck imputation*

In random hot deck imputation, imputation classes are formed based on categorical auxiliary variables. For each recipient unit i in a given imputation class, the group of potential donors consists of the units within the same class with y observed. Of these potential donors, one is selected at random – typically through equal-probability sampling – and used to impute the recipient. Note that this procedure implies that the donor and the recipient have exactly the same values on all auxiliary variables that are used to define the imputation classes. Conditional on these auxiliary variables, the donor is selected completely at random.

Sequential hot deck imputation also requires that the donor and the recipient have identical values on the auxiliary variables, but here the data set is not explicitly split into groups. Instead, one goes over the records in the data set in order and imputes each missing value by the last previously encountered observed value for a unit with the same scores on the auxiliary variables. Thus, the recipient is imputed using as a donor the last unit with y observed that belongs to the same imputation class and that comes before the recipient in the data file. Historically, the sequential hot deck method had the advantage that it can be carried out by a computer in a very efficient manner. The algorithm requires just one pass over the data set (Kalton and Kasprzyk, 1986). With the rise of computing power, this is no longer considered a real advantage for most practical applications.

For the sequential hot deck method, the imputations obviously depend on the order of the records in the data set. The method can be applied after a random sorting of the records; this yields stochastic imputations and is sometimes called ‘random sequential hot deck’. Alternatively, deterministic imputations may be obtained by sorting the records on one or more background characteristics. Either way, it is recommended to perform some form of explicit sorting before applying this method, because otherwise the results may be biased due to an implicit and unforeseen ordering of the units in the file.

Typically, the standard errors of means and totals of y will be inflated by random (sequential) hot deck imputation (Little and Rubin, 2002). In part, this may be due to the risk of outliers being ‘magnified’, which can be avoided by excluding outliers from the group of potential donors. More generally, it is desirable to avoid that the same unit can be used as a donor for many different recipients. In random hot deck imputation, this can be achieved by using a more elaborate selection mechanism, so that a repeated use of the same donor is only allowed once all or most of the potential donors within an imputation class have had a turn. In sequential hot deck imputation, a repeated use of the same donor may occur whenever there are several item non-respondents close together in the data file. One way to prevent this is to consider an extension of sequential hot deck imputation. Under this extension, one stores the last K observed values within an imputation class (for some $K > 1$). Whenever an item non-respondent is encountered, it is imputed by choosing at random one of the K potential donor values.

2.3 *Nearest-neighbour imputation*

In nearest-neighbour imputation, we drop the restriction that the donor and the recipient have identical scores on all auxiliary variables. Instead, the auxiliary variables are used to define a distance function $D(i, k)$ between units i and k , where i is the recipient and k is a potential donor. The *nearest neighbour* of unit i is defined as the respondent d that minimises this distance function. Formally,

$$d = \arg \min_{k \in obs} D(i, k), \quad (2)$$

where *obs* denotes the set of units with *y* observed, i.e., the set of potential donors.

Before going into the imputation method itself, we will briefly discuss possible choices of the distance function in formula (2). Assuming for now that the auxiliary variables (x_1, \dots, x_q) are all quantitative (but see Section 2.5), a frequently used family of distance functions is given by:

$$D_z(i, k) = \left(\sum_{j=1}^q |x_{ji} - x_{jk}|^z \right)^{1/z} \quad (3)$$

with $z > 0$. For $z = 2$, formula (3) yields the well-known Euclidean distance. For $z = 1$, it is just the sum of the absolute differences $|x_{ji} - x_{jk}|$; this is sometimes called the ‘city-block’ or ‘Manhattan’ distance. As z becomes larger, formula (3) places a higher penalty on large differences for individual auxiliary variables. In fact, by letting z tend to infinity in (3), we obtain the so-called ‘minimax’ distance given by

$$D_\infty(i, k) = \max_{j=1, \dots, q} |x_{ji} - x_{jk}|. \quad (4)$$

According to distance (4), the nearest neighbour should not deviate strongly from the recipient on any auxiliary variable x_j . Practical applications of nearest-neighbour imputation that involve distance function (3) with choices other than $z = 1$, $z = 2$, or $z \rightarrow \infty$ are rare.

A generalisation of (3) is obtained by including weight factors γ_j that express the importance of each auxiliary variable for the purpose of finding accurate imputations:

$$D_{z,\gamma}(i, k) = \left(\sum_{j=1}^q \gamma_j |x_{ji} - x_{jk}|^z \right)^{1/z}. \quad (5)$$

In addition, note that the contributions of the auxiliary variables to (3) or (5) are implicitly weighted if these variables are measured on different scales. For instance, if x_1 represents last year’s turnover in Euros and x_2 represents the number of employees, then the value of $D_1(i, k) = |x_{1i} - x_{1k}| + |x_{2i} - x_{2k}|$ will depend almost exclusively on the first term in practice. To prevent this, one should first standardise the auxiliary variables so that their variances are equal to 1. Alternatively, the so-called Mahalanobis distance could be used which also takes correlations between variables into account (see, e.g., Little and Rubin, 2002); this can be seen as a generalisation of the Euclidean distance $D_2(i, k)$.

In its basic form, the nearest-neighbour method imputes an item non-respondent by using its nearest neighbour as donor. This yields a deterministic imputation. As before, the underlying idea is that two units that are closely matched on relevant background characteristics [i.e., for which $D(i, k)$ has a small value] are likely to also have a similar score on the target variable.

A stochastic generalisation of nearest-neighbour imputation first selects the K units that are closest to unit i in terms of $D(i, k)$ – i.e., the K nearest neighbours – as potential donors and then draws one of these units at random. In some applications, unequal drawing probabilities are assigned to the K nearest neighbours so that within this group the units with smaller values of $D(i, k)$ are more likely to

be selected as donor. Following Bankier et al. (2000), an appropriate choice of drawing probability for the k^{th} potential donor is then given by:

$$p(k) \propto \left(\frac{D_{\min}}{D(i, k)} \right)^t, \quad (k = 1, \dots, K), \quad (6)$$

where $D_{\min} = \min_{k \in \text{obs}} D(i, k)$ denotes the distance of the nearest neighbour and $t \geq 0$ is a parameter determining the selection mechanism. Equal-probability selection is obtained as a special case of (6) with $t = 0$. The method coincides with ordinary deterministic nearest-neighbour imputation in the limit $t \rightarrow \infty$.

2.4 Predictive mean matching

Little (1988) described a variant of donor imputation known as predictive mean matching. In this imputation method, a linear regression is first performed of the target variable y on some auxiliary variables x_1, \dots, x_q . The regression model is fitted on the data of units without item non-response. Next, the resulting regression equation is used to obtain predicted values \hat{y} for all records, in accordance with formula (4) in the module “Imputation – Model-Based Imputation”. For item non-respondent i with predicted value \hat{y}_i , we select as donor the item respondent d for which the predicted value \hat{y}_d is as close as possible to \hat{y}_i . Finally, the *observed* value y_d of the donor is imputed, in accordance with formula (1) above. The latter feature makes this method a form of donor imputation rather than model-based imputation.

It should be noted that predictive mean matching is actually a special case of nearest-neighbour imputation. This is easily seen by considering the distance function

$$D_{\text{pmm}}(i, k) = |\hat{y}_i - \hat{y}_k|$$

and choosing the donor according to formula (2). Alternatively, this distance function can be expressed as a weighted sum of differences between the auxiliary variables used in the regression (De Waal et al., 2011, p. 253).

2.5 Practical issues

Random and sequential hot deck imputation require that the auxiliary variables are categorical, because these variables are used to construct imputation classes. Quantitative auxiliary variables can be included by first deriving ‘categorised’ versions of them (e.g., a size class variable based on the number of employees).

Nearest-neighbour imputation is used mainly with quantitative auxiliary variables. It is also possible to include categorical auxiliary variables, but this requires an appropriate extension of the distance function. One way to do this is to assign, for each categorical variable separately, a distance to each possible pair of values. For an auxiliary variable x_j with m categories, this ‘local’ distance function can be summarised in the form of an $m \times m$ matrix A_j . Next, we can define a ‘global’ distance function of the form (3) or (5), by replacing the absolute difference $|x_{ji} - x_{jk}|$ by the value

$A_j(x_{ji}, x_{jk})$ in these expressions. Similarly, a combination of quantitative and qualitative auxiliary variables can also be handled in nearest-neighbour imputation.

An alternative way to handle a combination of quantitative and qualitative auxiliary variables is to combine the random and nearest-neighbour hot deck methods. That is, we first use the categorical variables to construct imputation classes. Next, within each imputation class, we apply the nearest-neighbour method using a distance function of quantitative variables. In this case, the donor has to match the recipient exactly on the categorical variables but their scores on the quantitative variables may be different. The approach in the previous paragraph offers more flexibility.

It is possible to take sampling weights into account in the selection of the donor; see Kalton (1983) and Andridge and Little (2009). As discussed in “Imputation – Main Module”, there is no consensus of opinion on the necessity in general of incorporating sampling weights into imputation procedures. However, it is often useful to ensure that recipients are imputed from donors with similarly-sized weights. Effectively, donor imputation increases the weight of a donor by adding the weights of its recipients (Kalton, 1983). Therefore, if a donor with a small weight is used to impute a recipient with a much larger weight, the influence of that donor on the survey estimates increases disproportionately; as a result, the variances of these estimates will be inflated. To prevent this, the weighting variable – or the design variables that constitute the weighting model – may be included as auxiliary variables in the donor selection. Andridge and Little (2009) compared the performance of hot deck imputation with and without the inclusion of sampling weights in a simulation study.

3. Design issues

4. Available software tools

Several R packages are available that can perform hot deck donor imputation, including `StatMatch` and `mice`. The Banff system by Statistics Canada performs nearest-neighbour imputation for quantitative data. CANCEIS, another tool by Statistics Canada, offers more advanced nearest-neighbour imputation functionality for quantitative and qualitative data. It should be noted that CANCEIS is mainly aimed at social statistics, in particular the population census.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Andridge, R. R. and Little, R. J. (2009), The Use of Sampling Weights in Hot Deck Imputation. *Journal of Official Statistics* **25**, 21–36.

- Bankier, M., Lachance, M., and Poirier, P. (2000), 2001 Canadian Census Minimum Change Donor Imputation Methodology. Working Paper, UN/ECE Work Session on Statistical Data Editing, Cardiff.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- Israëls, A., Kuijvenhoven, L., van der Laan, J., Pannekoek, J., and Schulte Nordholt, E. (2011), *Imputation*. Methods Series Theme, Statistics Netherlands, The Hague.
- Kalton, G. (1983), *Compensating for Missing Survey Data*. Survey Research Center Institute for Social Research, The University of Michigan.
- Kalton, G. and Kasprzyk, D. (1986), The Treatment of Missing Survey Data. *Survey Methodology* **12**, 1–16.
- Little, R. J. A. (1988), Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics* **6**, 287–296.
- Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, second edition. John Wiley & Sons, New York.

Interconnections with other modules

8. Related themes described in other modules

1. Imputation – Main Module
2. Imputation – Model-Based Imputation

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.4: Impute

12. Tools explicitly referred to in this module

1. Banff
2. CANCEIS
3. R

13. Process steps explicitly referred to in this module

1. Imputation, i.e., determining and filling in new values for occurrences of missing or discarded values in a data file

Administrative section

14. Module code

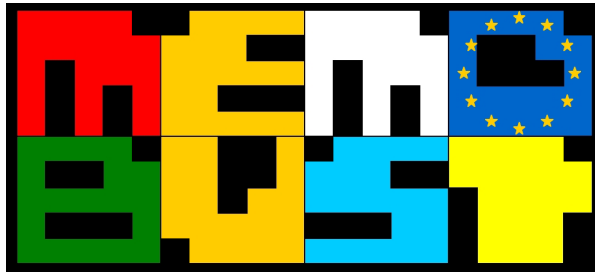
Imputation-T-Donor Imputation

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.2	15-07-2013	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	07-10-2013	improvements based on Norwegian review	Sander Scholtus	CBS (Netherlands)
0.3.1	21-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:16



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Imputation for Longitudinal Data

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Longitudinal data.....	3
2.2 Introduction to imputation for longitudinal data	3
2.3 Imputation methods	5
2.4 Evaluation techniques.....	9
2.5 Quality indicators of the output data	9
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

We refer to longitudinal data when the same variables of the same units are measured several times at different moments. The common trait is that the entity under investigation is observed or measured at more than one point in time, possibly regularly, in order to study how it develops over time. The data are collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each unit from historical records. Also data from registers can be referred to as longitudinal data, indeed it is possible to match historical data about the same units once they are available with some degree of regularity.

This theme is due to describe the methods for imputation of missing longitudinal data, that could be performed for all aforementioned types of data. Particular emphasis is focused on the Short Term Statistics context.

2. General description

2.1 Longitudinal data

Longitudinal data are typically the result of a repeated survey, whose purpose is to collect data on the same observation units along several years (e.g., every four years or biannual) or once a year (annually) or several times during the same year (e.g., quarterly or even monthly). In the context of business statistics, longitudinal data can be used both in structural and in short term surveys. The combination of the periodicity and the type of parameter to be estimated can determine the difference between Structural Business Statistics (SBS) and Short Term Statistics (STS) (see the modules “General Observations – Different Types of Surveys” and “Repeated Surveys – Repeated Surveys”). In a short-term statistics context the parameter to be estimated is usually the change of a certain indicator along time.

In general, longitudinal data can be represented as data collected on the same units several times in a consecutive sequence, hence for each unit $i=1,\dots,n$ belonging to the sample, there are $t=1,\dots,T$ different measurements, one for each wave of interview. The period t can be a month, a quarter or a year; the first two cases drive to intra-annual longitudinal data. It is clear that, given the period t , a vector of cross-sectional observations is available, while as regards the i -th observation a vector of longitudinal data on the same unit is available and a strong correlation is expected among these values (see the module “Statistical Data Editing – Editing for Longitudinal Data”).

2.2 Introduction to imputation for longitudinal data

In statistical surveys, respondents sometimes do not provide answers to one or more questions, while they are required to do that. Commonly, two cases are distinguished: the *item non-response* (or *partial non-response*) is when the unit answers to the survey, but it does not provide information about one or more questions; the *unit non-response* case is when the observation unit does not respond at all. In a longitudinal context, these cases can vary also with respect to the specific time t the data are related to, hence, the missing values come into two forms:

- a) scattered missing values: item or total non-response, because units do not answer to some questions or to the total questionnaire in one or more waves, but they deliver the whole records in other waves. Most of the times the high timeliness of the STS increases late answers with respect to the deadline, so that their data are available afterwards;
- b) panel drop-out: starting from a specific time t some units stop to answer. This phenomenon is called panel *attrition* (Kalton, 2009).

In the case of longitudinal data, the unit dropout is often the greatest concern, because it could hide a major reason for not answering and it should be considered to systematically behave in a different way compared to the units which give response to the survey, even if not at every wave. In these cases, it is suggested to investigate the event, to discover whether the unit has been modified by a demographic event (see the module “Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys”) that could change the composition of the panel.

Where imputation of missing values is required, there are two possible approaches according to the dimension. On one side, for each occasion t a set of cross-sectional data is available, for which all the described methods are applicable. On the other side, for each unit i a longitudinal vector is available, for which also other methods can be applied that would take into account the information from other measurements on the same units.

There are two main reasons to use longitudinal imputation techniques instead of the cross-sectional methods:

1. Earlier or later observations of the same object are generally very good predictors for the missing value. This means that the quality of the imputation can strongly be improved.
2. To correctly estimate changes of a variable over time (typically the final aim of a short-term survey), the imputation of missing values should take into account information about the previous and the future values of the given variable on the same unit under observation, that supplies useful evidence about their change over time.

It must be observed that the use of cross-sectional methods is unavoidable in case of missing or incorrect information referred to units included for the first time in a rotating panel, as no historical data are available for these units.

Imputation of missing values can be derived from other characteristics of the unit under study (see the module “Imputation – Deductive Imputation”), when also values recorded in other occasions are available the same rule can be applied. In other cases, auxiliary information is available and it makes prediction model of the missing values possible, which is supposed to generate the data (see the module “Imputation – Model-Based Imputation”). These models can be applied also in the case of a longitudinal context, once the proper auxiliary variable has been settled to be the measurement of the same variable on the given unit in another occasion. The choice of the imputation method usually depends on the characteristics of the variable under observation. In the longitudinal context the different pattern of seasonality should also be taken into account, as it determines important features of the variable (for instance, the number of monthly hours worked depends on the number of working days in the same month).

Many methods are based on the assumption that data are originated from a multivariate normal distribution. These methods should be applied carefully to data coming from business surveys,

because the above mentioned hypothesis is not valid in case of concentration of enterprises. In particular cases such as for very big enterprises, it is worth identifying a specific imputation method which takes into account the profile of the units themselves in order to improve quality of final estimates.

This is the reason why an a priori analysis of each variable under study is recommended, in order to choose the proper kind of historical data to be used for the imputation as auxiliary information

2.3 Imputation methods

Imputation methods considerably depend on the type of data set, its extent and the characteristics of the missingness mechanism. Those for longitudinal data usually take into account the historical information of each unit to define any type of imputation method (both for the deductive imputation and as auxiliary information). Let y_{it} be a missing value of unit i at period t on variable y . Then y -values of unit i at previous and subsequent periods can be used to create an imputed value \tilde{y}_{it} . The longitudinal imputation methods are briefly described in the following sections.

2.3.1 Last observation carried forward

In this case, the last observed value of a unit is used for the values of the later periods that must be imputed, that is called Last Observation Carried Forward (LOCF). It is often used in practice, even though it may have some problems (Israëls et al., 2011; Watson and Starick, 2011).

This method is mainly applicable to categorical variables, for which it is known that their change is very little over time. For the quantitative variables, it risks to produce an overly stable picture of the actual situation.

2.3.2 Interpolation or historical imputation

In this case, missing observations can be estimated from both previous and later observations; obviously, in the case of current surveys data can be imputed only using previous observations. Different versions of the method include correction based on a trend component (Israëls et al., 2011).

In the case data exhibit a specific periodical pattern, it is recommended to use data from the same period (in short-term statistics the historical data of one season ago, i.e., one year ago, one month ago).

For the unit i , \tilde{y}_{it} is determined by a function of the K observations from the past and L observations from the future. Interpolation can be used for quantitative variables in a situation where it is difficult to make any model assumption on the variable under study, because there is neither correlation with previous measurement of the same variable nor with other variables in the same context. For quantitative variables, the following rather general formula is suggested:

$$\tilde{y}_{it} = \frac{\sum_{k=1}^K w_{-k} y_{it-k} + \sum_{l=1}^L w_l y_{it+l}}{\sum_{k=1}^K w_{-k} + \sum_{l=1}^L w_l} \quad (1)$$

with weights $w_{-1} \geq w_{-2} \geq \dots \geq w_{-K}$ and $w_1 \geq w_2 \geq \dots \geq w_L$; this means that y_{it} has a smaller weight in both directions from period t , as periods k and l are further away from period t . The weights can be freely

selected, for example, it is possible to choose $K=L$ and $w_k=w_{-k}=1/k$. When only information from the past is used or in the case of panel drop-out, the weights w_1, \dots, w_L are all equal to zero.

If an intra-annual value has to be estimated, the interpolation formulas can be adjusted in order to take into account the seasonal pattern.

The general formula (1) can be applied in several cases, one example is the linear interpolation between the preceding and the subsequent observation of the same unit, for which the equality $w_1=w_{-1}$ is usually considered:

$$\tilde{y}_{it} = \frac{w_1(y_{it-1} + y_{it+1})}{2w_1} = \frac{y_{it-1} + y_{it+1}}{2} \quad (2)$$

A proposal to determine the weights w_{-1} and w_1 is based on the observed changes on the respondent units of the sample: an indicator variable is created which equals to 1 when the reported change between waves t and $t-1$ is smaller than the reported change between waves t and $t+1$ for the complete cases and 0 otherwise. Then, it is possible to calculate the proportion p , which is the share of the interviewed sample for which the change between waves t and $t+1$ is smaller than the change between the previous wave t and $t-1$. Hence, the weight $w_1=p$ reflects the probability to change between t and $t+1$, while $w_{-1}=1-p$ is about the change between $t-1$ and t , both reflecting the probabilities associated with the occurrence of change between waves found in the complete cases (Watson and Starick, 2011).

2.3.3 Mean imputation

A missing value is replaced by the mean of valid data. It can be applied both in the longitudinal and cross-sectional view. According to the first one it can be seen as a specific case of the interpolation, where the weights simply represent the presence of each data. The cross-sectional approach is very useful when longitudinal data are not available and the assumption of similar behaviour between respondents and not respondents is valid.

Let y_{it} the response for subject i at occasion t , let y_{it-k} and y_{it+l} be the response of the same unit at time $t-k$ and time $t+l$, and r_{it-k} and r_{it+l} equal to 1 if y_{it-k} and y_{it+l} are observed, 0 otherwise. If y_{it} is missing, it can be replaced by the mean of the nearest preceding and subsequent observations as follows:

$${}_L\tilde{y}_{it} = \frac{\sum_{k=1}^K r_{it-k} y_{it-k} + \sum_{l=1}^L r_{it+l} y_{it+l}}{\sum_{k=1}^K r_{it-k} + \sum_{l=1}^L r_{it+l}} \quad (3)$$

where the time t can vary both along previous observations or future observations. In this case, each missing unit will be replaced by a different value that is strictly correlated to its longitudinal profile. On the other side, the cross-sectional mean response for unit i at time t is equal to:

$${}_{CS}\tilde{y}_{it} = \frac{\sum_j r_{jt} y_{jt}}{\sum_{j \in obs} r_{jt}} \quad (4)$$

where y_{jt} is the observed value of the j -th respondent at time t and obs is the sample of respondent observations. In this case a cross-imputation is done and the same mean is imputed for each missing value; in this term, it can lead to a peak in the distribution. An alternative version of this method is to

impute a class mean, where the classes may be based on some explanatory variables. This method is influenced by the existence of patterns and similarities between enterprises and, therefore, it has to be carefully evaluated before being used. Anyway, it offers a very good tool in the case where new units have entered the panel and no longitudinal information is available for them. Disadvantages of such procedures are that distributions of survey variables are compressed and relationships between variables may be distorted (Little and Rubin, 2002).

2.3.4 Ratio imputation

Let us suppose that the variable y , to be imputed, is strongly correlated to a single auxiliary variable x and let a coefficient R represent the relationship between the variables y and x such that $y=Rx$ for every unit in the target population. For longitudinal data, the most common situation is that x measures a past observation of the same variable y , for which it is reasonable to take the assumption that the observation at period t is proportional to the observation at period $t-1$. To update the past value to the current time t the observed growth on the respondents is used, with respect to the past observed value at time $t-1$. After the pattern of the variable has been determined, it could happen that variable y is proportional to the same variable observed at the same month (or quarter) in the previous year, hence, the choice will fall on past observations referred to times $t-12$ or $t-4$ (an example is the case of the hours worked). As a consequence, a missing value can be estimated by increasing the previous observation according to the same proportion of the one observed on the respondent units from time $t-1$ to time t .

In these terms, the past value y_{it-1} can be used as the auxiliary information to impute y_{it} and the constant R is used to link the two historical values. Generally, R is not known and it is estimated at every t using only those units for which values at both occasion t and $t-1$ are known:

$$\tilde{y}_{it} = \tilde{R}_t y_{it-1} = \frac{\sum_{j \in obs} y_{jt}}{\sum_{j \in obs} y_{jt-1}} y_{it-1} \quad (5)$$

where y_{jt} is the observed value of the j -th respondent at time t and obs is the sample of respondents observations. According to the previous formula, the proportional constant is equal to the ratio between the means of y_t and y_{t-1} calculated using the units respondent in both periods¹.

2.3.5 Regression imputation

The regression of the variable of interest is based on covariates and the resulting equation is used to estimate the missing values. An advantage of longitudinal data is that, in general, the past and/or future observed values of a variable are very good predictors of missing values.

The regression imputation may use both quantitative and categorical variables, in the second case the logistic regression must be used instead of the linear regression. It is considered a good imputation method for business surveys (Kovar and Whitridge, 1995), but it should be controlled in case of new developments in the business cycle that are not included in the model.

¹ Where, for example, the variable y strongly depends on the number of working days in the reference period (nwd_t), the use of a further multiplier is recommended such as: nwd_t/nwd_{t-1} .

For a missing value y_i , a regression model is assumed for the prediction of y by means of information given by the observed value of the same variable y at previous time $t-1, t-2 \dots$. The regression model is as follows:

$$y_i = \alpha + \beta_{t-1} y_{it-1} + \dots + \beta_{t-k} y_{it-k} + \varepsilon_i \quad (6)$$

with $\alpha, \beta_{t-1}, \dots, \beta_{t-k}$ are unknown parameters, $\varepsilon_i \sim N(0, \sigma_i^2 I)$ is the unit residual which is supposed to follow a multivariate normal distribution, where I is the identity matrix and σ_i^2 is the unit model variance. In the presence of longitudinal data, we are generally interested in the correlation between the observations at different periods; therefore it is important that the imputation method retains the correlation between the observations. Where the changes over time are under study, if the disturbance term is not used, the significance of the changes will be strongly overestimated.

Model (6) can be seen as a particular case of the general regression model, where only the lagged values of the variable y are used as auxiliary variables. Regression imputation may also be applied including other auxiliary variables x correlated with the y under study in model (6) as well.

The mean imputation and the ratio imputation can be seen as special cases of the regression imputation (see the module “Imputation – Model-Based Imputation”): in the mean imputation no auxiliary variables are used; in the case of the ratio imputation the model is based also on another auxiliary variable x .

2.3.6 Donor imputation

The donor imputation methods involve replacing missing values with values from a “similar” responding unit of one or more variables for a non-respondent (called the recipient) with observed values from a respondent (the donor), that is similar to the non-respondent with respect to characteristics observed on both cases. In some versions, the donor is selected randomly from a set of potential donors, which we call the donor pool, as the *random hot deck method*. In other versions a single donor is identified and values are imputed from that case, as the *nearest neighbour method* based on some metric, where there is no randomness involved in the selection of the donor (see the module “Imputation – Donor Imputation”).

The missing variable values are replaced by the values of one of the respondents, the possibility to impute several values on the same unit, also in its longitudinal profile, makes these methods particularly suitable for longitudinal data. As a rule, one donor is chosen to ensure consistency within the same record. In nearest neighbour imputation, a distance $d(i,j)$ is defined between two objects i and j , where i is the item non-respondent and j an arbitrary item respondent. A possible measure for the similarity between a non-respondent enterprise and a possible neighbour is based on the correlation of historical data. An advantage of the method is that the results are plausible values, because the donor has been checked in advance and so not too many further controls are needed.

2.3.7 Little and Su method

The Little and Su method can be used for missing values in a quantitative variable y , which can be modelled as a combination of period effect and an individual effect and for which stochastic imputation is desired. It is a nearest neighbour technique, that takes into account both cross-sectional and longitudinal information in defining the nearest neighbours. Imputations can be based on row effects (units) and column effects (periods), where the sum of periods reproduces the whole

observation year. The residual is taken from another unit which, in terms of the row effect, is most similar to the unit that is imputed. The assumption is that units that are similar with respect to the row effect are also similar with respect to residuals. In the ideal case, the donor (of the residuals) has as many attributes equal to the recipient as possible.

This method is reasonably easy to use and can deal with different patterns of missing data, including multiple missing values per single unit. More details on the calculation method are described in the specific method module “Imputation – Little and Su Method”).

2.4 *Evaluation techniques*

An analysis of the imputed data is usually recommended, most of the proposed indicators are based on the comparison between the imputed values and the true values that the non-respondents would have supplied. In the STS context sometimes the non-response is actually a late answer, i.e., it is not in time with respect to the official deadline for the estimates, but it is available immediately after. Hence, such a comparison is possible at least on the set of late responses. On the other hand, a measurement can also be performed on data created randomly according to a simulation scheme, in this way data are not influenced by any characteristics of the late respondents, and the comparison would be done between the simulated data and the ones derived from the imputation method (Little and Su, 1989).

2.5 *Quality indicators of the output data*

The indicators are usually based on a measure of distance between the two kinds of data. They can be evaluated either at a micro level, or regarding a parameter elaborated at macro level or comparing the eventual difference between the distributions of the two final sets of data.

In general, the usual indicators are based on the following criteria:

- a. Predictive Accuracy:* to assess how the imputed value \tilde{y} (estimate) is close to the reference (true) value y^* :

a.1 the first evaluation criterion, based on the Pearson correlation between \tilde{y} and y^* , this criterion works well for data that are reasonably normal. As r gets closer to 1 the imputation method is judged to be good; if data are highly skewed this measure is not recommended as it could be influenced by the presence of outliers and influential values.

a.2 Another criterion assesses the preservation of the change between waves, by comparing the cross-wave correlations for the imputed and true values. The imputation method is better as the cross-wave correlations from the imputed data are closer to the true cross-wave correlations.

- b. Distributional accuracy:* to measure the distribution accuracy by analysing whether the imputation method preserves distribution of the true values:

b.1 the Kolmogorov-Smirnov distance is calculated between the empirical distribution for both the imputed and the true values. The imputation method is judged to be better as the distance is smaller.

b.2 It is also important to compare the distribution in the dataset that includes the imputed values with the one that includes only true values (this measure includes all cases rather than just those imputed). A measure is based on the change in the variable “decile group membership” from one wave to another. A Chi-Square test is used where the observed cell frequencies are those from the imputed

dataset and the expected cell frequencies are the true cell frequencies. The best imputation method will have the lowest χ^2 .

3. Design issues

4. Available software tools

Mean, ratio and regression imputation can be implemented using almost any statistical software. Several R packages are available that can perform imputation, for example, *StatMatch* and *Mice*.

In SAS there are IVEware (Imputation and Variance Estimation) and BANFF. The first uses a multivariate sequential regression approach for multiply imputing item missing values in a data set. The second is a generalised system for statistical editing and imputation developed at Statistics Canada.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.

EUROSTAT (2006), *Methodology of Short Term Business Statistics: Interpretation and Guidelines*. Methods and Nomenclatures.

Israëls, A., Kuijvenhoven, L., van der Laan, J., Pannekoek, J., and Schulte Nordholt, E. (2011), *Imputation*. Methods Series Theme, Statistics Netherlands, The Hague.

Kalton, G. (2009), Designs for Surveys over Time. In: *Sample Surveys: Design, Methods and Applications*, Elsevier, Amsterdam, 89–108.

Kennon, R., Copeland, K. R., and Valliant, R. (2007), Imputing for Late Reporting in the U.S. Current Employment Statistics Survey. *Journal of Official Statistics* **23**, 69–90.

Kovar, J. and Whitridge, P. (1995), Imputation of Business Survey Data. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, 403–423.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, second edition. John Wiley & Sons, New York.

Little, R. J. A. and Su, H.-L. (1989), Item Non-response in Panel Surveys. In: D. Kasprzyk, G. Duncan, and M. P. Singh (eds.), *Panel Surveys*, John Wiley and Sons, 400–425.

Watson, N. and Starick, R. (2011), Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey. *Journal of Official Statistics* **27**, 693–715.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Different Types of Surveys
2. Repeated Surveys – Repeated Surveys
3. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
4. Statistical Data Editing – Editing for Longitudinal Data
5. Imputation – Model-Based Imputation
6. Imputation – Donor Imputation

9. Methods explicitly referred to in this module

1. Imputation – Deductive Imputation
2. Imputation – Little and Su Method

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 5.4 Imputation

12. Tools explicitly referred to in this module

1. R
2. SAS

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

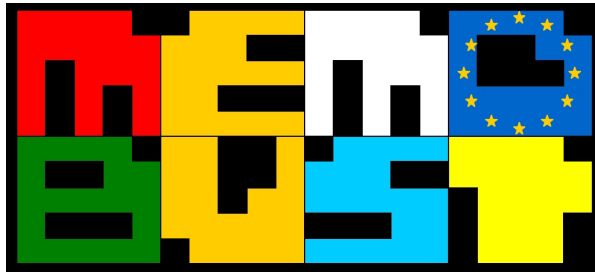
Imputation-T-Longitudinal Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-02-2013	first version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.2	23-08-2013	review based on the received comments	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.3	31-10-2013	review	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4	25-11-2013	review	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4.1	29-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:17



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Little and Su Method

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Introduction	3
2.2 Description of the Little and Su method.....	4
2.3 Conclusion.....	5
3. Preparatory phase	5
4. Examples – not tool specific.....	5
4.1 Example of the Little and Su method.....	5
5. Examples – tool specific.....	7
6. Glossary.....	7
7. References	7
Specific section.....	8
Interconnections with other modules.....	9
Administrative section.....	10

General section

1. Summary

In the following we describe the Little and Su method which is applicable to impute longitudinal data. It takes into account both trend information derived from the data and single units levels. In Section 2 some background of the method are given, while in Section 4 an example of an application of this imputation method is described.

2. General description of the method

2.1 Introduction

Text In the case of repeated measures on a single variable, relatively efficient and simple imputations can often be based on the variable classified by unit and by period (wave). In this context the Little and Su imputation method actually incorporates information about the overall trend of the data and the single unit levels of the unit under study (Little and Su, 1989). It is a nearest neighbour technique, that takes into account both cross-sectional and longitudinal information in defining the nearest neighbours. Furthermore, a residual component is taken from another unit which is most similar to the unit that is imputed in terms of the unit characteristics.

According to this method, the variable is classified by row (meant as unit level) and by column (meant as the period), on which the information about the unit and the trend, respectively, are elaborated.

The two main effects can be combined in different ways. If the missing values are well fitted by a model with additive row and column effects, then imputations may be based on an additive row + column fit:

$$\text{imputation} = (\text{row effect}) + (\text{column effect}) + (\text{residual}) \quad (1)$$

If a multiplicative model, or equivalently an additive model for the logarithm of the variable, seems more appropriate to fit missing values, then imputations may be based on a multiplicative row \times column fit:

$$\text{imputation} = (\text{row effect}) \times (\text{column effect}) \times (\text{residual}) \quad (2)$$

The choice of an additive or a multiplicative model depends on the characteristics of missing data, i.e., if data to be imputed have to be not negative, a multiplicative model has to be applied. This is the common case of data coming from business surveys: turnover, number of persons employed, wages and so on. An example can be found in Little and Su (1989).

In the Little and Su method the row and column effects are proportional to row and column means; the column effect describes the mean change over time and is therefore also called the ‘period effect’, while the row effect describes the single unit level corrected for the period effect (Frick et al., 2003).

In particular, the column effect for a certain period is based on the ratio between the period y mean and the average y mean calculated through the whole year: the higher the column effect, the higher the “seasonal” weight of the period concerned will be.

The row effect for a certain unit is given by the y mean of all the available longitudinal observations for that unit, where each period observation has been divided by its specific column (period) effect. The row effect is the “longitudinal profile” of the unit concerned.

The residual is taken from another unit which, in terms of the row effect, is the most similar to the unit of which data are going to be imputed. The assumption is that units that are similar with respect to the row effect are also similar with respect to residuals.

2.2 Description of the Little and Su method

As said before, this method (Israëls et al., 2011) can be used for missing values in a quantitative variable y , which can be modeled as a period effect combined to a single unit effect and for which imputation is desired. It is reasonably easy to use and can deal with different patterns of missing data, including multiple missing values for each unit. Some problems in applying this method can occur in cases of observed values are all equal to zero for rows with values to be imputed.

In the following an implementation of the model is described.

The column effect c_t gives the mean change of the variable y over time and is estimated by:

$$c_t = \frac{\bar{y}_t}{\frac{1}{M} \sum_{t=1}^M \bar{y}_t} \quad (3)$$

where \bar{y}_t is the mean of the observed y_{it} at period t , M is the number of periods (or waves) for which the average is considered to be significant. The row effect r_i for unit i is represented by:

$$r_i = \frac{1}{m_i} \sum_t \frac{y_{it}}{c_t} \quad (4)$$

where the sum is calculated over the m_i available y_{it} for unit i over all the periods it is observed.

The residual is derived considering all the units for which the periods, missing for unit i , are observed. All these units are sorted according to the row effect value and, among them, the one presenting a row effect closest to that of unit i , say unit j , is selected.

The residual of unit j is represented by:

$$e_{jt} = \frac{y_{jt}}{r_j c_t} \quad (5)$$

In the case of additive model (1), the final estimation is:

$$\tilde{y}_{it} = r_i + c_t + e_{jt} \quad (6)$$

on the other hand, in the case of multiplicative model (2), the final estimation is:

$$\tilde{y}_{it} = r_i c_t e_{jt} \quad (7)$$

It is important to notice that, in this case, a zero row effect will result in a zero imputed value.

In both (6) and (7) the three terms represent the row, column, and residual effects, respectively. In particular the first two terms estimate the predicted mean, and the last term is the component of the imputation from the matched case.

Considering (5), expression (7) can also be written as:

$$\tilde{y}_{it} = r_i c_t \frac{y_{jt}}{r_j c_t} = \frac{r_i}{r_j} y_{jt} \quad (8)$$

From (8) it can be derived that, if the multiplicative model (2) is applied, the final estimation is proportional to the y_{jt} value (y value for the closest unit), adjusted by the ratio between the row effects of the units i and j .

2.3 Conclusion

In general, the method has the following useful features:

- a) the imputed values incorporate information about trend from the column effects, and single unit level from the row effects;
- b) the method does not require separate modelling for different pattern of missing data, dealing with all patterns simultaneously;
- c) the method is comparatively easy to implement and this is an important consideration with large complex data sets.

3. Preparatory phase

4. Examples – not tool specific

4.1 Example of the Little and Su method

A practical example of the use of the Little and Su method in a longitudinal study can be found in this section. Suppose to have the following small sample of fictitious responses to current wages and salaries. In Table 1 there are all cases.

From this example, we see that observation 1 did not respond to the current wages and salaries questions in wave 1, but provided responses in subsequent waves. Observations 5 and 6 also partially responded and wages and salaries information are not provided in two and in one waves, respectively. The first step in the Little and Su method consists in calculating the column effects based on complete cases only, that is, units that were interviewed in 3 waves and responded in all 3 waves for the variables of interest; in the example there are 7 complete cases.

The Little and Su method incorporates trend information into the imputed amounts via the column effects. In this example, the wave 1 column effect of 0.70 indicates that the mean current wages and salaries in wave 1 is 30% lower than the overall mean current wages and salaries, and the means in waves 2 and 3 are 6% and 24% higher than the overall mean, respectively.

Table 1

OBS	Wages & salaries		
	Wave 1	Wave 2	Wave 3
1		400	420
2	675	235	700
3	345	690	800
4	200	480	210
5	200		
6	350	370	
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200

In the following, the row effects are calculated: for each unit the row effect is the mean (computed on the number of recorded cases) of the reported values divided by the correspondent column effect. In our example, the row effect for unit 1 is $((400/1.06+420/1.24)/2)$. The sample is then ordered by increasing row effects (Table 2). In this way, for each observation to be imputed, it is possible to identify the closest donor as the closest complete case.

Table 2

OBS	Wages & salaries			
	Wave 1	Wave 2	Wave 3	
10	135	130	200	159
5	200			286
4	200	480	210	303
1		400	420	358
6	350	370		425
7	400	450	470	458
8	0	790	790	461
9	360	450	600	474
2	675	235	700	584
3	345	690	800	596
	0.70	1.06	1.24	

The following step consists in imputing the missing value by multiplying the actual value for the variable of interest of the donor with the row effect of the recipient divided by the row effect of the donor. That is:

- Obs1 - Wave 1: $200 \cdot 358 / 303 = 236.30 \sim 236$
- Obs5 - Wave 2: $480 \cdot 286 / 303 = 453.07 \sim 453$
- Obs5 - Wave 3: $210 \cdot 286 / 303 = 198.22 \sim 198$
- Obs6 - Wave 3: $470 \cdot 425 / 458 = 436.14 \sim 436$

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Frick, J. R. and Grabka, M. M. (2003), *Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income Distribution*. DIW Berlin.
- Israëls, A., Kuijvenhoven, L., van der Laan, J., Pannekoek, J., and Schulte Nordholt, E. (2011), *Imputation*. Methods Series Theme, Statistics Netherlands, The Hague.
- Little, R. J. A. and Su, H.-L. (1989), Item Non-response in Panel Surveys. In: D. Kasprzyk, G. Duncan, and M. P. Singh (eds.), *Panel Surveys*, John Wiley and Sons, 400–425.

Specific section

8. Purpose of the method

Check erroneous values in microdata on logical grounds.

9. Recommended use of the method

- 1.

10. Possible disadvantages of the method

- 1.

11. Variants of the method

- 1.

12. Input data

- 1.

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 1. Not allowed. All observed values have to be correct.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

- 1.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

- 1.

17. Properties of the output data

- 1.

18. Unit of input data suitable for the method

19. User interaction - not tool specific

1.

20. Logging indicators

1.

21. Quality indicators of the output data

1.

22. Actual use of the method

1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Imputation – Imputation for Longitudinal Data

24. Related methods described in other modules

1.

25. Mathematical techniques used by the method described in this module

1.

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1.

28. Process step performed by the method

Administrative section

29. Module code

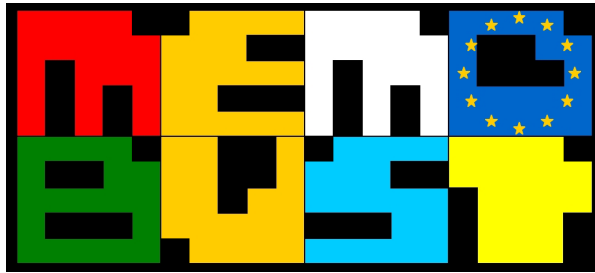
Imputation-M-Little and Su

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-02-2013	first version	Roberto Gismondi Fabiana Rocci Anna Rita Giorgi Maria Liria Ferraro	Istat (Italy)
0.2	20-08-2013	second version	Roberto Gismondi Fabiana Rocci Anna Rita Giorgi Maria Liria Ferraro	Istat (Italy)
0.3	30-10-2013	review	Roberto Gismondi Fabiana Rocci Anna Rita Giorgi Maria Liria Ferraro	Istat (Italy)
0.3.1	19-11-2013	preliminary release		
0.3.2	29-11-2013	minor corrections		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:17



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Imputation under Edit Constraints

Contents

General section	3
1. Summary	3
2. General description.....	3
2.1 Introduction	3
2.2 Imputation under edit constraints by direct modeling	4
2.3 Imputation under edit constraints by adjustment methods	7
3. Design issues	8
4. Available software tools.....	8
5. Decision tree of methods	8
6. Glossary.....	8
7. References	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

In the context of business surveys at National Statistical Institutes (NSIs), imputation of missing values is often complicated by the fact that the data should conform to a large number of edit rules. In this module, we consider two basic approaches to obtain imputations that satisfy edit rules. Under the first approach, the edits are incorporated directly in the imputation model, so that all imputations are automatically consistent. Unfortunately, this can lead to a very complex model. Therefore, in practice, another approach is often used, in which the missing values are first imputed without taking the edits into account. In a subsequent step, the initial imputations are then minimally adjusted to become consistent with the edits.

2. General description

2.1 Introduction

In the context of business surveys at NSIs, the imputation of missing values is often complicated by the fact that the data should conform to a large number of restrictions, known as *edit rules*, *edit constraints*, or *edits* (see also “Statistical Data Editing – Main Module”). For instance, if a survey includes the variables *turnover*, *costs*, and *profit*, then the edit rule

$$profit = turnover - costs$$

is supposed to hold for the corresponding values. In addition, there are edits stating that the values of *turnover* and *costs* should be non-negative. It is desirable to avoid imputations that are inconsistent with the edit rules, because data with obvious inconsistencies are likely to be rejected by most users, even if they could in fact be used to make valid statistical inferences (Pannekoek and De Waal, 2005). Särndal and Lundström (2005, p. 176) wrote: “Whatever the imputation method used, the completed data set should be subjected to the usual checks for internal consistency. All imputed values should undergo the editing checks normally carried out for the survey.”

Obviously, if a standard imputation method such as regression imputation (see “Imputation – Model-Based Imputation”) or random hot deck imputation (see “Imputation – Donor Imputation”) is applied without taking the edit rules into account, then one should generally not expect the resulting imputations to satisfy the edits. Unfortunately, taking edit rules into account directly in the imputations tends to introduce complications. De Waal et al. (2011) give the following simple example. Suppose that we are given a record with missing values on the variables x and y , and suppose that the following edit rules have been defined for these variables:

$$x \geq 50; \tag{1}$$

$$y \leq 100; \tag{2}$$

$$y \geq x. \tag{3}$$

If we first impute x , the only edit which can be evaluated at this stage is (1). Taking this edit into account, we might impute the value $\tilde{x} = 150$. The resulting edit rules for y given by (2) and (3) cannot be satisfied simultaneously: $y \leq 100$ and $y \geq 150$. Furthermore, if we start by imputing y ,

taking edit (2) into account, we might impute the value $\tilde{y} = 40$ and encounter a similar problem with the resulting edit rules for x . Thus, consistency with the edit rules is not guaranteed under this sequential procedure. The point is that if the variables are imputed sequentially, in general, edit rules involving variables that will be imputed later cannot be ignored.

There are two general approaches to imputation under edit constraints. The first approach is to, somehow, include the edit rules in the (implicit or explicit) model used for imputation, so that the imputed values automatically satisfy all constraints. The second approach is to apply a two-step procedure. In the first step, the missing values are imputed without taking (all) constraints into account. In the second step, the initially imputed values are minimally adjusted to satisfy all edits. These two approaches will be discussed further in Sections 2.2 and 2.3, respectively. Finally, it should be noted that values derived by deductive imputation methods (see “Imputation – Deductive Imputation”) trivially satisfy the edits that were used in the derivation. We will return to this point in Section 2.3.

2.2 *Imputation under edit constraints by direct modeling*

2.2.1 *Ratio hot deck imputation*

In general, imputation methods that take edit constraints into account directly tend to be complex. One exception is the ratio hot deck method. This is an extension of the ordinary hot deck donor imputation method (see “Imputation – Donor Imputation”) that is appropriate to impute missing values among a set of non-negative variables y_1, \dots, y_m that should satisfy a linear balance edit of the form:

$$y_1 + \dots + y_m = y_{tot}, \quad (4)$$

where it is assumed that the total value y_{tot} is always observed (or previously imputed). Basically, instead of imputing the donor values directly, we use the donor to distribute the total missing amount over the missing variables.

Consider the i^{th} record that requires imputation and suppose for notational convenience that the first t variables are observed (with values $y_{i,1}, \dots, y_{i,t}$) and the last $m-t$ values are missing. We first compute the total missing amount, $r_i = y_{i,tot} - y_{i,1} - \dots - y_{i,t}$. Next, using any of the ordinary donor imputation methods, we choose a donor from the completely observed records. The donor record should be consistent with the edits. We compute the sum of the donor values of the variables to impute, say, $r_d = y_{d,t+1} + \dots + y_{d,m}$. The ratio hot deck imputations are given by:

$$\tilde{y}_{i,j} = \frac{r_i}{r_d} y_{d,j}, \quad (j = t+1, \dots, m).$$

By construction, the imputed values are non-negative and consistent with edit (4):

$$y_{i,1} + \dots + y_{i,t} + \tilde{y}_{i,t+1} + \dots + \tilde{y}_{i,m} = y_{i,1} + \dots + y_{i,t} + \frac{r_i}{r_d} r_d = y_{i,tot}.$$

For an application of the ratio hot deck method in practice, see Pannekoek and Van Veller (2004) or Pannekoek and De Waal (2005). A straightforward generalisation of the method can be applied if the

balance edit contains coefficients unequal to 1 (De Waal et al., 2011). Unfortunately, the method cannot be used to obtain consistent imputations if there are multiple, inter-related restrictions.

2.2.2 Parametric imputation models

To introduce the direct modeling of edit constraints in a parametric model, it is useful to consider a small univariate example. Suppose that a certain variable y is to be imputed using the normal distribution $N(\mu, \sigma^2)$, and suppose in addition that we require the imputations to be non-negative; i.e., the edit rule $y \geq 0$ should hold. To make the example interesting, consider the case that μ and σ are such that the distribution $N(\mu, \sigma^2)$ has a significant probability of generating negative values (e.g., $\mu = 1$ and $\sigma = 2$). The edit would be failed quite often if we imputed values directly from $N(\mu, \sigma^2)$. An intuitively sensible approach to obtain consistent imputations in this case works as follows: obtain a random draw z from $N(\mu, \sigma^2)$. If it holds that $z \geq 0$, then impute $\tilde{y} = z$. Otherwise, repeat the procedure until a draw with $z \geq 0$ is obtained. By construction, all resulting imputations will satisfy the non-negativity edit. Technically, these imputations follow a so-called *truncated* normal distribution (Geweke, 1991).¹ The above iterative procedure for obtaining values from this distribution is known as *Acceptance/Rejection sampling* (Tempelman, 2007).

The univariate truncated normal distribution is a relatively simple example of a model that incorporates constraints on the modeled variables (in this case: one inequality constraint and one variable). The general idea of imputation under edit constraints by direct modeling is to find a model that incorporates all the relevant constraints on the variables to impute. The main advantage of this approach is that it avoids having to adjust the imputations later on to satisfy the edit rules. Two important disadvantages of the direct modeling approach are: (i) in most practical applications, the resulting imputation methods are mathematically complex and require heavy computational work; and (ii) as this methodology is relatively new, only a limited number of models have been developed.

Tempelman (2007) developed imputation models that can incorporate particular types of constraints:

- If all edits are linear inequalities (i.e., the restrictions can be written as $\mathbf{Qy} \geq \mathbf{b}$ for a given matrix \mathbf{Q} and vector \mathbf{b} of constants), then the multivariate truncated normal distribution can be used. The distribution is truncated to the region defined by the constraints $\mathbf{Qy} \geq \mathbf{b}$. This is a multivariate extension of the univariate example given above.
- If all edits are linear equalities (i.e., the restrictions can be written as $\mathbf{Ry} = \mathbf{a}$ for a given matrix \mathbf{R} and vector \mathbf{a} of constants), then the multivariate singular normal distribution can be used. This is a generalisation of the ordinary multivariate normal distribution $N_d(\mu, \Sigma)$ for the

¹ In general, a random variable with density function $f(x|\theta)$ can be truncated to any subdomain G of its original support by defining the truncated density function:

$$f(x|\theta; G) = \begin{cases} \frac{f(x|\theta)}{\int_G f(x|\theta)dx} & \text{if } x \in G \\ 0 & \text{if } x \notin G \end{cases}$$

case that the covariance matrix Σ is singular (Khatri, 1968). In fact, the covariance matrix is singular here because the constraints $\mathbf{Ry} = \mathbf{a}$ induce a linear dependence in this matrix.

- If both linear equalities and inequalities occur, then the multivariate truncated singular normal distribution can be used. This distribution combines the features of the two previous cases.
- For the special case of one linear equality with non-negativity edits for all variables involved – i.e., the case that can be handled by the ratio hot deck method –, an alternative model is given by the Dirichlet distribution (Wilks, 1962).

A full treatment of these models is beyond the scope of this module. We refer to Tempelman (2007) and De Waal et al. (2011, Ch. 9) for more details. An important theoretical limitation of the first three models is that they are only appropriate for data that are approximately normally distributed. Moreover, it is not useful here to apply a standard (non-linear) transformation to the data to obtain a closer resemblance to a normal distribution, because the edits for the transformed data would not have the linear structure ($\mathbf{Qy} \geq \mathbf{b}$ and/or $\mathbf{Ry} = \mathbf{a}$) of the original edits.

2.2.3 The elimination approach

In the above approaches, a joint model is used to impute all variables with missing values in a record at once. A somewhat different, less complex approach was proposed by Coutinho et al. (2007). They used a technique called *Fourier-Motzkin elimination* (Williams, 1986; De Waal et al., 2011) to reduce the problem of consistent imputation to a sequence of univariate problems. This elimination technique is used more traditionally in algorithms for automatic error localisation. We refer to the module “Statistical Data Editing – Automatic Editing” for a brief description of Fourier-Motzkin elimination.

A full discussion of the elimination approach is beyond the scope of this module. Here, we will only give a small example. Consider again the example from Section 2.1, where the objective is to impute the variables x and y in such a way that the edits (1), (2), and (3) are satisfied. Before we can start imputing, we have to posit and estimate a joint model for the data. In contrast to Section 2.2.2, this model need not incorporate the edit constraints, which makes the modeling task much easier. Following Coutinho et al. (2007), we will use an ordinary bivariate normal distribution in this example for simplicity:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 60 \\ 55 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix} \right). \quad (5)$$

We begin by applying Fourier-Motzkin elimination to the original edits (1)–(3) to eliminate x from these edits. In this particular example, this yields two implied edits for the remaining variable y :

$$y \geq 50; \quad (6)$$

$$y \leq 100. \quad (7)$$

We would now like to impute y from its posited $N(55,100)$ distribution, in such a way that the imputed value satisfies the inequalities (6) and (7). This can be achieved, as in the example from Section 2.2.2, by drawing from a truncated normal distribution by means of Acceptance/Rejection

sampling. That is, we draw random values from the $N(55,100)$ distribution until we obtain a value that lies between 50 and 100. Suppose that we obtain the value $\tilde{y} = 70$.

In the next step, we substitute the imputed value $\tilde{y} = 70$ for y in the original edits (1)–(3). This yields two reduced edit rules that involve only x :

$$x \geq 50 ; \quad (8)$$

$$70 \geq x . \quad (9)$$

Finally, x is imputed by drawing from the posited $N(60,100)$ distribution until we obtain a value that complies with edits (8) and (9). (In general, we would use the conditional distribution of x , given the previously imputed value for y , but the two variables are uncorrelated in this example.) This might yield the value $\tilde{x} = 52$. In this manner, we obtain the imputed record $(\tilde{x}, \tilde{y}) = (52, 70)$ which is consistent with the original edits (1)–(3).

By a fundamental property of the Fourier-Motzkin elimination technique, the above method always yields imputations that are consistent with the edit rules (Coutinho et al., 2007). Note that according to model (5), the mean of x is larger than the mean of y . In this sense, the posited model does not comply with edit rule (3). Nevertheless, the elimination approach yields consistent imputations, as was illustrated by the example. However, it should be noted that if the model strongly disagrees with the edit rules, the procedure of Acceptance/Rejection sampling from a truncated distribution may become very inefficient. In fact, an appropriate model for the data should not strongly disagree with the edit rules, provided that these rules are substantively meaningful.

For a general description of the elimination approach to consistent imputation, we refer to Coutinho et al. (2007) and De Waal et al. (2011, Ch. 9). Extensions of this method have been considered by Pannekoek et al. (2008, 2013) and Coutinho et al. (2013).

2.3 *Imputation under edit constraints by adjustment methods*

Since most of the methods discussed in Section 2.2 have limited practical applicability, a less complex approach is often applied in practice. Under this approach, the variables with missing values are first imputed by any method that produces a complete data set with good statistical properties, without taking (all) edit constraints into account. That is to say, any appropriate method discussed in the other modules on imputation can be used. Denote the initial imputed record by \hat{y} . Next, an adjusted imputed record \tilde{y} is obtained from \hat{y} as the solution to a constrained minimisation problem:

$$\begin{aligned} &\text{Minimise } D(\hat{y}, \tilde{y}) , \\ &\text{so that } \tilde{y} \text{ satisfies all edit constraints.} \end{aligned} \quad (10)$$

Here, D is a function that measures the distance between the initial imputed record \hat{y} and the adjusted record \tilde{y} . It is customary to demand that only the imputed values may be adjusted under this minimisation problem, i.e., the variables that were originally observed retain their original values.

Adjusting the imputed values for consistency with the edit constraints is a special case of the general problem of *data reconciliation*. Methods for this more general problem are treated in “Micro-Fusion – Reconciling Conflicting Microdata” and in particular the underlying method module “Micro-Fusion –

Minimum Adjustment Methods”. The reader is referred to these modules and to De Waal et al. (2011, Ch. 10) for more details.

In the special case that all edits are linear equalities (written as a linear system of the form $\mathbf{R}\mathbf{y} = \mathbf{a}$), one could also apply the methodology discussed in the module “Imputation – Deductive Imputation” to obtain a consistent record in the second step above. Suppose that the initial imputed record $\hat{\mathbf{y}}$ is partitioned as $\hat{\mathbf{y}} = (\hat{\mathbf{y}}'_o, \hat{\mathbf{y}}'_m)'$ and the imputed values in $\hat{\mathbf{y}}_m$ are suppressed (i.e., replaced by missing values). The matrix \mathbf{R} is partitioned accordingly as $\mathbf{R} = [\mathbf{R}_o \quad \mathbf{R}_m]$. If \mathbf{R}_m has full rank, it follows that the missing values are imputed consistently by $\tilde{\mathbf{y}}_m = \mathbf{R}_m^{-1}(\mathbf{a} - \mathbf{R}_o \hat{\mathbf{y}}_o)$; see “Imputation – Deductive Imputation” for more details. Thus, we should choose $\hat{\mathbf{y}}_m$ in such a way that \mathbf{R}_m has full rank.² Since this choice is not unique in practice, we may randomly vary the selection of $\hat{\mathbf{y}}_m$ for each imputed record; thereby, we avoid the introduction of a systematic effect in some variables. The resulting approach may be seen as a heuristic approximation to minimisation problem (10). However, if appropriate software is available, finding the optimal solution to (10) directly should be relatively straightforward and there is little to be gained from a heuristic approach.

3. Design issues

4. Available software tools

There are no generally available tools that have the imputation methods described in this module as standard functionality. Some NSIs have developed dedicated tools for particular applications. On the other hand, the methods are relatively easy to implement in statistical computing environments such as R and SAS, using the existing functionality available in these environments. Some standard tools do exist for solving problem (10) in the adjustment step of Section 2.3; e.g., the R package `rspa`, as well as commercial solvers such as CPLEX and Xpress.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Coutinho, W., de Waal, T., and Remmerswaal, M. (2007), Imputation of Numerical Data under Linear Edit Restrictions. Discussion Paper 07012, Statistics Netherlands, The Hague.

² It seems undesirable to suppress and impute values that were originally observed. To avoid this, one should restrict the system $\mathbf{R}\mathbf{y} = \mathbf{a}$ to those edits that involve at least one imputed value (the other edits should already be satisfied by the observed values). The partitioning can and should then be made in such a way that $\hat{\mathbf{y}}_m$ contains only variables that were initially imputed.

- Coutinho, W., de Waal, T., and Shlomo, N. (2013), Calibrated Hot Deck Imputation Subject to Edit Restrictions. *Journal of Official Statistics* **29**, 299–321.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- Geweke, J. (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities. Report, University of Minnesota.
- Khatri, C. G. (1968), Some Results for the Singular Normal Multivariate Regression Models. *Sankhyā Series A* **30**, 267–280.
- Pannekoek, J. and de Waal, T. (2005), Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* **21**, 257–286.
- Pannekoek, J. and van Veller, M. G. P. (2004), Regression and Hot-Deck Imputation Strategies for Continuous and Semi-Continuous Variables. In: J. R. H. Charlton (ed.), *Methods and Experimental Results from the EUREDIT Project*. (<http://www.cs.york.ac.uk/euredit/>)
- Pannekoek, J., Shlomo, N., and de Waal, T. (2008), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.
- Pannekoek, J., Shlomo, N., and de Waal, T. (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. *Annals of Applied Statistics* **7**, 1983–2006.
- Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester.
- Tempelman, D. C. G. (2007), *Imputation of Restricted Data*. PhD Thesis, University of Groningen.
- Wilks, S. S. (1962), *Mathematical Statistics*. John Wiley & Sons, New York.
- Williams, H. P. (1986), Fourier's Method of Linear Programming and Its Dual. *The American Mathematical Monthly* **93**, 681–695.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Imputation – Main Module
3. Imputation – Model-Based Imputation
4. Imputation – Donor Imputation

9. Methods explicitly referred to in this module

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Micro-Fusion – Minimum Adjustment Methods
3. Statistical Data Editing – Automatic Editing
4. Imputation – Deductive Imputation

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.4: Impute

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Imputation, i.e., determining and filling in new values for occurrences of missing or discarded values in a data file

Administrative section

14. Module code

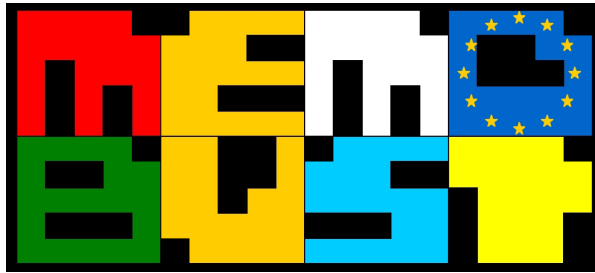
Imputation-T-Imputation under Edit Constraints

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.2	10-07-2013	improvements based on Norwegian review	Sander Scholtus	CBS (Netherlands)
0.2.1	19-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:17



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Weighting and Estimation – Main Module

Contents

General section.....	3
1. Summary	3
2. General description.....	4
2.1 Weighting – basic weighting	4
2.2 Weight adjustment for non-response.....	5
2.3 Weight adjustment.....	6
2.4 Robust estimation in the presence of outliers	7
2.5 Model-based estimators.....	8
2.6 Panel surveys.....	9
2.7 Preliminary estimates	11
2.8 Small area estimation	12
2.9 Integration of administrative sources in the statistical production	13
2.10 Departure from ideal conditions: imperfect frames.....	14
3. Design issues	15
4. Available software tools.....	16
5. Decision tree of methods	17
6. Glossary.....	17
7. References	17
Interconnections with other modules.....	20
Administrative section.....	22

General section

1. Summary

The present module gives an overview of the methods that can be used to obtain estimates for parameters such as totals, means or ratios, from the observed sample data. It is assumed that data have already been processed to treat potential errors and item non-response (see the modules “Statistical Data Editing – Main Module” and “Imputation – Main Module” for introduction to treatment of errors and item non-response).

Commonly, in official statistics, probability-based sampling designs are carried out, and a design weight can be associated to each sampled unit. This design weight equals the inverse of the inclusion probability. It can be thought as the number of population units each sample unit is representative of. Hence, a simple method to obtain estimates of the target parameters is to use these design weights to inflate the sample observations (see subsection 2.1). Design weights are strictly related to sampling design implemented for the survey (see the module “Sample Selection – Main Module”). Moreover, design weights can be adjusted also to consider non-response (see subsection 2.2), and/or they can be modified to take into account of auxiliary information (Särndal et al., 1992). An example of use of external information is given by the calibration estimator (see the module “Weighting and Estimation – Calibration”) or the GREG estimator (see the module “Weighting and Estimation – Generalised Regression Estimator”), which is a special case of calibration estimator.

The previous estimators are unbiased or approximately unbiased in a randomisation approach (or design-based approach: properties are assessed on the set of all possible samples). Note that even if, in some cases a model is assumed (as for GREG), the properties of the estimators do not depend on the model and the estimators remain design unbiased even in case of model failure. For this reason, this class of methods is robust. However, their efficiency depends strongly on model assumptions and relationships on auxiliary variables affect their variances.

In fact, when the distribution of the target variable in the population is highly skewed, as it often happens in business surveys, representative outliers may occur in the sample. The values of such units are true values and then they do not need to be edited (see the topic “Statistical Data Editing”). Nevertheless, even if estimators remain unbiased, presence of these outlying units has a large impact on variance estimators. The module “Weighting and Estimation – Outlier Treatment” gives an overview of methods that have been suggested in literature for reducing variance of the estimates, while controlling for the presence of bias.

A relevant approach for estimation is given by model-based approach: differently from design-based approach, where, as stated above, properties are assessed on the set of all possible samples, in this framework, the assumption of a model is the basis to obtain estimators that are the best in terms of model Mean Square Error: Best Linear Unbiased Predictor (Royall, 1970, Vaillant et al., 2000). In official statistics, the class of model-based estimator is applied in specific situations, such as when the sample size is not large enough to obtain estimates with sufficient accuracy (small area estimators, see also the module “Weighting and Estimation – Small Area Estimation”). A second important field of application of model-based estimation is given by preliminary estimation, when for short term statistics a provisional estimate is calculated on a sub-sample of the sample units. The auto-selection of units in the preliminary sample may be the most relevant issue for preliminary estimates. Moreover,

when the sample is selected with a non-probabilistic mechanism, model-based estimates can be applied for inference, and model-based variance can be evaluated.

The peculiarity of panel surveys is also highlighted. In panel surveys, the same units are observed in several occasions (waves), allowing for reduction of estimators' variance and estimation of longitudinal parameters (e.g., gross change and measure of frequency). Cross-sectional and longitudinal weights have to be determined according to the target parameters (see subsection 2.6).

Finally, the use of administrative data is mentioned in subsection 2.9.

To conclude the review of relevant issues in weighting and estimation, subsection 2.10 underlines some of the most typical matters in applied cases.

2. General description

2.1 Weighting – basic weighting

A very important methodology in sampling strategy is provided by the use of weights to obtain estimates of the parameter of interest such as totals (levels), means, differences (or ratios), *etc.* In official statistics, probabilistic sample designs are regularly implemented and a design weight equal to the inverse of the inclusion probability can be associated to each sample unit.

The design weight can be thought as the number of units in the population, a unit in the sample represents.

On this basis, a very simple principle to obtain estimates is to use the design weights. Estimates are produced by summing up the sample data multiplied by their design weights, i.e., the data are inflated with the weights for reproducing the whole population.

Let y_i be the value of the target variable associated to the i -th unit, and let d_i be the weight equal to the inverse of the inclusion probability, an estimation of the total of Y is given by:

$$\hat{Y}_{HT} = \sum_{i \in s} d_i y_i . \quad (1)$$

The resulting estimator is called the Horvitz-Thompson estimator.

The Horvitz-Thompson estimation of the mean is

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{i \in s} d_i y_i .$$

Whereas, in estimating the mean value, if the amount of population is estimated as well, the Hájek estimator is obtained

$$\hat{\bar{Y}}_H = \frac{1}{\hat{N}} \sum_{i \in s} d_i y_i = \frac{1}{\sum_{i \in s} d_i} \sum_{i \in s} d_i y_i . \quad (2)$$

The use of design weights is relevant in particular whenever the sample design assigns different inclusion probabilities to units in the population, e.g., to account for different size of population units if size is thought to be related with the main target variable.

In the case of stratified simple random sampling design (see the module “Sample Selection – Main Module”), for unit i belonging to stratum h :

$$d_{ih} = \frac{N_h}{n_h},$$

\hat{Y} reduces to

$$\hat{Y} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_i,$$

whereas

$$\hat{Y}_{HT} = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_i = \sum_{h=1}^H \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{i=1}^{n_h} y_i \right) = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h.$$

More complex indicators can be estimated by replacing true values by their respective HT estimators. For example, estimation of change of variable Y in a given lag-time l is given by

$$\hat{Y}_{t+l} - \hat{Y}_t. \quad (3)$$

Estimation of relative change is given by

$$\frac{\hat{Y}_{t+l} - \hat{Y}_t}{\hat{Y}_t}, \quad (4)$$

where \hat{Y}_t and \hat{Y}_{t+l} are the estimates of Y at different times t and $t+l$, obtained by applying formula (1).

2.2 Weight adjustment for non-response

The principle of weighting is also applied to account for unit non-response of sample units. In fact, design weights can be adjusted also to consider non-response in order to reduce the possible bias of resulting estimates, which may arise when there is a different propensity in answering for different groups. For example, the sample can be partitioned into sub-groups of units where the response rates are assumed to be constant, and where it can be assumed that non-respondents behave similarly to respondents. More precisely, the method is based on the assumption that the non-response depends on variables that define the sub-sets, but conditionally on these variables it is independent of the target variable (non-response is missing at random, MAR, see Little and Rubin, 2002). This grouping may differ from the sampling strata and cut across them.

A response rate, possibly weighted by the initial sampling weights, is determined in each class and a new weight is defined as the product of the design weight and the inverse of the response rate. The new weights are used in the weighting process of respondent sample units in order to get the estimates.

Let us assume for simplicity in notation that sample design is stratified and that sub-groups (or post-strata) coincide with design strata, the response rate in stratum h is evaluated as:

$$r_h = \frac{n_{rh}}{n_h}.$$

Then the initial weight of unit i in stratum, $d_{hi} = \frac{N_h}{n_h}$, is replaced with the new weight

$w_{hi} = \frac{d_{hi}}{r_{hi}} = \frac{N_h}{n_{rh}}$ and the usual HT is given by:

$$\hat{Y}_{HT} = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{n_{rh}} \sum_{i=1}^{n_{rh}} y_i = \sum_{h=1}^H \frac{N_h}{N} \left(\frac{1}{n_{rh}} \sum_{i=1}^{n_{rh}} y_i \right) = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_{rh}.$$

Occasionally, unit non-response can also be treated by imputation methods (see the module “Imputation – Main Module”).

2.3 Weight adjustment

Besides the modification of weights for handling non-response, weights adjustment may also be carried out to take into account of auxiliary information, for example by means of the calibration estimator (see the module “Weighting and Estimation – Calibration”) or GREG estimator (see the module “Weighting and Estimation – Generalised Regression Estimator”). The use of auxiliary information can have the aim to insure consistency among estimates of different sample surveys. Indeed, when good covariates are available, some improvement in the precision of estimators may be achieved by exploiting the relationship between target variable and extra information.

Auxiliary data can be used to improve the precision of the estimators as long as the values of the auxiliary variables are collected for all surveyed units and known population totals are available for these variables from another reliable source in case a linear relationship is assumed. Otherwise, totals do not suffice; see comments in the module “Weighting and Estimation – Generalised Regression Estimator”.

A general method for exploiting auxiliary information is calibration estimators (see the module “Weighting and Estimation – Calibration”). The weights are adjusted so that applying the estimators on the auxiliary variables, one is able to reproduce the known covariates totals. Calibration includes well-known estimators such as the regression, the ratio and the raking-ratio estimators (Deville and Särndal, 1992).

However, the calibration estimator may introduce high variability in weights and consequently an increase in variance of the estimator which may be relevant whenever the auxiliary variables are not enough correlated with the target variable. In particular, in official statistics, where the same set of weights is used for several target variables, it may happen that the set of covariate used in determining the final weights is not appropriate for reducing the variance of the estimators of a sub-set of the target variables.

Besides the aim of actually improving the accuracy of the sample estimators, calibration is often applied in practice to attain consistency of estimates obtained with different sources. In fact, the estimates calculated with a survey should be consistent with information on known totals obtained, for example, from a larger survey or from reliable administrative sources. Though, problems to achieve consistency can be encountered in practice if weights are also forced to lie within a given range.

An important use of calibration estimators is for further reducing the effect of unit non-response on the estimators and the possible coverage error of the sampling frame (see Lundström and Särndal, 2001).

In fact, calibration estimators may offer some protection against non-response bias when non-response is related with variables used in calibration. Poststratification and regression estimation, both special cases of calibration estimators, are widely used techniques to attempt to reduce non-response bias in sample surveys. Särndal and Lundström (2005) suggest the use of calibration to handle non-response.

Finally, weighting can be applied to combine different samples (sources) and produce estimators that are more accurate than the estimators based on any of the single samples individually (e.g., see Renssen and Nieuwenbroek, 1997, Houbiers et al., 2003).

Once the weights are obtained, estimators of totals, means are easily obtained as described above.

2.4 Robust estimation in the presence of outliers

In business surveys, the statistical distribution of target variables is often highly skewed, hence in observed sample observations that differ substantially from most of the other observations occur. These units, referred as *representative outliers* (see Chambers, 1986), are true values in the finite population and should not be considered as gross errors.

In particular, presence of this kind of outlying values in the sample does not affect the bias of the HT or calibration estimators described in 2.2 and 2.3. However their occurrence has usually a great impact on their variability.

Outlier treatment at estimation stage (robust estimation) aims at reducing the effect on variance of outliers, also controlling the possible bias of the estimator.

The methods for dealing with outliers can be broadly classified as winsorisation, modification of weight, and M estimation, i.e., methods for robust estimation in classical theory properly adapted in the finite population estimation framework.

In particular, winsorisation consists in modifying the outlying observations so that they have less impact on the estimation. Sample observations whose values lie outside certain pre-set cut-off values are set equal to the cut-off (type I winsorisation) or are transformed as a linear combination of the observed value and the cut-off (type II winsorisation) with coefficients for the observed values equal to the inverse of the sampling weights (see Gross et al., 1986, Kokic and Bell, 1994).

In case of simple random sample (s.r.s.), the winsorised estimator is

$$Y_{WR} = \frac{N}{n} \sum_{i \in s} Y_i^*$$

where

$$Y_i^* = \begin{cases} fy + (1-f)K & y > K \\ y & \text{otherwise} \end{cases}$$

and f is a coefficient in $[0,1]$. When $f = 0$, the winsor estimator is said winsor of type I, whereas, when $f = n/N$, a winsor type II estimator is obtained.

An extension for a stratified sampling design is in Gross et al. (1986). Choice of cut-off under superpopulation models are in Kokic and Bell (1994), Chambers and Kokic (1993). See also the module “Weighting and Estimation – Outlier Treatment” for a detailed description of winsorisation

estimator and the choice of cut-off for a general sampling design. Once the data are transformed the estimation process consists in applying the chosen estimator (e.g., GREG) to the new set of data.

The cut-off values are chosen to approximately minimise the MSE of the resulting estimator, usually under model assumptions (e.g., see Kokic and Bell, 1994, for optimal cut-off in stratified sampling design), the efficacy of this method is highly dependent on the goodness of cut-off(s) choice.

An alternative class of methods relates to modification of sampling weights, i.e., weights are reduced to decrease the impact of outlying units. Various methods for weight reduction have been proposed (see Hidirolou and Srinath, 1981, Lee, 1995).

For s.r.s., Hidirolou and Srinath (1981) suggested

$$\hat{Y}_{HS} = \lambda \sum_{i \in s_2} Y_i + q(\lambda) \sum_{i \in s_1} Y_i ,$$

where s_1 is the sub-sample of *inliers* and s_2 is the subsample of outliers, $q(\lambda)$ is a function of the downweighting factor λ of the outlying units, such that

$$\lambda n_2 + q(\lambda) n_1 = N .$$

Hidirolou and Srinath (1981) proposed a method to determine λ in order to minimise the conditional mean squared error, which is difficult to apply in practice. Chambers (1986) obtained an optimum value that minimises the model-based mean squared error. This method requires estimation of unknown parameters of the two different models underlying the subpopulation of inliers and the subpopulation of outliers.

Finally, the class of M estimators (Huber, 1981) is applied to HT or GREG estimators in the finite population sampling framework (e.g., see Chambers, 1986, Hulliger, 1995, Beaumont and Alavi, 2004).

See Beaumont and Rivest (1999) for a description of the methods and a presentation of practical issues. The module “Weighting and Estimation – Outlier Treatment” in this handbook provides a review of methods for dealing with outliers at estimation stage focusing on winsorisation methods.

2.5 *Model-based estimators*

The weighting methods described previously rely on inference that is based on the randomisation introduced by the sampling mechanism. This approach is more robust to model failure, i.e., less dependent on model assumptions on super-population¹ relationships between the target variable and auxiliary variables and for this reason commonly applied in official statistics. Though, model-based framework for inference in finite population sampling (see Valliant et al., 2000) is applied in specific fields of application, as it can produce more reliable estimators than those obtained with the traditional design-based (or model-assisted²) approach, and it may be preferable in cases where the sample size is very limited. We mention here some circumstances where model-based estimation is applied in official statistics.

¹ A mechanism generating the realised finite population.

² The model-assisted approach assumes a super-population relationship, as well. However, on the contrary to model-based approach, the properties of the estimators are still based on the randomisation approach. The calibration estimator described in the previous subsection is an example of model-assisted estimator.

An important field of application of model-based estimators, as we will see in subsection 2.8, is on small area estimation³. The issue of small area estimation arises whenever the sample size of a target domain is not large enough, so the direct estimator has too large variability to be published (see EUROSTAT, 2013, for some examples of threshold on reliability of the estimators). A large development in terms of methods and software, as well as real applications in official statistics has been produced in recent years (see Rao, 2003, EURAREA project, and WP2 and WP6 reports of ESSnet SAE).

Model-based estimation has also been proposed for the dissemination of short term statistics where the need for timely estimates conflicts with the need to observe the whole planned sample (Rao et al., 1986). In this case, besides the problem of estimation in presence of few observed data, one has to deal with risk of presence of (auto) selection bias. See the module “Weighting and Estimation – Preliminary Estimates with Model-Based Methods” for model-based methods to tackle the preliminary estimation issue.

Finally, note that whenever the sample is selected without a randomisation mechanism but units are chosen purposely, model-based estimation represents the framework for assessing the obtained estimators. More specifically, the implicit model that is underlying the estimation method can be evaluated in order to give support to, or, on the contrary, to invalidate the strategy used for estimation (see Kalton, 1983). For example, the ratio estimator is commonly associated with cut-off sampling, which is often chosen for convenience and cost consideration. This strategy can be justified under a ratio model. Validity of this model can be verified to assess the whole sampling strategy and measures of variability can be provided following this approach (see Valliant et al., 2000, and the module “Sample Selection – Main Module” for a review of sampling designs). See also Benedetti et al. (2010) where a model-based estimator is proposed for the unobserved subpopulation in a cut-off framework.

Models that are used at micro level to cope, for example, with non-response or to edit units (for these issues, see the topics “Statistical Data Editing” and “Imputation”) are not reported within the weighting and estimation topic.

2.6 *Panel surveys*

Short term surveys make use of repeated surveys (see the module “Repeated Surveys – Repeated Surveys”) to produce estimation of monthly or quarterly changes. For this reason, overlapping of samples, instead of renewing the sample at each occasion of the same survey, is applied as it allows reducing the variance of estimation of net changes. In fact, variations over time are measured more accurately with overlapping samples with respect to the case where samples on different occasions do not overlap (see for example Eurostat, 2013 for estimation of variance of changes when samples overlap). Actually, standard errors of the estimate of changes over time are minimised by using complete overlap of samples (Kish, 1965) if the correlation between observations in different periods is positive, as it is usually the case. Estimation of changes is a relevant objective for short term statistics and the use of panel (or rotating panels, see Kish, 1987), where the same set of units is observed each month or quarter of the year(s), is a mean to attain the aim of reducing its variance.

Note that, whereas, in a repeated survey with independent samples at each occasion, net change reflects a combination of changing values and changing population composition, on the contrary in a

³ Model-based estimators are not the only class of methods applied in this field, even if they have a central role.

panel survey, unless steps are taken to incorporate new entrants at later waves as in rotating panels, net change reflects only changing values but refer to the initial population. See Duncan and Kalton (1987) for a comprehensive review of the design and analysis of longitudinal data. See also the module “Repeated Surveys – Repeated Surveys” for a discussion of possible alternative sampling designs to be applied in repeated surveys. In this module focus is given to panel surveys and in particular way to issues in estimation and in determination of sampling weights.

An important preliminary matter when a panel survey is conducted is the definition of continuity rules in order to establish whether an enterprise represent the same unit over the different sampling occasions (waves). This definition, of course, affects definition of target population and statistical units and have effect on sample definition. The interested reader can refer to the modules “Repeated Surveys – Repeated Surveys” and “Dynamics of the Business Population – Business Demography” where aspects of continuity rules are discussed, here we focus on relevant issues in the determination of sampling weights. In this respect, let us note that in a panel survey, two types of weights can be calculated: cross-sectional weights and longitudinal weights. This distinction depends on the nature target populations and related parameters. Cross-sectional weights refer to a population of a given wave and are used to estimate parameters of the given population. Longitudinal weights are used to estimate parameters referring to the longitudinal population, i.e., the population in different occasions. Examples of the latter are gross change⁴ and measures of frequency, timing and duration of events occurring within a given time period.

Definition of cross-sectional weights for each wave of the survey to reproduce the target population proceed similarly to the standard cross-sectional surveys, but may require specific computation when using panel surveys.

Evaluation of cross-sectional weights for the first occasion of the survey follows the standard steps described in subsections 2.1, 2.2 and 2.3: determination of a design weight equal to the inverse of the inclusion probability and subsequent adjustment for non-response and for improving estimators.

It has to be underlined that if no-renewal is done in the panel, the sample is in fact representative only of the initial population. Moreover, even if the population is fixed, after the first wave, determination of weights should take attrition into account. Then, at each subsequent wave, the first operation should consist in adjusting the first wave weights for non-response due to attrition. On the other hand, as population is subject to changes, it is important to modify weights to reflect these changes, as well. If updated totals are available then calibration to new totals can reduce presence of bias (see also subsection 2.10).

If, on the contrary, a refreshment of panel is done to represent the population dynamics, the sample in a given wave is composed of different parts. To obtain estimate of cross-sectional indicators, two different approaches may be applied. One approach is determining weights for each component and then combine the estimates, the second consists in pooling the two samples by assigning a unique weight. This second method may be less straightforward to apply in practice due to complexity in computation of inclusion probability in both samples.

⁴ On the contrary, estimation of net change requires use of cross-sectional weights at each wave and proceed as described in sub-section 2.1.

Determination of longitudinal weights requires first definition of the target population, which may be for example the set of units present both at time t_0 and at time t , or the initial population at time t_0 only. In the first case, for example, one assigns weights only to overlapping units in the two different samples and the longitudinal weight is given by the product of cross-sectional weight in t_0 and the conditional weights to units in t that belong to t_0 .

Use of panel survey is much more established in sampling surveys on households where examples of weights definition can be found (e.g., Verma et al., 2006). An example of panel in a business survey with discussion of different weights usage can be found in Australian Bureau of Statistics (2000, pages 9-20).

It has been mentioned at the beginning of this subsection, that repeated observations on the same units allows reduction of variance of the estimators of changes. An additional advantage of observing repeatedly the same units consists in the possibility of explicitly exploiting the temporal correlation which arises between observations on the same units at the different occasions of the survey to improve estimators on the basis of models which take into account the autocorrelation between observation on the same units at different time points to estimate cross-sectional measures. Model-based estimators for panel data are for example proposed in Fabrizi et al. (2007).

Before concluding this subsection, it is important to note that overlapping of samples induced with (rotated) panels require also special concern for variance estimation both for estimation of levels when more sampling occasions are involved (e.g., means of quarters in a year) and for estimation of changes. In fact, for example, when estimating a measure of change, the variance estimation of this estimate has to account for the correlation between estimators at different times of the repeated survey. See for example, Nordberg, (2000) for a proposal for coordinated sample with permanent random numbers, Qualité and Tillé (2008) for a proposal of variance estimation of changes in repeated surveys, Berger (2004) and finally, Knottnerus and Van Delden (2012) for variance estimation of changes in rotating panels. See EUROSTAT (2013) for review of variance estimation methods and key references.

2.7 *Preliminary estimates*

As already mentioned in subsection 2.5, timeliness in disseminating the estimates is a very important aspect of the quality of short term statistics and it is also one of the main peculiarities of them.

For short term statistics, in fact, the planned sample may occur to be partially observed when the estimates have to be disseminated. Preliminary (provisional or early) estimates are the estimates that are computed using the statistical information available on the basis of the preliminary sample (PS), i.e., the subset of the planned final sample (FS) that is observed at time of first release of the estimates.

The main problem that has to be faced in a short-term preliminary estimation context concerns the possible self-selection of early respondents, since self-selection can lead to biased estimators of the unknown population mean and variances. Early respondents may have systematically different (e.g., lower) values in terms of the target variables from late respondents.

Preliminary estimation methods may be classified in function of the stage on which the preliminary method is applied.

In fact, it is possible to identify different methods according to the stage they are implemented in:

1. the sampling design stage, by selecting a preliminary subsample of the planned sample (see the module “Sample Selection – Subsampling for Preliminary Estimates”);
2. the estimation stage, in the following ways:
 - a) by means of imputation techniques of missing data, that are applied to non-respondent units in FS but not in PS;
 - b) by means of weighting adjustment, i.e., modifying the sampling weights assigned to the units in PS in order to take into account non respondents of the FS;
 - c) by applying direct and indirect estimators, using known population totals of auxiliary variables and/or time series of preliminary and final estimates of the variable of interest (see the modules “Weighting and Estimation – Preliminary Estimates with Design-Based Methods” and “Weighting and Estimation – Preliminary Estimates with Model-Based Methods”).

The different approaches can be compared in terms of bias and revision error, i.e., the difference between preliminary and final estimates.

See the module “Weighting and Estimation – Preliminary Estimates with Design-Based Methods” for a description of design methods, in particular for a method proposed in Rao et al. (1989) which at time t exploit time t and $t-1$ data aiming at minimising the mean square error of the estimator. Moreover, see the module “Weighting and Estimation – Preliminary Estimates with Model-Based Methods” for a description of a model-based estimator proposed by Rao et al. (1989), which introduces model that use disaggregated auxiliary information coming from survey data at previous times and/or administrative register data. For these, the relationship between the variable of interest and the auxiliary variables is usually formalised through domain level models in which the auxiliary information is expressed in terms of domain known totals or estimates. An estimation technique of the latter class was developed by Rao et al. (1989). In their proposal, preliminary estimates are computed on the basis of a first order autoregressive model for final estimates and revision errors.

2.8 *Small area estimation*

The aim of small area (domain) estimation methods is to produce reliable estimators for the variable of interest under budget and time constraints. In fact, National Statistical Office surveys are usually planned for large domains. Hence, whenever more detailed information is required, the sample size may be not large enough to guarantee the release of direct estimators at the desired level of disaggregation. In the most extreme cases direct estimator cannot be calculated when no units belonging to the domain occur in the observed sample. For instance, one is interested in the overall amount of industrial turnover for the whole population of business enterprises, and also in estimating analogous parameters with respect to relevant population sub-sets, i.e., sub-populations corresponding to geographical partitions (e.g., administrative areas) or sub-populations associated to economic cross-classification (e.g., enterprise size and sector of activity).

When domain estimates based on direct estimator cannot be disseminated because of unsatisfactory quality, an ad hoc class of methods, called *small area estimation* (SAE) methods, is available to solve the problem. These methods are usually referred as *indirect estimators* since they cope with poor

information for each domain by borrowing strength from the sample information belonging to other domains, resulting in increasing the effective sample size for each small area, i.e., the sample size that affects variances.

This means that their variability does not depend on the sample size of domain d , but on sample size of a larger area (see Rao, 2003).

More precisely, the increase in efficiency of SAE is obtained by means of information on units belonging to other areas considered geographically close or similar with respect to structural characteristics to the small area of interest. In practice, an improvement in the efficiency of the estimators can be achieved by assuming, implicitly or explicitly, a relationship which links together sampling units in the small area of interest and sampling units in the small areas which behaves similarly to the small area of interest. Enhanced methods are involved when applying model using complex spatial or temporal information. In particular, the model using temporal information may be useful in case of repeated surveys, i.e., when several survey occasions are available. In fact, in this case it would be possible to use the information from the previous survey occasions or times.

An account of small area estimation is given in the module “Weighting and Estimation – Small Area Estimation”. Specific small area methods, both design-based and model-based, are described in the modules “Weighting and Estimation – Synthetic Estimators for Small Area Estimation”, “Weighting and Estimation – Composite Estimators for Small Area Estimation”, “Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)”, “Weighting and Estimation – EBLUP Unit Level for Small Area Estimation”, and, finally, “Weighting and Estimation – Small Area Estimation Methods for Time Series Data”.

Area and unit level EBLUP are both based on linear mixed model assuming a random area (domain) effect to take into account extra variability between areas not accounted for by the linear relationship between target and auxiliary variables. Both estimators are a linear combination of the direct estimator and the synthetic prediction resulting from the model. The area level EBLUP can be applied also when only macrodata referred to domain level are available, in this case variance of the direct estimator has to be (or assumed to be) known. Furthermore, to exploit temporal information a dedicated method module “Weighting and Estimation – Small Area Estimation Methods for Time Series Data” is provided. Some of these methods are based also on linear mixed models, in which time random effect is introduced or alternatively on auto-regressive specifications.

For a review of recent developments on small area estimation, see Pfeffermann (2013).

2.9 *Integration of administrative sources in the statistical production*

Nowadays there is an increasing interest in using administrative data for production of official statistics. The administrative data are meant not only as a source of auxiliary information or as a tool for building sampling frames, but also as a source of statistical information itself in place of sample surveys and censuses (Wallgren and Wallgren, 2007), in order to reduce costs and statistical burden.

Hence, though, traditionally, administrative records are used to support the survey work, now more and more increasingly, administrative records are given a central role in the statistical process, to completely replace the collection of survey data. Sample surveys are now part of a more complex system where more sources and surveys are combined together. In some cases they represent the

supplementary data that may be used to adjust for data quality (see Eltinge, 2011) or to complement administrative data when coverage issues arise.

Having administrative data acquired a relevant role in the production of official statistical output, the issue of establishing a framework for assessing, measuring, documenting and reporting on quality of administrative data sources and its statistical potential usability has received a considerable attention. An example of a framework for assessment of quality of administrative data can be found, for example, in Daas et al., 2011, mainly developed within the European project BLUE-ETS (<http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.1.pdf>; Laitila et al., 2011). In the present handbook, the module “Weighting and Estimation – Estimation with Administrative Data” reports the main aspects to be considered when administrative data are used to replace in part or completely sample surveys.

More in general, the issue of integrating administrative and sample sources has emerged. Chambers et al. (2006) designate the model-based estimation approach as a natural framework for integrating sources in the statistical production. In this context, a proposed solution is fitting a model on the sample and applying it to predict values for units on the unobserved part of the sample using information obtained from administrative data. The ESSnet on Administrative Data (<http://www.cros-portal.eu/content/admindata-sga-3>) reports various experimental applications of this approach. The module “Weighting and Estimation – Estimation with Administrative Data” describes practical uses of administrative data in business statistics and gives suggestions on which methods is more appropriate according to the informative context (timeliness and coverage) of the administrative source providing data.

2.10 Departure from ideal conditions: imperfect frames

In this subsection some of the most typical departures from the ideal conditions are the basis of sampling and estimation methodologies are highlighted.

The most common case is when sample is selected from a not updated frame. In this context, it may occur that some values of the stratification variables used for sampling differ from those observed in the selected units. For instance, the observed enterprise size, measured as the number of employees, can change from the measure registered in the sample frame. When new totals are available, a post-stratified estimator can be applied in order to take into account of this updated information during the estimation phase. See the module “Weighting and Estimation – Calibration” for more details on how the post-stratification estimator can be applied.

A second important concern arises due to demography of the statistical units (enterprises). Since business population experiences rapid changes, the sampling frame is affected by some degree of overcoverage and undercoverage, i.e., units in the frame are no longer in the target population or vice versa. For units not covered by the frame (undercoverage) there is a zero probability, and this feature may cause biased estimators. See Lundström and Särndal (2001, page 139) for a formalisation of the context and for possible solutions, in particular in Section 11.3 the calibration approach is described when updated known totals can be used at estimation stage.

Finally, an important issue arises when mergers and splits occurs between the construction of the frame and the surveys. In this case sampled units do not correspond to units recorded in the sampling frame. This population dynamics affects the sampling inclusion of the final units and may introduce

bias in the final estimates, if it is not properly taken into account. This context can be formalised with indirect sampling (Lavallée, 1995, Deville and Lavallée, 2006). In fact, the sampling step is carried out on a frame not containing the target population units but linked to them. The major difficulty with this approach consists in recognising the links between sampling list and target population produced by mergers or splitting and in determining the correct weights.

The Generalised Weight Share Method (GWSM) has been developed by Lavallée (1995) and Deville and Lavallée (2006). Lavallée and Labelle-Blanchet (2013) present the method for skewed populations. The Swiss Federal Statistical Office applied the weight share method for the estimations of the Quarterly Job Statistics:

http://www.bfs.admin.ch/bfs/portal/fr/index/infothek/erhebungen_quellen/methodenberichte.html?publicationID=3217%20.

More details on imperfect frames are reported in the module “Weighting and Estimation – Design of Estimation – Some Practical Issues”.

3. Design issues

The choice of estimations methodology is strictly related to the main aspects of quality described in the module “Quality Aspects – Quality of Statistics”. The main features are accuracy, coherence, timeliness. Choice of estimation methodology is also highly related to characteristics of the sampling design (e.g., probabilistic or cut-off sampling). Here we give a brief summary of the quality and sampling factors to be considered. First of all, one should determine if administrative data are available, accessible and can be used for direct production of statistical output according to schema in the module “Weighting and Estimation – Estimation with Administrative Data”.

Whenever administrative data are not available, and sampling is carried out to achieve the required information, in order to choose the proper estimation methodology, one should take into account of the sampling mechanism. If a non-random mechanism is applied, a model-based estimator can be applied. However, the risk of this sampling strategy is that when the model is not valid bias may be present.

On the other hand, bias can be present, also in case of a probabilistic sampling design, when there are no-respondents and no proper measure are taken, for example, by means of adjustment of weights.

Moreover, when external constraints are given then benchmarking to these external constraints can be obtained with calibration estimators and rereighting. This solution is not always applicable in practice and problems can be encountered in practice to meet too many constraints.

Similarly, to improve accuracy of estimators, in particular to reduce variance one may want to use of auxiliary information correlated with the target (e.g., calibration estimators). However, in practice many target indicators are produced by a survey and a single weight is used for a single survey. Then, the auxiliary variables of the calibration estimator or GREG estimator will likely relates only with one (or few) of the target variables. Another class of methods to (further) improve accuracy when large variability of HT (or calibration) estimators is caused by skewness of the distribution of the target variable is robust estimation.

If calibration estimators still does not satisfy the needed accuracy level (this will often occur with unplanned domain, but also in case of large no-response), small area estimators may represent a possible solution to guarantee the desired degree of information.

Similarly, if there is a need to obtain estimates with incomplete sample observations to meet timeliness, preliminary estimators may be applied.

Many of the aspects recalled in this section, conflict with each other and compromise solutions should be considered between the different competing needs, aiming at guaranteeing the quality of the estimates.

The module “Weighting and Estimation – Design of Estimation – Some Practical Issues” provides more details on practical issues to be considered in designing the estimation methodology.

4. Available software tools

There are several software tools to perform estimation using basic weights or calibration estimators together with variance estimation (see the topic “Quality Aspects”). In the following we classify some of them requiring open source R or the commercial software SAS and SPSS.

The following packages R are available from the R-CRAN archives:

Package survey, <http://cran.r-project.org/web/packages/survey/index.html>

Package sampling, <http://cran.r-project.org/web/packages/sampling/index.html>

A full-fledged R system for design-based and model-assisted analysis of complex sample surveys *REGENESEES* is available at <http://joinup.ec.europa.eu/software/regenesees/release/all>

The following programs allows to calibrate weights and calculate variance estimation:

BASCULA <http://www.cbs.nl/en-GB/menu/informatie/onderzoekers/blaise-software/blaise-voor-windows/productinformatie/bascula-info.htm>

CALMAR is a SAS macro developed by the French National Statistics Office (INSEE),

CLAN is a system of SAS macros developed by Statistics Sweden.

GENESEES (SAS macro by Italian statistical Institute <http://www.istat.it/it/strumenti/metodi-e-software/software/genesees>)

GES, developed in Statistics Canada, is also a system of SAS macros

g-Calib (SPSS by Statistics Belgium)

See Eurostat (2013) for further details on these software tools.

Many tools are available to perform small area estimation methods as well

1. The collection of SAS macros included in the zip file [The EURAREA 'Standard' estimators and performance criteria](http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html) of the EURAREA project (<http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html>)
2. the R functions produced by ESSnet SAE (ESSnet/sae portal http://www.cros-portal.eu/sites/default/files/R_codes_%26_documentations_3.zip)
3. R package sae2 (BIAS project website: <http://www.bias-project.org.uk/>)
4. SAMPLE project codes in <http://www.sample-project.eu/it/the-project/deliverables-docs.html>

A description of functions and software for small area estimation can be found in WP4 final report of ESSnet SAE.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Australian Bureau of Statistics (2000), Business Longitudinal Survey, Confidentialised Unit Record File, 1994-95, 1995-96, 1996-97, 1997-98.
- Beaumont, J.-F. and Alavi, A. (2004), Robust generalized regression estimation. *Survey Methodology* **30**, 195–208.
- Beaumont, J. F. and Rivest, L. P. (2008), Dealing with Outliers in Survey Data. In: C. R. Rao and D. Pfeffermann (eds.), *Handbook of Statistics, Design, Methods and Applications*, Vol. 29.
- Benedetti, R., Bee, M., and Espa, G. (2010), A Framework for Cut-off Sampling in Business Survey Design. *Journal of Official Statistics* **26**, 651–671.
- Berger, Y. G. (2004), Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics* **32**, 451–467.
- Chambers, R. L. (1986), Outlier robust finite population estimation. *Journal of the American Statistical Association* **81**, 1063–1069.
- Chambers, R. L. and Kokic, P. (1993), Outlier robust sample survey inference. Invited paper, *Proceedings of the ISI 49th session, Firenze, Italy, August 1993*, 55–72.
- Chambers, R., van den Brakel, J. A., Hedlin, D., Lehtonen, R., and Zhang, L.-C. (2006), Future Challenges of Small Area Estimation. *Statistics in Transition* **7**, 759–769.
- Daas, P. and Ossen, S. (2011), List of quality groups and indicators identified for administrative data. Deliverable 4.1, FP7 BLUE-ETS project.
- Deville, J.-C. and Lavallée, P. (2006), Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology* **32**, 165–176.
- Deville, J.-C. and Särndal, C.-E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.
- Duncan, G. and Kalton, G. (1987), Issues of Design and Analysis of Surveys across Time. *International Statistical Review* **55**, 97–117.
- Ferrante, M. R. and Pacei, S. (2004), Small Area Estimation for Longitudinal Surveys. *STATISTICAL METHODS & APPLICATIONS* **13**, 327–340.

- Gross, W. F., Bode, G., Taylor, J. M., and Lloyd-Smith, C. W. (1986), Some finite population estimators which reduce the contributions of outliers. *Pacific Statistical conference: Proceedings of the congress*, Auckland, New Zealand, 20–24.
- Eltinge, J. L. (2011), Two approaches to the use of administrative records to reduce respondent burden and data collection costs. UNECE.
http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.42/2011/mtg1/USA_TwoApproaches.pdf
- EUROSTAT (2013), *ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys*. Methodologies and working papers.
- Fabrizi, E., Ferrante, M. R., and Pacei, S. (2007) Small area estimation of average household income based on unit level models for panel data. *Survey Methodology* **33**, 187–198.
- Hidiroglou, M. A. and Srinath, K. P. (1981), Some estimators of population total from simple random samples containing large units. *Journal of the American Statistical Association* **76**, 690–695.
- Houbiers, M., Knottnerus, P., Kroese, A. H., Renssen, R. H., and Snijders, V. (2003), Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005, Statistics Netherlands, Voorburg / Heerlen. <http://www.cbs.nl/NR/rdonlyres/6C31D31C-831F-41E5-8A94-7F321297ADB8/0/discussionpaper03005.pdf>
- Huber, P. J. (1981), *Robust Statistics*. John Wiley & Sons, New York.
- Hulliger, B. (1995), Outlier robust Horvitz-Thompson estimators. *Survey Methodology* **21**, 79–87.
- Kalton, G. (1983), Models in the Practice of Survey Sampling. *International Statistical Review* **51**, 175–188.
- Kokic, P. N. and Bell, P. A. (1994), Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics* **10**, 419–435.
- Knottnerus, P. and van Delden, A. (2012), On variances of changes estimated from rotating panels and dynamic strata. *Survey Methodology* **38**, 43–52.
- Laitila, T., Wallgren, A., and Wallgren, B. (2011), *Quality Assessment of Administrative Data*. Research and Development – Methodology reports from Statistics Sweden, 2.
- Lavallée, P. (1995), Cross-sectional weighting of longitudinal surveys of individuals and households using weight share method. *Survey Methodology* **21**, 25–32.
- Lavallée, P. and Labelle-Blanchet, S. (2013), Indirect sampling applied to skewed populations. *Survey Methodology* **39**, 183–215.
- Lee, H. (1995), Outliers in business surveys. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York.
- Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd edition. Wiley, New York.
- Lundström, S. and Särndal, C.-E. (2001), Estimation in the presence of Nonresponse and Frame Imperfections. Statistics Sweden.

- Nordberg, L. (2000), On Variance Estimation for Measures of Changes When Samples are Coordinated by Use of Permanent Random Numbers. *Journal of Official Statistics* **16**, 363–378.
- Pfeffermann, D. (2013), New Important Developments in Small Area Estimation. *Statistical Science* **28**, 40–68.
- Qualité, L. and Tillé, Y. (2008), Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology* **34**, 173–181.
- Rao, J. N. K., Srinath, K. P., and Quenneville, B. (1986), Estimation of Level and Change using Current Preliminary Data. In: Kasprzyk, Duncan, Kalton, and Singh (eds.), *Panel Surveys*, John Wiley & Sons, New York, 457–485.
- Rao, J. N. K. (2003), *Small Area Estimation*. John Wiley and Sons, New York.
- Renssen, R. H. and Nieuwenbroek, N. J. (1997), Aligning Estimates for Common Variables in Two or More Sample Surveys. *Journal of the American Statistical Association* **92**, 368–374.
- Royall, R. M. (1970), On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–387.
- SAE ESSnet (2012), Deliverables of the project. <http://cros-portal.eu/projectdetail/1392>.
- Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley and Sons, New York.
- Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992), *Model Assisted Survey Sampling*. Springer Series in Statistics, Springer-Verlag, New York.
- Vaillant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference, a Prediction Approach*. Wiley, New York.
- Verma, V., Betti, G., and Ghellini, G. (2006), Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC. Working paper, 67, University of Siena, available at http://www.econ-pol.unisi.it/dmq/pdf/DMQ_WP_67.pdf.
- Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics – Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester, England.

Interconnections with other modules

8. Related themes described in other modules

1. User Needs – Specification of User Needs for Business Statistics
2. Repeated Surveys – Repeated Surveys
3. Dynamics of the Business Population – Business Demography
4. Sample Selection – Main Module
5. Statistical Data Editing – Main Module
6. Imputation – Main Module
7. Weighting and Estimation – Design of Estimation – Some Practical Issues
8. Weighting and Estimation – Small Area Estimation
9. Weighting and Estimation – Estimation with Administrative Data
10. Quality Aspects – Quality of Statistics

9. Methods explicitly referred to in this module

1. Sample Selection – Subsampling for Preliminary Estimates
2. Weighting and Estimation – Calibration
3. Weighting and Estimation – Generalised Regression Estimator
4. Weighting and Estimation – Outlier Treatment
5. Weighting and Estimation – Preliminary Estimates with Design-Based Methods
6. Weighting and Estimation – Preliminary Estimates with Model-Based Methods
7. Weighting and Estimation – Synthetic Estimators for Small Area Estimation
8. Weighting and Estimation – Composite Estimators for Small Area Estimation
9. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
10. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation
11. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 5.5 Calculate weights
2. 5.6 Calculate aggregates

12. Tools explicitly referred to in this module

1. Software tools for estimation, calibration of weights, variance estimation, application of small area methods

13. Process steps explicitly referred to in this module

1. Sampling, Estimation and Evaluation of Accuracy

Administrative section

14. Module code

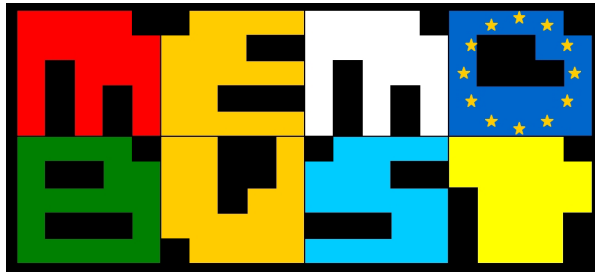
Weighting and Estimation-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-03-2012	first version	Loredana Di Consiglio	ISTAT
0.2	05-07-2012	after first review	Loredana Di Consiglio	ISTAT
0.2.1	03-09-2012	new version to take into account the modified workplan	Loredana Di Consiglio	ISTAT
0.3	14-05-2013	revised version	Loredana Di Consiglio	ISTAT
0.4	08-10-2013	after review	Loredana Di Consiglio	ISTAT
0.5	10-01-2014	after CH-HU-SE reviews	Loredana Di Consiglio	ISTAT
0.6	24-02-2014	after HU-SE-EB reviews	Loredana Di Consiglio	ISTAT
0.6.1	28-02-2014	minor revisions	Loredana Di Consiglio	ISTAT
0.6.2	11-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:31



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Design of Estimation – Some Practical Issues

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 The setting of a business survey and some typical issues	3
2.2 Estimation principles	4
2.3 A few estimation characteristics.....	5
2.4 Some comments on sampling design	6
2.5 Estimation with non-response and skewed distributions.....	7
2.6 Estimation with over- and under-coverage.....	7
2.7 Using auxiliary information in both sample design and estimation	8
2.8 Handling organisational changes in a sample.....	8
2.9 Outliers in the sample.....	10
2.10 Cut-off sampling.....	11
2.11 Models: early estimates	11
2.12 Models: small area estimation	11
2.13 Use of administrative and other accessible data.....	12
2.14 Evaluation of the design	12
3. Design issues	13
4. Available software tools.....	13
5. Decision tree of methods	13
6. Glossary.....	13
7. References	13
Interconnections with other modules.....	15
Administrative section.....	16

General section

1. Summary

This module discusses a set of issues to consider in the design of the estimation of a business survey with emphasis on practical problems rather than on theory. Decisions have to be made together with other design decisions. The issues discussed mostly have neither obvious nor simple solutions. There is some theoretical ground to use, though, together with practical experience. The design decisions should be reasonable from these perspectives. The estimation procedure needs to work in practice also with a pressure on timeliness. It is important to save information about the procedure, including its weak points, in order to be able to improve successively.

There are different types of statistical inference depending on the combination of data and models used. The practical problems include, for instance, skewed distributions, outliers, coverage deficiencies, non-response, organisational changes, early estimates, small domains of estimation, and handling of administrative data.

The requests on quality include, for example, accuracy, coherence, and timeliness. Most business surveys are used both for primary statistics and as input to the National Accounts. In order to achieve coherence similar methods must be used in all business surveys.

Several types of statistical units are used in business surveys. Here the term ‘enterprise’ is used in most of the cases discussed, out of convenience.

2. General description

Some typical characteristics for a business survey are described in Section 2.1; see also Rivière (2002). Some of the corresponding design considerations are discussed in Sections 2.2–2.13 with emphasis on practical issues. Finally, there is a section on evaluation.

2.1 *The setting of a business survey and some typical issues*

There is a statistical Business Register (BR) providing frames and information about different unit types used for business statistics, see, for instance, the handbook module “Statistical Registers and Frames – Survey Frames for Business Surveys” and other modules in that topic.

The distribution of most continuous variables (like turnover) is strongly skewed. A fairly small or limited number of large enterprises account for a noticeable or considerable portion of the total turnover, the total number of employees etc. See, for instance, figures given by Assoulin (2009, p. 2) and by the Blue-Ets-project (2013, p. 39). It is also typical that enterprises change quickly due to reorganisations, births, and deaths.

It is frequently the case that a statistical office has a special unit or group devoted to data collection from the largest enterprises, a group that keeps up-to-date with the organisational changes of these enterprises, serves as point of contact between each enterprise and many surveys, simplifies their data provision etc. See, for instance, the handbook modules “Data Collection – Design of Data Collection Part 2: Contact Strategies” and “Data Collection – Mixed Mode Data Collection” (about tailoring).

There is an increased interest in reducing direct data collection and in using other, already existing, data, typically administrative data such as tax data; see the handbook module “Data Collection –

Collection and Use of Secondary Data”. There are both positive and negative implications of such use: lower costs, lower response burden, and a data set that is close to a census, but also a dependence on the alternative source with its unit type(s), population(s), variables, and reference times. The estimation possibilities become both greater and more restricted than with a sample survey with direct data collection. For instance, estimation for small areas and other small domains may become possible. On the other hand, it may not be possible to influence variables, editing, and production time.

There are some different types of business statistics and surveys in the European Statistical System, see, for instance, the handbook module “General Observations – Different Types of Surveys”. There are annual statistics with a high degree of detail, such as the Structural Business Statistics (SBS). There are short-term statistics with a high pressure on timeliness and often with focus on changes over time, such as the STS: Short-Term business Statistics on industry, construction, retail trade and other services. There are further, secondary, statistics like the National Accounts (NA), which build on the primary business statistics just mentioned. There are many requests on the output quality of the primary statistics, stemming from European Regulations, national needs, the NA etc. Different needs often have to be balanced. The pressure on timeliness for short-term statistics leads to a system where there are preliminary statistics, revisions, and final statistics. See, for instance, the handbook module “Repeated Surveys – Repeated Surveys”.

2.2 *Estimation principles*

A statistical estimation procedure starts with some data and arrives at a set of statistics; to do so it uses some principle for the statistical inference. The principle usually includes an element of randomness. Statistical offices have a fairly long tradition of drawing a random sample, collecting data, and making inference to a finite population. Selection probabilities are then an essential ingredient. This is called a design-based method or inference.

The design-based principle may be extended to model-assisted estimation, where auxiliary information is used. Improvement of the accuracy is a major aim. Some form of assumption – implicit or explicit – is included. There may, for instance, be a model for the survey variable(s) in terms of the auxiliary variables. Non-sampling errors, like non-response, may be included in the modelling approach. Still, the sampling design is the foundation for the estimation.

Another type of principle uses a statistical model with its assumptions as the basis for the statistical inference. This is a model-based method. It is sometimes called a prediction approach. The model plays the major role in the inference, and the sampling procedure (if any) is less important. Model-based methods use, for instance, an assumption about a non-sampled part of the population or about relationships between variables or over time. They may be preferred, adequate, or even necessary to use. Small area estimation, early estimates, and non-random samples provide examples.

In general, no or negligible bias and a small variance are the basic demands for the choice of an estimator, possibly summarised in terms of a minimum or small mean squared error (MSE). Obviously, the type of statistical inference must be stated first, in order for bias, variance etc. to have a meaning. Design-based, model-assisted, and model-based methods use different approaches to randomness, and they include somewhat different information. The estimators are different, and so are their properties, since they depend on different assumptions about random elements.

If administrative data are used, some modelling and adjustments may be necessary, for instance to derive or model statistical variables based on the administrative variable definitions and to go from administrative units to statistical units. There are not (yet) generally agreed inference principles for this type of data. It is essential to state clearly the target for the inference, for instance the target population. This is not always done. There may be coverage deficiencies, missing data, and other non-sampling errors to handle in the estimation. It is important to be aware of this and transparent.

The estimation procedure has to be designed together with other parts of the survey design. Important enablers and issues to consider are the BR, further accessible data sources, data collection, sampling method (if any) and questionnaire design (when relevant), accuracy, and other output quality requests. The targets of the statistical output and the statistical inference to be used need to be stated; see, for example, the handbook modules “Overall design – Overall Design” and “Weighting and Estimation – Main Module”. There is a lot of literature on estimation principles. A recent description is given by Valliant (2013), who comments on a summary report on non-probability sampling made by Baker, Brick, Bates, Battaglia, Couper, Dever, Gile, and Tourangeau (2013). There are further comments and a rejoinder by Baker et al. (2013). Even if these discussions are more relevant for household statistics, there are many general comments, which are useful also for business statistics and as a starting point to further literature. See also Zhang (2012), who provides a description of register-based statistics.

2.3 A few estimation characteristics

The statistical output consists of many estimates; there are normally many tables, each of which has many table cells. Hence priorities have to be made among all estimators and accuracy requests expressed, more or less explicitly. There are balances to make, such as follows. Which is more important in the trade-offs below? This should be considered before choosing the estimator – and choosing the sampling method, including the sample allocation.

- The overall table “total”, the margins of the table, or the table cells.
- The current level or the change since a previous period.
- The absolute level or the relative level.

There are always differences between the current reality and the information in the BR. This is often due to delays in reporting about births, deaths, and organisational changes. There may also be incorrect information in the BR, for instance about the industry/activity of the enterprise. It is obviously important for the BR to have good and frequently updated sources. The BR will still be a bit behind. The deficiencies and their implications are asymmetric for the survey. The current sample includes over-coverage, which may be possible to detect, whereas outside information is needed to find under-coverage. There are two main ways to obtain and use information that is more up-to-date than the frame information. Firstly, there may be a more recent register – an updated version of the BR or the frame, or a similar register – to use as support in the estimation procedure. Secondly, differences from the frame information may be observed in the survey. Such information often needs to be taken into account, after adequate checking. The cases with negligible or small effects are fairly rare.

Size is mostly an essential factor in the survey design. It is essential to obtain data of good quality from the large enterprises. Small enterprises that have grown in comparison with the information used in the design may be disturbingly influential if straight-forward estimation formulas are used. Hence, care should be taken to have recent and adequate measures of size.

2.4 *Some comments on sampling design*

The sampling and estimation methods should be chosen together with regard to the targets of the survey and the priorities made. Choudhry, Rao, and Hidirolou (2012) show how the sample can be allocated to satisfy accuracy requests when estimating a set of sub-population means; this is just one example to illustrate the strong link.

The reference time of the target population usually agrees with the reference time of the survey variables, for instance in a monthly survey. If estimates of change are important, it is normally favourable to include the same unit repeatedly in the sample. However, some exchange is necessary, both to include changes in the target population and with regard to response burden for the enterprises selected.

Selecting a new sample from an updated frame leads to more recent information about the population than continuing with the old sample from the old frame. On the other hand it means that enterprises may move in and out of the sample, which is inconvenient for such enterprises and an addition to their response burdens. It means additional work also for the statistical office. The frequency of updates of frames and samples is discussed for example in the handbook modules “Statistical Registers and Frames – Survey Frames for Business Surveys” and “Repeated Surveys – Repeated Surveys”. Many countries renew or complement their samples fairly frequently. An addition is typically selected from population parts that would otherwise be under-coverage.

Rather many countries use some form of sample co-ordination, at least over time, but possibly also between surveys. See the handbook module “Sample Selection – Sample Co-ordination”. There are two main reasons, as already indicated. Firstly, accuracy is improved through positive co-ordination of samples (which means more overlap than expected in the case with independent samples; typically used over time for each survey and between related surveys at the same time). Secondly, response burden is spread over enterprises through negative co-ordination (less overlap; typically used between unrelated surveys at the same time). Time aspects and a full set of surveys are taken into account. The principal aim is that an enterprise is selected for one or possibly a few sample surveys during a period and then, ideally, has “survey holidays”. The success depends on the number of surveys, the number of enterprises to select from, and the methodology chosen. The response burden can, of course, not be spread unless there are a (fairly) large number of enterprises. The positive co-ordination over time is limited with regard to response burden, for instance expressed as an enterprise being selected a targeted number of years.

There is some similarity between a sample survey with positive co-ordination over time and a panel survey. However, the latter has a pre-determined pattern of overlap, whereas the former contains random elements. Therefore, the term panel is normally not used in systems with sample co-ordination.

With panels it may be necessary to consider both cross-sectional and longitudinal weights; see, for instance, the handbook module “Weighting and Estimation – Main Module”. As described in the handbook module “Sample Selection – Sample Co-ordination” there are techniques to co-ordinate samples such that each sample can be considered as random, which make the estimation design straight-forward. There is, however, dependence between samples over time, which has to be taken into account in the estimation of variance for estimators of change. See also the handbook module “Repeated Surveys – Repeated Surveys” for a discussion and a few references.

2.5 *Estimation with non-response and skewed distributions*

There are two main methods to handle non-response in the estimation. It is frequently the case that imputation is used for item non-response; see the handbook module “Imputation – Main Module”. Similarly, reweighting is often used for unit non-response; see the handbook module “Weighting and Estimation – Main Module”, where response propensity is described and furthermore weight adjustments like calibration. This approach with imputation and reweighting is a reasonable starting point for the design of estimation. However, the nature of the distribution of the target variables must also be considered. This is a general aspect, which is necessary to take into account also when it comes to non-response.

For continuous variables which are highly skewed, say turnover, it is of great importance to aim at no unit non-response for the largest enterprises. If there still is unit non-response among the largest enterprises, different treatments of the non-response for different segments of the target population should be considered. The design choice may be a split into two major methods. Non-response among the smaller enterprises is handled by reweighting, which is with or without auxiliary information. Since the largest enterprises normally are unique, it may be wise to consider imputation for them. It needs to be a careful imputation, which builds on information for the particular enterprise, such as previous values. It is often manual. See also the handbook methodological module “Statistical Data Editing – Manual Editing”.

It should be observed that non-response may be related to over-coverage or to organisational changes, as further discussed in the next sections.

2.6 *Estimation with over- and under-coverage*

A possibility to reduce coverage problems – in the estimation procedure rather than in the sampling procedure – is to use auxiliary information from an updated frame or from a similar register when building the estimator. Calibration against updated auxiliary information is a way to handle the problems with over-coverage and under-coverage, as far as the new version of auxiliary information goes.

There are, however, some practical drawbacks with the calibration method. Unexpected results may be difficult to penetrate and resolve, since the estimator is rather complicated. Moreover, with quick production rounds in short-term statistics, the difficulties with matching microdata should not be underestimated.

Hence, under-coverage may be reduced through the sampling or estimation procedure or both. If neither is used in short-term statistics and if over-coverage is set to zero, the level of the estimates will decrease gradually as time goes and the effects of over-coverage increase. This may be balanced through some model assumption. ONS (2013, p. 6) and ONS (2012, pp. 41–44) are two examples showing how design weights are adjusted to unit non-response and to births and deaths with an assumption about the births-to-deaths ratio. The basic assumption used in these cases is that the ratio is equal to 1. Large enterprises, which are included with probability one, are a natural exception. A further possibility is to analyse birth and deaths rates and make some extrapolation over time. BLS (2013, p. 37) illustrates variation over time for some variables, and the estimation methodology used for a large employment survey is described in these technical notes.

The coverage issues just discussed are relevant not only for the population as a whole but also for sub-populations, for instance industries/activities within the target population. An erroneous or changed NACE code may be detected in the sample. Again there is a design choice whether to use the “original” NACE code in the frame or to use the correct one. Several issues should be considered, including bias and variance of the possible estimators. The enterprise may have both principal and secondary activities. To “move” such an enterprise between domains of estimation may exaggerate the changes that have occurred.

2.7 Using auxiliary information in both sample design and estimation

Auxiliary information may be used both in the sample design, for instance in a stratification, and in the estimation, for example through calibration. The auxiliary information should in both cases be highly correlated with the target variable(s).

It is possible to use the same variables as auxiliary information in both stratification and calibration. As an example, consider the situation using the number of employees as stratification variable. The variable is categorised into some classes, and this categorised variable is used as a stratification variable. However, the original variable still contains information, which can be used in calibration in order to reduce non-response bias.

The choice of which variable(s) to use in the sampling and which variable(s) to use in the estimation – the same or different variables – is based on an analysis of the possible auxiliary variables and their correlations with important target variables.

Calibration means that certain estimates are consistent with known totals. This may be a further aim of the calibration; to increase coherence and consistency. It may make it easier for a user to combine several sets of statistics.

2.8 Handling organisational changes in a sample

2.8.1 General discussion

The data collection will show that the sampled enterprises differ somewhat from the frame information, and the changes will most likely increase with time. These differences must be handled. It is especially important when using an estimator with no auxiliary information and an “aging” sample. Then other methods than calibration should be used in order to overcome coverage issues. Some such situations are described below together with possible assumptions and methods. There is little literature about this type of methodology and no established practice.

If, for instance, two enterprises merge after the time of the frame information, the resulting enterprise could not have been sampled directly, but it may be sampled via either of the original enterprises. This situation may be regarded as indirect sampling, and it may be handled as such. The weight construction depends in general on the sampling method and the possible routes from the frame and sampling units to the target and collection units. This methodology is used when the unit types are deliberately different. See, for instance, Lavallé and Labelle-Blanchet (2013).

However, such estimation may become quite complex. Here, the sampling units and the target units do not differ systematically but due to certain events, which affect a limited number of units. A simplified, pragmatic, approach may then be considered and possibly chosen. This may be rational if

the causing type of event is fairly rare. For instance, with stratified random sampling, the ordinary point and variance estimators are quite simple, and it may be tempting to essentially keep these estimators. There are further options, such as weight-sharing, as just mentioned above, Lavallé and Labelle-Blanchet (2013). This may be natural, for instance when working with panels. Knottnerus (2011) studies estimators for totals and growth when panels are used, including situations where businesses change, merge, or demerge/split.

Black (2001) discusses different ways to adjust weights and handle changes in numbers of units. He notes that making changes can require a significant amount of processing and possibly create revisions that are unnecessary. With small net effects it may be better not to make changes or to postpone them.

Please observe that the descriptions below are meant as food for thought in the case of an aging sample. In a case where an updated register or frame is used for calibration, there is some but often not full information about changing units. For each unit the current situation may or may not have reached that register or frame. The term enterprise is again used here in a generic sense, out of convenience. From an estimation point of view activities are interesting, not identification numbers and other administrative information. The latter are important as such, for instance for contacts, for comparisons with later versions of registers and frames, including calibration, and also when using administrative data. Lindblom and Nordberg (2004) describe and discuss birth and death rates and also calibration.

2.8.2 Births, deaths, and re-constructions

When information about completely new large enterprises is found in media or elsewhere, then such enterprises can be put in a special stratum or group, with design weight one. This method should only apply to very large, new, enterprises. It is an enterprise that almost certainly would have been included in a “take-all-stratum” had it been known at the sampling occasion. For births of smaller enterprises, see Section 2.6 above.

When an enterprise ceases to exist, a common method is to code it as over-coverage and set its variables values to zero. Using this method in a successive set of estimates from the same sample without compensation for under-coverage means a tendency with decreasing estimates, as more and more enterprises die. Another possibility is to code and treat the dead enterprise as non-response. With simple reweighting (no calibration) this corresponds to an assumption that there is a birth rate equal to the death rate; compare the examples given in Section 2.6 above. This is a somewhat dangerous assumption, especially when there are quick movements in the economy (the business cycle with up-and-down movements in economic activity).

If an enterprise is reconstructed – in the sense that it has new owner(s) but is otherwise unaffected – it should be treated as unchanged in the estimation procedure. It is only identification number(s) and some administrative information that change. This is simply a continuity rule.

2.8.3 Mergers and splits

If two enterprises merge, it may be a reasonable assumption for the estimation that the larger enterprise has “taken over” (bought) the smaller one: say that A buys B resulting in C. With this simplifying assumption A has changed into C and B no longer exists. The estimation procedure then handles the units involved so. If A belongs to the sample, then C (the new A) is surveyed. If B belongs to the sample, then B is coded as over-coverage with value zero. Hence, the resulting unit C has the

original sampling probability of unit A (belongs to that stratum in case of a stratified random sample). This is not feasible if C is large enough to be an outlier (Section 2.9).

If, instead, an enterprise splits into two (or more) enterprises, this may be considered as a situation where both (all) enterprises are tied to the original sampling unit and its weight. In a simplified approach, as above for mergers, the new enterprises are kept in the sample and surveyed. It may be reasonable to combine the data collected from the parts into data that correspond to the original enterprise. This could be the case with simple variables, like turnover. However, there may be complicating facts with information showing, for instance, that the parts belong to different domains of estimation. The influence on estimates has to be considered. With fairly large enterprises some tailor-made solution is needed.

See also the handbook module “Weighting and Estimation – Main Module”, which recommends a strict approach with indirect sampling and weight-sharing, and Lavallé and Labelle-Blanchet (2013).

2.9 *Outliers in the sample*

An outlier is a value that deviates considerably from the “bulk” of observations. It may be due to the skewed distribution, to the current size of the unit differing from the frame information, or to an erroneous value. Especially the last category should be handled and eliminated in the statistical editing. The large enterprises, which are included with probability one, are considered on their own; compare Section 2.5 above. The remaining outliers are usually treated as representative. They may still be too influential in a simple estimation procedure, depending on their weights. There are methods to handle such outliers, for instance by winsorisation of the value or by weight modification. Normally, the variance is decreased by using such methods, but a bias is introduced at the same time. The handbook module “Weighting and Estimation – Main Module” gives an overview of methods, and the method module “Weighting and Estimation – Outlier Treatment” provides more detailed information. There is a recent article by Beaumont, Haziza, and Ruiz-Gazen (2013).

The design should (i) try to prevent or at least reduce the occurrence of outliers and (ii) choose an appropriate method for handling the outliers that still occur. An appropriate and up-to-date (as far as possible) measure of size is essential for the first aim and well worth a study during the design.

For some variables, which are strongly skewed, there will be outliers due to their nature, like investment in buildings, which for a small enterprise may be high at rare occasions. The estimation procedure should foresee such outliers in the design stage and choose an appropriate method to handle them. Especially in a repeated survey there is information about outliers to first collect and then utilise in later production rounds. Another possibility is to use information from annual surveys to make rough estimates for short-term surveys about occurrences and then choose appropriate method(s). There may in both situations be differences, though, between different parts of a business cycle (up-and-down movements in economic activity). Some care may be needed with different levels of detail, for instance for domains of estimation. Higher levels of aggregates are less sensitive than lower levels.

The choice of method has to consider (i) the information needed and available for the estimator being robust and also (ii) the complexity which is added in comparison with the “basic” estimator that would otherwise be used, if there were no outliers.

Some surveys use a technique with a “surprise-stratum”, which means that they put found outliers there with weight one. If the outlier really can be considered unique, this handling is reasonable.

Otherwise – when the outlier is representative – such a procedure introduces a bias, and it is not recommended in comparison with the methods described.

The two previously used examples from ONS are different. ONS (2013, p. 6) is a survey using winsorisation. There scaling is used to achieve consistency when one winsorised variable is the sum of two others. ONS (2012, pp. 42–44) describes how outliers are first identified and then treated as non-representative, using also a post-stratification. There are also model-based methods in use; see the handbook module “Weighting and Estimation – Outlier Treatment” for a description and references.

2.10 Cut-off sampling

As indicated in the handbook module “Sample Selection – Main Module”, it may be motivated not to include a certain part of the population in the sample. Benedetti, Bee, and Espa (2010) provide a framework for such sampling and estimation. They show, for instance, how the estimator can be constructed for the target population, which includes the unobserved population below the cut-off threshold. A model needs to be used, such as an assumption that the share below the threshold of the total of a survey variable is the same as for an auxiliary variable known, for instance, from the frame.

In this case the estimation design determines important parameters, such as the threshold value, the auxiliary variable, and the model assumption to be used. Again, the choice of the size variable is important to get a reasonable model and estimator. Later on, when the BR has been updated, it may be possible to assess the estimator, at least partly.

There is some similarity between this estimation procedure and the use of administrative data for small enterprises, see Section 2.13.

2.11 Models: early estimates

A preliminary (early) estimate may be required already at a time when the response rate normally is fairly low. This affects not only the variance (a smaller number of respondents), but also the potential bias, due to the early respondents possibly being different from later ones. The handbook module “Weighting and Estimation – Main Module” describes three different possibilities: to take a random sub-sample with a short response time, to use a design-based estimator based on early respondents, and to use a model-based estimator. The estimators can be compared in terms of assessed possible biases and estimated variances, and – perhaps most important – revision sizes. Again, there may be differences between different parts of a business cycle with its up-and-down movements.

The outcomes of comparisons either when planning (if possible and meaningful) or later – when revisions can be computed – are useful ingredients when choosing an appropriate estimator. If it is a repeated survey, information should be collected, analysed, and used for improvements of the design.

Similar situations may occur when using administrative data, see Section 2.13.

2.12 Models: small area estimation

If there are small domains of estimation, design-based estimators will have a large variance, so model-based estimators are often preferred or even necessary. This topic is described in rather much detail in the handbook module “Weighting and Estimation – Main Module” with references to further modules. Hence, there is no specific discussion here. It is important to note, though, that statistical disclosure control may come into play, see handbook module “Statistical Disclosure Control – Main Module”.

2.13 Use of administrative and other accessible data

The handbook module “Weighting and Estimation – Main Module” has a short section on integration of administrative data, and there are special handbook modules on administrative and secondary data in the topics “Data Collection”, “Statistical Data Editing”, and “Weighting and Estimation”. Especially “Weighting and Estimation – Estimation with Administrative Data” is relevant here.

There are several important issues to consider in the design when considering the possible use of administrative data:

- The administrative and the statistical units.
- The coverage of the administrative source(s) in comparison with the target population.
- Variable definitions and how to derive or model the target variables from the administrative variables.
- Measurement errors and other non-sampling errors (deficiencies).
- Timeliness, which may depend on the size of the unit. (Small legal units may not report frequently to tax authorities, for instance.)

It may be a good solution to use direct data collection for large enterprises and administrative data for medium-sized and small enterprises. This eliminates the response burden for the latter group. There may be problems, though, with some variables and with timeliness. Models need to be introduced, studied, and chosen with respect to quality aspects such as accuracy, coherence, and timeliness.

There is a large set of deliverables from an ESSnet project, see ESSnet on AdminData (2013) with different aims, such as data quality and estimation in different situations, depending on variables and timeliness, for instance. Lewis (2012), Paavilainen (2012), and Brinkley, Preston, and Scott (2012) are three different examples of the use of administrative data in surveys, showing also difficulties in practice and how model assumptions can be included. The two first examples are from the ESSnet project. Lewis (2012) illustrates how to find solutions when variables are not directly available from administrative sources. Paavilainen (2012) describe that administrative data may change, that some of them may be late for short-term statistics with short production time, and how an index can be constructed with mixed sources. Brinkley et al. (2012) is a broad description of the inclusion of administrative data with emphasis on sample design and estimation.

2.14 Evaluation of the design

The design of the estimation procedure means work with requests on quality, in particular accuracy. At the end of the production round the achieved quality (for instance accuracy) should be studied and compared with the planned quality (including accuracy). If there are differences, the causes should be analysed and possible actions should be considered. This is particularly the case in a repeated survey, for possible improvements in later production rounds. Experiences may also be shared across surveys.

If changes of the design are motivated, not only the estimator should be considered. It may be easier and better to modify the allocation of the sample, just as a simple example. There may be many further issues to consider, for instance how to handle non-sampling errors in the estimation and in other parts of the production process. See, for example, the handbook modules “Repeated Surveys – Repeated Surveys” and “Overall Design – Overall Design”.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Assoulin, D. (2009), Choosing an Imputation Method for Large Firms. European Establishment Statistics Workshop, Efficient Methodology for Producing High Quality Establishment Statistics (Stockholm, Sweden).

Baker, R., Brick, M., Bates, N., Battaglia, M., Couper, M., Dever, J., Gile, K., and Tourangeau, R. (2013), Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology* **1**, 90–105 (and Rejoinder 137–143).

Beaumont, J.-F., Haziza, D., and Ruiz-Gazen, A. (2013), A unified approach to robust estimation in finite population sampling. *Biometrika* **100**, 555–569.

Benedetti, R., Bee, M., and Espa, G. (2010), A Framework for Cut-off Sampling in Business Survey Design. *Journal of Official Statistics* **26**, 651–671.

Black, J. (2001), Changes in Sampling Units in Surveys of Businesses. Federal Committee on Statistical Methodology (FCSM), 2001 FCSM Conference Papers.

BLS (2013), Technical Notes for the Current Employment Statistics Survey. US Bureau of Labor Statistics. (Link from 24-01-2014: <http://www.bls.gov/ces/cestn.pdf>.)

Blue-Ets-project (2013), *Best practice recommendations on variance estimation and small area estimation in business surveys*. Deliverable 6.2. BLUE-Enterprise and Trade Statistics. (Link from 24-01-2014: <http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable6.2.pdf>.)

Brinkley, E., Preston, J., and Scott, A. (2012), Using Administrative Taxation Data to Improve Sample Design and Estimation - An ABS Perspective. Paper presented at the International Conference on Establishments, ICES-IV.

Choudhry, G. H., Rao, J. N. K., and Hidirolou, M. A. (2012), On sample allocation for efficient domain estimation. *Survey Methodology* **38**, 23–29.

ESSnet on AdminData (2013), *Use of Administrative and Accounts Data for Business Statistics*. Project in three time periods (SGAs 1-3). Deliverables are accessible. (Link from 24-01-2014: <http://www.cros-portal.eu/content/use-administrative-and-accounts-data-business-statistics>.)

- Knottnerus, P. (2011), *Panels – Business Panels*. Statistical Methods (201119), Statistics Netherlands, The Hague/Heerlen.
- Lavallé, P. and Labelle-Blanchet, S. (2013), Indirect sampling applied to skewed populations. *Survey Methodology* **39**, 183–215.
- Lewis, D. (2012), Methods for using administrative data to estimate survey variables not directly available from administrative sources. Paper presented at the International Conference on Establishments, ICES-IV.
- Lindblom, A. and Nordberg, L. (2004), On adjustment for coverage problems in short-term business surveys. Paper contributed to the European Conference on Quality and Methodology in Official Statistics (Q 2004).
- ONS (2012), *Annual Business Survey (ABS): Technical Report*. Issued by Office for National Statistics, August 2012.
- ONS (2013), *Quality and Methodology Information on Business Register and Employment Survey (BRES)*. Information paper. Issued by Office for National Statistics, 20th November 2013.
- Paavilainen, P. (2012), Efficient use of administrative data in the production of economic statistics in Finland. Paper presented at the International Conference on Establishments, ICES-IV.
- Rivière, P. (2002), What Makes Business Statistics Special? *International Statistical Review* **70**, 145–159.
- Valliant, R. (2013), Comment (on Baker et al., 2013). *Journal of Survey Statistics and Methodology* **1**, 105–111.
- Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* **66**, 41–63.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Different Types of Surveys
2. Overall Design – Overall Design
3. Repeated Surveys – Repeated Surveys
4. Statistical Registers and Frames – Survey Frames for Business Surveys
5. Sample Selection – Main Module
6. Sample Selection – Sample Co-ordination
7. Data Collection – Main Module
8. Data Collection – Design of Data Collection Part 2: Contact Strategies
9. Data Collection – Mixed Mode Data Collection
10. Data Collection – Collection and Use of Secondary Data
11. Statistical Data Editing – Main Module
12. Imputation – Main Module
13. Weighting and Estimation – Main Module
14. Weighting and Estimation – Estimation with Administrative Data
15. Statistical Disclosure Control – Main Module

9. Methods explicitly referred to in this module

1. Statistical Data Editing – Manual Editing
2. Weighting and Estimation – Outlier Treatment

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

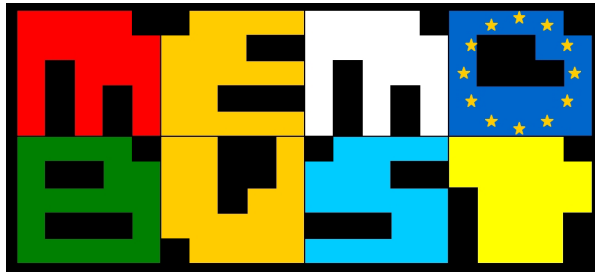
Weighting and Estimation-T-Design of Estimation

15. Version history

Version	Date	Description of changes	Author	Institute
0.0.1	16-12-2013	first plan	Marianne Ängsved and Eva Elvers	Statistics Sweden
0.0.2	03-01-2014	some sections drafted	Eva Elvers	Statistics Sweden
0.0.3	06-01-2014	further sections	Marianne Ängsved	Statistics Sweden
0.0.4	08-01-2014	merging, expanding	MÄ+EE	Statistics Sweden
0.0.5	10-01-2014	further additions	EE+MÄ	Statistics Sweden
0.0.6	25-01-2014	reviews from HU, CH, IT	EE+MÄ	Statistics Sweden
0.0.7	28-01-2014	continued improvements	MÄ+EE	Statistics Sweden
0.1.0	07-02-2014	reviews EB, IT	EE+MÄ	Statistics Sweden
0.1.1	12-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:31



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Calibration

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	7
4. Examples – not tool specific.....	7
5. Examples – tool specific.....	7
6. Glossary.....	9
7. References	9
Specific section.....	12
Interconnections with other modules.....	17
Administrative section.....	20

General section

1. Summary

Weighting is a statistical technique commonly used and applied in practice to compensate for nonresponse and coverage error. It is also used to make weighted sample estimates conform to known population external totals. In recent years a lot of theoretical work has been done in the area of weighting and there has been a rise in the use of these methods in many statistical surveys conducted by National Statistical Offices around the world. This module describes in detail calibration as a method of adjusting initial weights in surveys based on sampling in order to estimate known population totals of all auxiliary variables perfectly. This method can also be used in surveys as a possible solution for treatment of unit nonresponse and enables gain on efficiency in term of variance when strong correlation between the variable of interest and auxiliary variables exists. It is worth noting that this is one of many weighting methods which can be used in practice. Others include weighting, poststratification, raking, GREG weighting, logistic regression weighting, mixture approach and logit weighting. A review of the weighting method with examples can be found in Kalton and Flores-Cervantes (2003). More information can also be found in “Weighting and Estimation –Main Module”.

Calibration estimation, whereby sampling weights are adjusted to reproduce known population totals, is commonly used in survey sampling. The milestone was the article by Deville and Särndal (1992) in which calibration was described in details. Calibration can be treated as an important methodological instrument, especially in large-scale production of statistics. Many national statistical agencies have developed software designed to compute final weights, usually calibrated using auxiliary information available in administrative registers, censuses and other accurate sources. Calibration as a method of weighting has been described in detail in many articles. A full definition of calibration approach was formulated by Särndal (2007). According to Särndal, the calibration approach to estimation for finite populations consists of:

- (a) the computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s);
- (b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units;
- (c) satisfying an objective of obtaining nearly design unbiased estimates given that nonresponse and other non-sampling errors are absent.

2. General description of the method

We will assume that we are interested in computing the total value of variable Y (see formula 1). Let us assume that the whole population $U = \{1, \dots, N\}$ consists of N elements. From this population we draw, according to a certain sampling scheme, a sample $s \subseteq U$, which consists of n elements. Let π_i denote first order inclusion probability, i.e., $\pi_i = P(i \in s)$ and $d_i = \frac{1}{\pi_i}$ the design weight. Let $\pi_{ij} = P(i, j \in s)$ denote the second-order inclusion probability. We assume that our main goal is to estimate the total value of variable y :

$$Y = \sum_{i=1}^N y_i, \quad (1)$$

where y_i denotes the value of variable y for i -th unit, $i = 1, \dots, N$.

One well known, classical estimator of the total value (1) is the Horvitz-Thompson estimator, which is given by the formula:

$$\hat{Y}_{HT} = \sum_s d_i y_i = \sum_{i=1}^n d_i y_i. \quad (2)$$

If, in addition to y_i , auxiliary variables x_1, \dots, x_k are available from the sample and the population totals $\mathbf{X}_j = \sum_{i=1}^N x_{ij}$, $j = 1, \dots, k$ are known, it may occur that:

$$\sum_s d_i x_{ij} = \sum_{i=1}^n d_i x_{ij} \neq \mathbf{X}_j \quad (3)$$

where x_{ij} denotes the value of j -th auxiliary variable for the i -th unit.

Let \mathbf{X} denote the known vector of population totals for the vector of auxiliary variables:

$$\mathbf{X} = \left(\sum_{i=1}^N x_{i1}, \sum_{i=1}^N x_{i2}, \dots, \sum_{i=1}^N x_{ik} \right)^T. \quad (4)$$

This vector is often called the vector of calibration totals or calibration benchmarks.

Let w_i denote calibration weight $i = 1, \dots, n$. Our main goal is to look for new weights w_i which are as close as possible to the design weights d_i and satisfy

$$\mathbf{X} = \tilde{\mathbf{X}} \quad (5)$$

where

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^n w_i x_{i1}, \sum_{i=1}^n w_i x_{i2}, \dots, \sum_{i=1}^n w_i x_{ik} \right)^T. \quad (6)$$

The calibration estimator for totals (1) takes the form

$$\hat{Y}_{cal} = \sum_{i=1}^n w_i y_i, \quad (7)$$

and weights w_i fulfill the so called calibration equation given by formula (5).

The process of constructing calibration weights depends on the properly chosen so called distance function G , which measures the difference between initial weights d_i and final weights w_i . This function must satisfy the following regularity conditions:

- $G(\cdot)$ is strictly convex and twice continuously differentiable,
- $G(\cdot) \geq 0$,
- $G(1) = 0$,
- $G'(1) = 0$,
- $G''(1) = 1$.

The calibration problem involves searching for new weights for a given sample s which are as close as possible to the initial weights and satisfy calibration equations and possibly the boundary constraints. This problem can be formulated as a non-linear optimisation problem, see Vanderhoeft (2001):

(C1) Minimise the distance:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^n d_i G\left(\frac{w_i}{d_i}\right) \rightarrow \min \quad (8)$$

(C2) subject to k calibration equations:

$$\sum_{i=1}^n w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \dots, k, \quad (9)$$

(C3) subject to boundary constraints:

$$L \leq \frac{w_i}{d_i} \leq U, \text{ where } 0 \leq L \leq 1 \leq U, \quad i = 1, \dots, n. \quad (10)$$

The first constraint (C1) says that calibration weights w_i should be as close as possible to initial weights d_i in terms of distance function G , which measures the difference between both weights. It means that the ratio between final weights and initial weights should not be very different from one. In a special situation, where $w_i = d_i$, no correction is required. The second constraint (C2) is fundamental and constitutes the essence of the calibration approach. According to this constraint, calibration weights must perfectly estimate the totals of all auxiliary variables taken into account in the calibration procedure. This means that the totals of all auxiliary variables are estimated with zero variance using calibration weights. The third constraint (C3) is optional and it may be added whenever calibration weights are negative or extreme. In such a situation, the ratio between final and initial weights should be limited to a carefully specified range.

There is also some freedom in choosing the function G , i.e., this function can be chosen conveniently. The following functions are the most commonly used in practice

$$G_1(x) = \frac{1}{2}(x - 1)^2, \quad (11)$$

$$G_2(x) = \frac{(x-1)^2}{x}, \quad (12)$$

$$G_3(x) = x(\log x - 1) + 1, \quad (13)$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \quad (14)$$

$$G_5(x) = \frac{1}{2\alpha} \int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt, \quad (15)$$

where α is a positive parameter, which is used to control the degree of dispersion of calibrated weights in relation to initial weights and \sinh denotes the hyperbolic sinus function.

In many statistical packages the problem of finding calibration weights is implemented using different G functions. For example, in CALMAR, which is a macro written in 4GL in SAS four distance functions were implemented, i.e.:

- the linear method, which is based on formula (11),
- the raking ratio method, which is based on the distance function given by (13),
- the logit method, which provides lower limits L and upper limits U on the weight ratios w_i/d_i . In this case, the G function can be expressed as follows:

$$G(x) = \left[(x - L) \log \frac{x-L}{1-L} + (U - x) \log \frac{U-x}{U-1}\right] \frac{1}{A}, \quad (16)$$

where

$$A = \frac{U-L}{(1-L)(U-1)}, \quad (17)$$

- the truncated linear method, which is based on formula (11), but constraints on the weight ratios w_i/d_i are imposed, i.e., $L \leq \frac{w_i}{d_i} \leq U$.

In CALMAR 2, which is a later version of CALMAR, the distance function (15) is also implemented. The method expressed by the formula (16) and the truncated linear method are used to control the range of weight ratios. They are used when negative or large weights occur, which may happen when the linear method is taken into account.

The linear method is often used in practice because negative or extreme weights do not occur. This is also the fastest procedure, because it does not need an iterative approach to the problem of finding calibration weights. It can be proved that estimators based on this method are equal to *generalised regression* (GREG) *estimators* (Deville and Särndal 1992). More information about GREG estimators can also be found in Cassel, Särndal and Wretman (1976) and in the module “Weighting and Estimation – Generalised Regression Estimator”.

Let us assume that the distance function is expressed by the formula (11). In this situation we have:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^n d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^n d_i \frac{1}{2} \left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2} \sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i}. \quad (18)$$

This kind of formula allows us to find calibration weights in an explicit form. We can prove that if the matrix $\sum_{i=1}^n d_i \mathbf{x}_i \mathbf{x}_i^T$ is nonsingular, then the solution of the minimisation problem (8), subject to the calibration constraint (9) is a vector of calibration weights $\mathbf{w} = (w_1, \dots, w_n)^T$, whose elements are described by the formula:

$$w_i = d_i + d_i (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum_{i=1}^n d_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i, \quad (19)$$

where

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^n d_i x_{i1}, \sum_{i=1}^n d_i x_{i2}, \dots, \sum_{i=1}^n d_i x_{ik} \right)^T, \quad (20)$$

and

$$\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^T, \quad (21)$$

is the vector consisting of values of all auxiliary variables for the i -th unit in the sample $i = 1, \dots, n$.

All of calibrated estimators \hat{Y}_{cal} have the same asymptotical precision, regardless of the distance function G used. It was proven that the family of calibration estimators \hat{Y}_{cal} is asymptotically equivalent to the GREG-estimator (see Deville and Särndal, 1992). From this point of view, the variance of any calibration estimator \hat{Y}_{cal} can be estimated using the following formula for estimating the variance of the GREG estimator (see Deville and Särndal, 1992):

$$\hat{V}(\hat{Y}_{cal}) = \sum_{i \in s} \sum_{j \in s} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}} \right) (w_i e_i) (w_j e_j) \quad (22)$$

where e_i are residuals, which are calculated from a sample using weighted linear regression of y on calibration variables x_1, \dots, x_k , i.e.,

$$e_i = y_i - \mathbf{x}_i' \mathbf{B}_s, \quad (23)$$

$$\mathbf{B}_s = (\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i')^{-1} (\sum_{i \in s} w_i \mathbf{x}_i y_i). \quad (24)$$

For any distance function, as stated above, the variance is similar to that of the generalised regression estimator. This variance is given by residuals of a regression of the target variable y on auxiliary variables x_1, \dots, x_k . If the variable of interest is strongly correlated with the auxiliary variables the gain on precision will be noticeable.

3. Preparatory phase

4. Examples – not tool specific

Examples of how to use calibration can be found in a paper written by McCormack and Sautory (2003). Examples relate to the CALMAR/CALMAR2 macro written in 4GL in SAS language. A function with examples for applying calibration in the R environment can be found in Lumley (2012) and in SPSS software in Vanderhoeft (2001, 2002).

We refer to Wallgren and Wallgren (2007) for examples of applying calibration estimators in register-based statistics. In this book the method of determining calibration weights is presented step by step using operations on matrices.

One example of using calibration with the CALMAR2 macro to determine final weights was also described in detail in a section on tools in this module.

5. Examples – tool specific

Presented below is a detailed description of how to use the CALMAR2 macro to determine calibration weights.

We consider an artificial population of enterprises of size $N=1000$ from which a simple random sample of size $n=20$ is drawn. Hence design (initial) weights are equal, $N/n=1000/20=50$. We also consider a numerical variable x_1 (for instance, monthly revenue of enterprise) and one categorical variable x_2 (for instance, enterprise size, i.e., large - L and medium - M). In this example it will only be shown how calibration weights should be computed. We do not take into account the variable of interest y which is not necessary to compute calibration weights and would be necessary to calculate the variance of the estimator. Monthly revenue of enterprise and enterprise size are chosen as auxiliary variables.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Enterprise size	M	M	M	L	L	L	M	M	L	M	M	M	L	M	M	M	M	M	L	M
Monthly revenue	18	14	16	35	30	10	15	23	23	12	18	16	22	15	15	10	18	18	35	16

Source: artificial data set

The weighted sum of variable x_1 is equal to 18950. Number of medium and large enterprises according to this survey is equal to 700 (14 medium enterprises times 50) and 300 (6 large enterprises times 50), respectively. The exact population total of monthly revenue is known and equals to 19000 and the real number of medium and large enterprises are equal to 720 and 280, respectively. We would like to calibrate the design weights in such a way that known auxiliary totals will be reproduced. In other words, we would like to slightly modify the initial weights so that the sum of x_1 based on the new weights is equal to 19000 and weighted sum of medium and large enterprises is equal to 720 and 280, respectively. We will use the CALMAR2 code in SAS to solve this problem. The SAS code for creating the preliminary datasets and recalling the macro CALMAR2 command is given below.

```

/*****Library containing CALMAR*****/
libname calm 'D:\Calibration';
optionsmstoredsasmstore=calm;

/*****Creation of input dataset with drawn units*****/
data sample;
input enterprise $ size $ revenue weight;
cards;
ent01 M      18      50
ent02 M      14      50
ent03 M      16      50
ent04 L      35      50
ent05 L      30      50
ent06 L      10      50
ent07 M      15      50
ent08 M      23      50
ent09 L      23      50
ent10 M      12      50
ent11 M      18      50
ent12 M      16      50
ent13 L      22      50
ent14 M      15      50
ent15 M      15      50
ent16 M      10      50
ent17 M      18      50
ent18 M      18      50
ent19 L      35      50
ent20 M      16      50
;
run;

/*****Creation dataset with known population totals*****/
data totals;
inputvar $ n mar1 mar2;
cards;
size 2 280 720
revenue 0 19000 .
;
run;

/*****Call to CALMAR*****/

```

```
%CALMAR2(DATAMEN=sample, POIDS=weight, IDENT=enterprise,
MARMEN=totals, M=1,DATAPOI=wcal, POIDSFIN=cal_weights )

/*****Printing final result*****/
procprint data=wcalnoobs;
run;
```

The following dataset, with the final weights, is printed:

enterprise	cal_weights	enterprise	cal_weights
ent01	52.2750	ent11	52.2750
ent02	50.5821	ent12	51.4286
ent03	51.4286	ent13	45.0443
ent04	50.5462	ent14	51.0054
ent05	48.4301	ent15	51.0054
ent06	39.9657	ent16	48.8893
ent07	51.0054	ent17	52.2750
ent08	54.3911	ent18	52.2750
ent09	45.4675	ent19	50.5462
ent10	49.7357	ent20	51.4286

CALMAR2 changed the design weights so that the weighted total of variable x_1 is equal to 19000 and weighted number of medium and large enterprises is equal to 720 and 280, respectively. In this example a linear method was used ($M=1$; 1 – linear, 2 – raking ratio, 3 – logit, 4 – truncated linear, 5 – sinus hyperbolic). The macro parameter DATAMEN contains information about input dataset, POIDS contains information about design weights, IDENT contains the name of an identifying variable for the units in the sample dataset, MARMEN stores information about known totals of all auxiliary variables, M is the identifier of the calibration method that was used, DATAPOI is the name of a new dataset which will be created and will contain calibration weights.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Andersson, C. and Nordberg, L. (2000), *CLAN – A SAS Program for Computation of Point and Standard Error Estimates in Sample Surveys*. Statistics Sweden.
- Casciano, M.C., Giorgi, V., Oropallo, F., and Siesto, G. (2012), Estimation of Structural Business Statistics for Small Firms by Using Administrative Data. *Rivista Di Statistica Ufficiale*, N. 2-3.
- Cassel, C.M., Särndal, C-E., and Wretman, J. (1976), Some Results on Generalized difference estimation and Generalized Regression Estimation for Finite Populations. *Biometrika* **63**, 615–620.

- Cassel, C., Lundquist, P., and Selén, J. (2002), *Model-based calibration for survey estimation, with an example from expenditure analysis*. R&D Report, Research-Methods-Development, Statistics Sweden.
- Central Statistical Office in Poland (2011), *Incomes and Living Conditions of the Population in Poland (report from the EU-SILC survey of 2009)*. Statistical Information and Elaborations, Warsaw 2011.
- Deville, J.-C. and Särndal, C.-E. (1992), Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* **87**, 376–382.
- Duchesne, P. (1999), Robust calibration estimators. *Survey Methodology* **25**, 43–56.
- Éltető, Ö. and Mihályffi, L. (2002), Household Surveys in Hungary. *Statistics in Transition* **5**, 521–540.
- Estevao, V., Hidirolou, M.A., and Särndal, C.-E. (1995), Methodological Principles for a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics* **11**, 181–204.
- Eurostat (2004), *Description of target variables: Cross-sectional and Longitudinal*. EU-SILC 065/04.
- Gambino, J. (1999), *Discussion of Issues in weighting household and business surveys*. Statistics Canada, Household Survey Methods Division.
- Kalton, G. and Flores-Cervantes, I. (2003), Weighting methods. *Journal of Official Statistics* **19**, 81–97.
- Lumley, T. (2012), *Package survey*. Version 3.28 of this package is available at the website <http://cran.r-project.org/web/packages/survey/survey.pdf>.
- McCormack, K. (2011), *The calibration software CALMAR – What is it?* Central Statistics Office Ireland.
- MEETS (2011), *Use of Administrative Data for Business Statistics*. Final Report, Poznań (Poland).
- Nieuwenbroek, N. and Boonstra, H. (2001), *Bascula 4.0 Reference Manual*. Technical paper 3554-99-RSM, Statistics Netherlands.
- Sautory, O. (2003), Calmar 2: A New Version of the Calmar Calibration Adjustment Program. Program of Statistics Canada's Symposium 2003, Challenges in Survey Taking for the Next Decade.
- Särndal, C.-E. and Estevao, V. M. (2000), A Functional Form Approach to Calibration. *Journal of Official Statistics* **16**, 379–399.
- Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley & Sons.
- Särndal, C.-E. (2007), The Calibration Approach in Survey Theory and Practice. *Survey Methodology* **33**, 99–119.
- Vanderhoeft, C. (2001), *Generalised Calibration at Statistics Belgium. SPSS Module g-CALIB-S and Current Practises*. Statistics Belgium.
- Vanderhoeft, C. (2002), *g-Calib Generalised Calibration Under SPSS*. Statistics Belgium.
- Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley.

Zhang, L.-C. (1998), Post-Stratification and Calibration – A Synthesis. Statistics Norway, Research Department.

Specific section

8. Purpose of the method

The main purpose of the method is to adjust, using auxiliary variables, initial weights and construct final weights (calibration weights), which estimate perfectly the totals of all auxiliary variables taken into account in the calibration process in such a way that the final weights are as close as possible to the initial weights in terms of the distance function used. One of the main reasons why calibration should be used in survey sampling is the efficiency of estimates, which can be achieved by exploiting external information and can lead to a small variance of estimators which are based on calibration weights. As a result of calibration, potential improvements in the precision of estimates can be expected. Other reasons for using calibration and purposes of this method include, see Gambino (1999):

- balance, which can be understood to mean that following calibration, the sample “looks” like the population,
- consistency of estimates – after calibration each unit of the sample has a unique final weight, which ensures consistency in the sense that when weights are applied to auxiliary variables, they will conform to (will be consistent with) known aggregates for the same auxiliary variables, i.e., weighted parts will add up to totals and mutual consistency between estimated tables will be guaranteed,
- convenience and transparency – this is a particularly important purpose of calibration from the user’s point of view, since the resulting estimates are easy to interpret and calibration based on known totals is natural and leads to slightly modified design weights, which can reproduce in a transparent way known benchmarks,

and, see Deville and Särndal (1992), Särndal and Lundström (2005), Särndal (2007):

- potential reduction in bias in the presence of nonresponse and coverage error,
- potential improvements to the precision of estimates,
- coherent estimates based on data coming from different sources.

9. Recommended use of the method

1. Missing data are one of the major types of non-random errors in statistical surveys. They produce significantly biased results and can considerably affect the survey quality. As a rule, this problem is evident in all kinds of surveys conducted by statistical offices of many countries where the lack of response to certain survey questions is quite normal, although definitely undesirable from the point of view of estimation. In view of the above, recent years have seen a growing interest in various methods, which are designed to offset the negative effect of missing data. One of these methods is calibration, which is successfully used by statistical offices of many countries and recommended in many articles and books as a method to handle unit nonresponse. For details on how to use calibration as a method of estimation in surveys with missing data, see Särndal and Lundström (2005).

2. The calibration approach should be recommended and taken into account in all surveys based on sampling, because it can help to reduce bias due to unit nonresponse and variance of estimators. When auxiliary data are strongly correlated with variables of interest, calibration can allow an important gain in precision.
3. In many practical situations, especially involving economic surveys, the distribution of target variables is often asymmetric and some units might have extreme values compared to others (outliers). From one point of view a complete elimination of such units could lead to biased estimates. On the other hand, retaining them with their original weight could make the estimators used highly variable. Duchesne (1999) proposes robust calibration estimators in the case of outliers. This approach is an extension obtained by Deville and Särndal (1992) for the class of calibration estimators based on quantile regression technique, which are discussed in detail in this module. The approach could be extremely useful in business surveys, where distribution of variables, such as income or revenue is highly asymmetric. A broad discussion of the problem of outliers and their negative impact on final results can also be found in the module “Weighting and Estimation – Outlier Treatment”.

10. Possible disadvantages of the method

1. For some distance functions it is possible to receive quite large or negative calibration weights, which is very undesirable in terms of estimation. Such cases should be avoided, i.e., weights have to be positive and should lie within specific desirable limits in order to be as close as possible to original design weights. In any case, it is possible to fulfil this requirement by taking into account an appropriate chosen distance function which can exclude negative or large calibration weights while satisfying given calibration equations. For example, the function given by the formula (16) or (11) with constraints on the weight ratios can be a good remedy when large or negative weights occur.
2. When using the distance function, which helps to restrict the range of weights, it should be remembered that as a result of imposing too strong restrictions on calibration weights with respect to initial weights, the algorithm of finding adjusted weights may not converge.
3. The presence of outlying values in the auxiliary variables may produce extreme calibration weights, which differ a lot from original design weights. In such a situation calibration estimators can be highly variable.
4. In the presence of weak auxiliary information calibration may fail and lead to abnormally high or low weights and, as a consequence, can adversely affect the estimation process.
5. In the presence of some categorical auxiliary variables complete cross-classification may lead to small cells and, as a result, abnormal weights are possible.

11. Variants of the method

1. Variants of the method depend on the chosen distance function. All the calibrated estimators are asymptotically equivalent to the calibrated estimator obtained with the linear method. For more details, see Deville and Särndal (1992).

2. Final results depend on the availability and the choice of efficient auxiliary variables which, according to Särndal and Lundström (2005), should explain the response probability, the main study variable and identify the most important domains. If not, calibration may not be effective and may not bring any improvement or give inefficient or implausible estimators.

12. Input data

1. The input data generally corresponds to the information which is available in the sample and the margins known on the level of the population on which calibration will be done. The input data set usually contains some tables. For example, in CALMAR2, which is a macro written in 4GL in SAS, a table with sample data is required. This table should contain some important variables, e.g., initial weights for units in the sample, an identifying variable, values of the auxiliary variables. Another table should contain information with auxiliary variables, their names, the number of categories and associated margins.

13. Logical preconditions

1. Missing values
 1. When one wants to find calibration weights for a domain which is empty in the sample, it is impossible to create new adjusted weights or any linear estimator of the weighted form $\sum_{i=1}^n w_i y_i$. However, this can be done using over-weighting methods, e.g., the raking approach. When the problem of nonresponse concerns only some units in the domain, it is possible to apply calibration as a method of reducing bias and high variance of estimators. It can lead to reliable estimation provided that auxiliary information is used efficiently. For details, see Särndal and Lundström (2005).
2. Erroneous values
 1. Standard calibration methods do not take into consideration errors in variables. Possible misspecification of variables or corrections of variables are generally not taken into account. However, it is possible to construct robust calibration estimators, which can be very helpful in the presence of outliers and highly asymmetric distributions of variables under study. It can be especially important in business statistics, since in such surveys distributions are affected by extreme or erroneous values, e.g., monthly income of enterprises. For details, see Duchesne (1999).
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. When the sample is small, the linear approach to calibration may produce negative weights, which is undesirable; instead, restricted calibration methods based on iterative algorithms should be applied. For example, the function given by the formula (11) with additional constraints on weight ratios (lower and upper bounds) requires an iterative procedure of determining final calibration weights.

14. Tuning parameters

1. Parameters for the convergence of iterative methods used in the context of the calibration approach are: the maximum number of iterations, convergence criterion, choice of the method (distance function), choice of the lower and upper limits of the calibrated weights. Details of tuning parameters and how to establish them are described in detail in many publications; see, e.g., Lumley (2012), Nieuwenbroek and Boonstra (2001), Sautory (2003), Vanderhoeft (2002) and Zhang (1998).

15. Recommended use of the individual variants of the method

1. Since the linear method provides asymptotically the common linear approximation to all calibration estimators, in many cases it would be the best solution, because it does not need any iterative procedures and in this respect is the fastest one. Another reason why the linear method should be used first is the fact that in many surveys calibrated results often differ fairly little from one method to another, see Zhang (1998). It should be also underlined that other methods of calibration are widely used in practice (for instance, raking ratio) and give good results. Anyway when negative or extreme weights occur, other distance functions, which need iterative algorithms should be considered. In such cases, special attention should be paid to the choice of lower and upper limits of calibrated weights. Restricting the range of weights too much may prevent the algorithm of the calibration procedure from converging.

16. Output data

1. An output dataset depends on the program used and usually contains table(s) with the following information: number of iterations, number of negative weights after each iteration, termination criterion, information about the comparison between margins estimated from the sample (initial weights), using calibration weights and real margins in the population, a set of final (calibration) weights, information about the method used, coefficients of vector lambda of Lagrange multipliers after each iteration, ratios of weights (final weights/initial weights), statistics for ratios of weights, histograms with the distribution of initial and final weights, tables of estimates including estimates of standard errors.

17. Properties of the output data

1. The final output usually contains some tables written to separate files in the format compatible with input data sets (e.g., a file with calibrated weights) and information about the whole process of calibration written and exported to an appropriate file, e.g., pdf or html format. In this output one can find information about properties of calibration weights (number of iterations, number of negative weights, etc.). The user should check in detail the quality of estimates based on calibration weights and their knowledge of the investigated phenomenon, standard errors and bias of estimates.

18. Unit of input data suitable for the method

In order to compute calibration weights, information about initial weights and auxiliary variables should be available for all units in the sample (sample level). Unit level data are also necessary to compute variance estimation of the calibration estimator (input for the method).

19. User interaction - not tool specific

1. Select method of calibration (distance function). In the approach which needs limits for calibration weights establish the lower and upper limit of the range for ratio of initial and calibration weights.
2. Choose carefully potential auxiliary variables to be included into the calibration process.
3. Choose the right software and program to perform the process of calibration.
4. Establish tuning parameters (e.g., convergence criteria, number of iterations).
5. After the use of calibration, quality indicators should be checked and verified in order to evaluate the final results (existing negative or extreme weights, distribution of initial and final weights, correlation coefficient between initial and final weights, ratio of initial and final weights).

20. Logging indicators

1. Run time of the application.
2. Number of iterations to reach convergence in the calibration process.
3. Characteristics of input and output data.

21. Quality indicators of the output data

1. Information about negative or extreme calibration weights.
2. Tables of estimates including estimates of standard errors.
3. Basic statistics for ratios of weights (final weights/initial weights), e.g., mean, median, mode, standard deviation, variance, range, quantiles, interquartile interval.
4. Basic statistics for final weights, e.g., mean, median, mode, standard deviation, variance, range, quantiles, interquartile interval.
5. Histogram of distribution of initial and final weights.
6. Coefficient of correlation between initial and final weights.
7. Tables with margins estimated from the sample (initial weights), margins estimated using calibration weights and real margins in the population.

22. Actual use of the method

1. Calibration as a method of weighting is used by many statistical offices in many surveys. For instance, the Central Statistical Office in Sweden uses calibration in The Survey on Life and Health. This method was also used in Swedish household budget surveys to estimate average consumer expenditures. For details, see Särndal and Lundström (2005), Cassel, Lundquist and Selén (2002). The Hungarian Central Statistical Office (HCSO) adopted this approach in its Household Budget Survey in 1994 and in the Labour Survey in 1995 to compensate for nonresponse and for coverage deficiencies. HCSO uses this method in the form of the so called generalised iterative scaling (raking). For details, see Éltető and Mihályfi (2002).

2. In Poland the calibration approach is also used by the Central Statistical Office. For instance, the surveys which make use of calibration to compensate for the high percentage of nonresponse are the European Survey on Income and Living Conditions (EU-SILC) and the National Census of Population and Housing 2011. For details, see the Central Statistical Office in Poland (2011).
3. It is also worth noting that in many surveys calibration as a method of weighting and adjusting initial weights in order to reconstruct the known totals of auxiliary variables is recommended by Eurostat. This recommendation concerns primarily the European Union Survey on Income and Living Conditions (EU-SILC), where Eurostat recommends the method of integrated calibration. The idea of this approach is to use auxiliary variables defined at both household and individuals levels in such a way as to ensure consistency between households and individual estimates. After calibration households members will have the same household cross-sectional weight as the personal cross-sectional weight. This approach is used by many statistical offices in practice in EU-SILC. For details see Eurostat (2004).
4. In business statistics calibration is also used in practice. This method was used, for instance, by ISTAT, in the survey of Structural Business Statistics for small-medium enterprises. For more details see Casciano, Giorgi, Oropallo and Siesto (2012). Calibration was also used as a weighting technique for the Structural Business Survey on enterprises at Statistics Belgium. For details, see Vanderhoeft (2001). As a method of treating nonresponse, calibration was used in the MEETS project in a simulation study aimed at checking how it could improve the process of estimation for business data. For details, see MEETS (2011).

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Main Module
2. Weighting and Estimation – Design of Estimation – Some Practical Issues
3. Weighting and Estimation – Small Area Estimation

24. Related methods described in other modules

1. Weighting and Estimation – Generalised Regression Estimator
2. Weighting and Estimation – Outlier Treatment

25. Mathematical techniques used by the method described in this module

1. Advanced knowledge of linear algebra (including operations on matrices) and differential calculus is required. To find calibration weights the method of Lagrange multipliers is required. In many cases, when calibration weights should be bounded, optimisation algorithms, e.g., the Newton-Raphson approach should be used.

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate weights

2. 5.7 Calculate aggregates

27. Tools that implement the method described in this module

Calibration as a method of weighting is implemented in many statistical programs and described in details in many articles. Presented below is a short description of the most popular software devoted to calibration and an example of how to use CALMAR2 to determine calibration weights.

1. **Bascula 4.0** – the statistical tool developed in the Delphi language by Statistics Netherlands for the calculation of estimates of population totals, means and ratios. This program uses the so called Balanced Repeated Replication method to adjust weights and Taylor series methods for variance estimation. For details, see Nieuwenbroek and Boonstra (2001).
2. **Caljack** – this is a SAS macro written and developed by Statistics Canada and is an extension of the Calmar macro. This macro provides all the calibration methods which are available in Calmar and is able to calculate variance for many statistics like totals, ratios etc.
3. **CALMAR/CALMAR 2** – the statistical software developed by INSEE. Calmar is a SAS macro program that implements the calibration approach and adjusts weights assigned to individuals using auxiliary variables. Calmar 2 is the newest version of this software and was developed in France in 2003. It implements the generalised calibration method of handling nonresponse. For details, see Sautory (2003).
4. **CALWGT** – this is a freely distributed program for calibration written by Li-Chun Zhang in S-plus for Unix. The user is given the possibility to choose one of the methods, i.e., linear or multiplicative with many options (unrestricted, truncated or restricted approach etc.).
5. **CLAN 97** – the statistical software designed to handle surveys in Statistics Sweden. This is a SAS program (macro) written in 4GL language which is designed to compute point and standard error of estimates in sample surveys. For details, see Andersson and Nordberg (2000).
6. **G-Calib 2** – the statistical software developed in the SPSS language by Statistics Belgium. For details on how to implement this program, see Vanderhoeft (2002).
7. **GES** – this is a SAS-based application with a Windows-like interface which was developed in SAS/AF by Statistics Canada. Details related to GES can be found in Estevao, Hidioglou and Särndal (1995).
8. **R** – this is a free statistical software. The *calibrate* function, which can be found in the *survey* package, reweights the survey design weights and also adds additional information about estimated standard errors. For details, see Lumley (2012).
9. **ReGenesees System** – ReGenesees (R evolved Generalised software for sampling estimates and errors in surveys) – this is an R-based, full-fledged software system for design-based and model-assisted analysis of complex sample surveys with a user friendly interface which is very required especially by non R users. For details see web page <https://joinup.ec.europa.eu/software/regenesees/description>.

28. Process step performed by the method

Calibration of weights and estimation of parameters

Administrative section

29. Module code

Weighting and Estimation-M-Calibration

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	30-06-2012	first version	Marcin Szymkowiak	GUS (Poland)
0.2	04-12-2012	second version	Marcin Szymkowiak	GUS (Poland)
0.3	18-03-2014	third version	Marcin Szymkowiak	GUS (Poland)
0.3.1	19-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:32



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Generalised Regression Estimator

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Properties of GREG estimator.....	4
2.2 Particular cases and extensions	5
3. Preparatory phase	5
4. Examples – not tool specific.....	5
4.1 The Small-Medium Enterprises Survey and the current sampling strategy	5
5. Examples – tool specific.....	10
6. Glossary.....	10
7. References	10
Specific section.....	12
Interconnections with other modules.....	14
Administrative section.....	16

General section

1. Summary

The basic estimator (see “Weighting and Estimation – Main Module”) of a target parameter expands the observed values on the sample units using direct weights, which are the inverse of the inclusion probabilities. Generalised regression estimator is a model assisted estimator designed to improve the accuracy (see “Quality Aspects – Quality of Statistics”) of the estimates by means of auxiliary information. GREG estimator guarantees the coherence between sampling estimates and known totals of the auxiliary variables, as well. In fact, it is a special case of a calibration estimator (see “Weighting and Estimation – Calibration”) when the Euclidean distance is used.

2. General description of the method

In the estimation phase, the sample values are weighted to represent also unobserved units. When auxiliary information is available at unit or domain level, a GREG estimator can be used in order to reduce the variance of the estimates by using the relationship between the target variable and the auxiliary variables. At the same time the resulting weights allow calibration to the known totals.

Let y , x be the target variable and the vector of auxiliary variables, respectively.

The GREG estimator (Cassel, Särndal, and Wretman, 1976) can be expressed as a sum of the Horvitz Thompson estimator (HT) (see “Weighting and Estimation – Main Module”) and a weighted difference between known totals and their HT estimator:

$$\hat{t}_{GREG} = \sum_{i \in s} d_i y_i + \hat{\beta}' \left(t_X - \sum_{i \in s} d_i x_i \right), \quad (1)$$

where d_i , $i=1, \dots, n$, is the direct weight equal to the inverse of the inclusion probability, t_X is the vector of known population totals; moreover $\hat{\beta}$ is an estimate of the vector of regression coefficients of y on x , given by

$$\hat{\beta} = \left(\sum_{i \in s} d_i q_i x_i x_i' \right)^{-1} \sum_{i \in s} d_i q_i x_i y_i,$$

with q_i scale factors chosen properly, e.g., to account for heteroscedasticity. For example, when the variability of the target y depends on enterprises' size, z , the q_i can be chosen as $\sqrt{z_i}$.

In general, z may also be one of the covariates in the regression model.

Alternatively, the GREG estimator can be formulated in terms of predicted values for the target variables calculated on the basis of a linear relationship between y and x . More specifically, these predicted values are used in the estimation together with the residuals from the model, evaluated for sample units, i.e., the GREG estimator can be written as

$$\hat{t}_{GREG} = \sum_{i=1}^N \hat{y}_i + \sum_{i \in s} d_i e_i = \sum_{i=1}^N \hat{y}_i + \sum_{i \in s} d_i (y_i - \hat{y}_i), \quad (2)$$

where $\hat{y}_i = x_i' \hat{\beta}$ is the predicted value according to the linear model that relates y and x , e_i is the evaluated residual for a unit in the sample.

Finally, the GREG estimator can be conveniently formulated as a weighted sum of sample values:

$$\hat{t}_{GREG} = \sum_{i \in S} d_i g_{i,s} y_i = \sum_{i \in S} w_{i,s} y_i \quad (3)$$

where the correction factor g_{is} of the direct weights is given by

$$g_{i,s} = 1 + \left(t_X - \sum_{j \in S} d_j x_j \right)' \left(\sum_{j \in S} d_j q_j x_j x_j' \right)^{-1} q_i x_i \quad (4)$$

which does not depend on the target variable y .

2.1 Properties of GREG estimator

A fundamental property of the GREG estimator is that it is nearly design unbiased (Särndal et al., 1992).

The linear GREG estimator is motivated via the linear assisting model (Särndal et al., 1992)

$$E(y_i) = \beta' x_i, \quad V(y_i) = \sigma_i^2. \quad (5)$$

However, the knowledge of all x values is not necessary to evaluate linear GREG, because the knowledge of totals suffices to calculate the new weights, $w_{i,s}$ (see section 12 below).

The regression coefficient in (5) can be estimated at national level or for a disaggregated level, e.g., NUTS2. This level is referred as model group $\{U(p)\}$. In case of sub-national model group, the known totals need to be available at this level.

An important feature of the linear GREG is that the weighting system does not depend on the target variable but only on x values, as (4) shows.

The GREG estimator is calibrated to the known totals of the assisting model, that is

$$\sum_{i \in S} w_{i,s} x_i = t_X.$$

In fact GREG is a particular case of a calibration estimator (see “Weighting and Estimation – Calibration”) when using the Euclidean distance. Moreover, all the calibration estimators can be asymptotically approximated by the GREG (Deville and Särndal, 1992).

Another relevant property of GREG estimator is that the evaluation of its variance (see “Quality Aspects – Quality of Statistics”) is based on the variance of the residuals $(y_i - \hat{y}_i)$ (Särndal et al., 1992). As a consequence of this, the higher the fitting of the linear working model the lower the variance of GREG estimator and therefore the higher its accuracy. On the contrary, if the model underlying the GREG is not appropriate for the target variable, a too large variation of weights may increase the variance with respect to the HT estimator. In fact, variability of weights unrelated with the target variable can increase the variance of the estimates, an approximation of this impact is given (Kish, 1995) by

$$1 + \text{CV}(w_{i,s})^2, \quad (6)$$

where CV stands for the coefficient of variation of final weights.

A possible drawback of the GREG estimator is that it can produce negative weights (cf. section 10 below); on the contrary, in the framework of the calibration estimator, it is possible to obtain weights always positive using different distance functions (see “Weighting and Estimation – Calibration”).

2.2 Particular cases and extensions

The ratio estimator is a special case of GREG assisted by a model with only one covariate, obtainable if the variance of the target variable is assumed to be a linear function of the auxiliary variable $V(y_i) = \sigma^2 x_i$ (Deville, Särndal, 1992).

Extended GREG estimators are defined replacing the assisting model (4) with more general (non-linear, generalised, or mixed) models.

The non-linear GREG estimators (e.g., Lehtonen and Veijanen, 1998) require a separate model fitting for every target variable; hence, an important drawback of this kind of model assisted estimators is that they do not produce a unique system of weights uniformly applicable.

On the other side, the nonlinear GREG may give a considerably reduction in variance, as a result of the more refined models that can be considered when there is complete unit level auxiliary information.

3. Preparatory phase

4. Examples – not tool specific

4.1 The Small-Medium Enterprises Survey and the current sampling strategy

Small and Medium-sized Enterprises (SME) sample survey is carried out annually by sending a postal questionnaire with the purpose of investigating profit-and-loss account of enterprises with less than 100 persons employed, as requested by SBS EU Council Regulation n. 58/97 (Eurostat, 2003) and n. 295/2008. The units involved in the survey have also the possibility to fill in an electronic questionnaire and transmit it to Istat via web.

The survey covers enterprises belonging to the following economic activities according to the Nace Rev.1.1 classification:

- Sections C, D, E, F, G, H, I, J (division 67), K;
- Sections M, N and O for the enterprises operating in the private sector.

Main variables of interest asked to the SME sampled enterprises are Turnover, Value added at factor cost, Employment, Total purchases of goods and services, Personnel costs, Wages and salaries, Production value. They are also asked to specify their economic activity sector and geographical location in order to test the correctness of the frame with respect to these information. Totals of variables of interest are estimated with reference to three typologies of domains of study.

4.1.1 Frame of interest

For the reference year 2007, the population of interest for SME sample surveys is about 4.5 millions active enterprises.

The frame for SME survey is represented by the Italian Statistical Business Register (SBR). It results from the logical and physical combination of data from both statistical sources (surveys) and administrative sources (Tax Register, Register of Enterprises and Local Units, Social Security Register, Work Accident Insurance Register, Register of the Electric Power Board) treated with statistical methodologies. Variables in the register are both quantitative (Average number of employees in the year $t-1$, Number of employees in date 31/12/year $t-1$, Independent employment in date 31/12/year $t-1$, Number of enterprises) and qualitative (Geographical location, Economic activity according to Nace Rev.1.1- 4 digit). From the Fiscal Register is also provided the VAT Turnover, which represents a good proxy of the variable Turnover asked to the sampled enterprises by questionnaire.

4.1.2 Sampling design (allocation and domain of estimates)

SME is a multi-purpose and multi-domain survey and it produces statistics on several variables (mainly economic and employment variables) for three types of domains, each defining a partition of the population of interest (see Tables 1 and 2).

Table 1: Types of SME Survey domains

Type of domain		Number of Domains
Code	Description	
DOM1	Class of economic activity (4-digit Nace Rev.1*)	461
DOM2	Group of economic activity (3-digit Nace Rev.1) by size-class of employment	1.047
DOM3	Division of economic activity (2-digit Nace Rev.1) by region	984

*Nace Rev.1 = Statistical Classification of Economic Activities in the European Communities

Table 2: Definition of Size-classes of employment for domain DOM3 of SME Survey

Nace Rev.1.1 2-digit level	Size-classes of employment
10-45;	1-9; 10-19; 20-49; 50-99;
50-52;	1; 2-9; 10-19; 20-49; 50-99;
55;60-64;67;70-74;	1; 2-9; 10-19; 20-49; 50-99;
80; 85; 90; 92; 93;	1-9; 10-19; 20-49; 50-99;

Sampling design of the SME survey is a one stage stratified random sampling, with the strata defined by the combination of the modality of the characters Nace Rev.1.1 economic activity, size class and administrative region. A fixed number of enterprises are selected in each stratum without replacement and with equal probabilities. The number of units to be selected in each stratum is defined as a solution of a linear integer problem (Bethel, 1989).

In particular, the minimum sample size is determined in order to ensure that the variance of sampling estimates of the variable of interest in each domain does not exceed a given threshold, in terms of coefficient of variation. The variables of interest used for sample allocation are *Number of persons employed*, *Turnover*, *Value added at factor cost*, whose mean and variance are estimated in each stratum by data from the frame and data collected from the previous survey, respectively.

About 103,000 of small and medium-sized enterprises (units) are included in the sample. The sampling units are drawn by applying JALES procedure (Ohlsson, 1995) in order to take under control the total statistical burden, by achieving a negative co-ordination among samples drawn from the same selection register.

4.1.3 The weighting procedure

After calculating the total non-response correcting factors as the ratio of the number of sampled units and the number of responding units belonging to appropriate “weighting adjustment cells”, the weight of every single enterprise is further modified in order to match known or alternatively estimated population totals called benchmarks. In particular, known totals of selected auxiliary variables on the Business Register (Average number of employees in the year t-1, Number of enterprises) are currently used to correct for sample-survey nonresponse or for coverage error resulting from frame undercoverage or unit duplication.

Practical aspects in the application of the weighting procedure in the contest of SME survey

The evaluation of final weights for SME survey is usually carried out using the selected auxiliary variables, for the three types of domains described in Table 1. The optimisation problem underlying the GREG estimation process can be therefore formulated in the following way:

- the model group $\{U(p)\}$ is defined as the division of economic activity (2-digit Nace Rev.1.1) of the frame (the updated Business Register);
- the domains of interest are represented by the three typologies of partitions (described in Tables 1 and 2);
- the auxiliary variables are identified by
 - $x_1 = \text{Number of enterprises}$
 - $x_2 = \text{Average number of employees in the year } t-1$;
- for each enterprise, the vector \mathbf{x}_i of the auxiliary variables has been defined as follows:

$\mathbf{x}'_i = (\mathbf{x}'_{1i}, \mathbf{x}'_{2i})$, combination of two vectors \mathbf{x}'_{1i} and \mathbf{x}'_{2i} whose form is, respectively:

$$\mathbf{x}'_{1i} = \{\lambda_i(j_d)\},$$

$$\mathbf{x}'_{2i} = \{\alpha_i \lambda_i(j_d)\} \quad \text{with } d=1, \dots, 3; j=1, \dots, J_d,$$

where, according to the updated Business Register information:

- $\lambda_i(j_d)$ is a dichotomous variable whose value is equal to 1 if the unit i belongs to domain j_d and equal to 0 otherwise;
- α_i is the number of employed of enterprise i ;
- for each model group $\{U(p)\}$, i.e., for each division of economic activity (2-digit Nace Rev.1.1), the known population totals calculated on the updated frame, are expressed by:

$$X_{U(p)} = \sum U(p) \mathbf{x}'_i = (\sum U(p) \lambda_i(j_1), \dots, \sum U(p) \lambda_i(j_3), \sum U(p) \alpha_i \lambda_i(j_1), \dots, \sum U(p) \alpha_i \lambda_i(j_3)).$$

An example

In Table 3A the NACE code of every single domain of interest is listed in each cell; in the input data set of the weighting procedure each of them is replaced by the respective population total, in terms of the auxiliary variable *Average number of employees in the year $t-1$* (a similar specification is done in terms of the auxiliary variable *Number of enterprises*):

Table 3A: Example of benchmark specification (known totals)

DOMAIN	DOM1: Nace-4 digit (codes)					DOM2: Nace-3 digit * Size class (codes)					DOM3: Nace-2digit *Nuts		
Nace 2 digit	Tx1	Tx2	Tx3	Tx4	Tx _{jd}	Tx15	Tx16	Tx17	Tx18		Tx180	Tx181	Tx182
10	1010	1030	0	0	..	101*1-9	101*10-19	101*20-49	103*1-9	..	north	central	south
11	1111	1112	1113	1120	..	111*1-9	111*10-19	112*10-19	112*20-49	..	north	central	south
13	1310	1320	0	0	..	131*1-9	131*10-19	133*50+	132*1-9	..	north	central	south
14	1411	1412	1413	1421	..	141*1-9	141*10-19	141*20-49	142*50+	..	north	central	south
15	1511	1512	1513	1520	..	151*1-9	151*10-19	151*50+	152*1-9	..	north	central	south
16	1600	0	0	0	..	160*50+	0	0	0	..	north	central	south
17	1711	1712	1713	1715	..	171*1-9	171*10-19	171*20-49	172*1-9	..	north	central	south
...
93	9301	9302	9303	9304		930*1-9	930*1-9	930*10-19	930*50+		north	central	south

For each responding unit (enterprise), the vector of the auxiliary variable *Average number of employees in the year $t-1$* can be expressed as in Table 3B, whether or not the unit belongs to the domain represented on the cell:

Table 3B: Example of sample data specification (α_k = number of persons employed of enterprise k)

				DOM1: Nace-4 digit (x_2 -values)					DOM2: Nace-3 digit * Size class (x_2 --values)					DOM3: Nace-2digit *Nuts		
Unit identifier	Domain Nace2	q_k	Direct weight	x1	x2	x3	x4	x _{jd}	x15	x16	x17	x18		X180	X181	X182
1	10	α_1	22	α_1	0	0	0	..	α_1	0	0	0	..	0	α_1	0
2	10	α_2	1,4	0	α_2	0	0	..	0	α_2	0	0	..	0	0	α_2
3	93	α_3	10,5	0	0	0	α_3		0	0	0	α_3	..	0	α_3	0
4	14	α_4	3	α_4	0	0	0	..	α_4	0	0	0	..	α_4	0	0
5	17	α_5	6,4	0	α_5	0	0	.	0	0	α_5	0	..	α_5	0	0
...	...	α_k
n_p	14	α_{np}	18	0	0	α_{np}	0	..	α_{np}	0	0	0	..	α_{np}	0	0

The overall number of benchmarks (constrained estimates) in the optimisation process is equal to 182. In spite of the considerable number of constraints to be satisfied, the weighting process ends with a good convergence between final estimates and know population totals (see Table 4).

*Table 4A: Example of output of the weighting procedure for a domain of interest (2 digit NACE):
Check on the auxiliary variables*

Domain code =14						
Constraint code	Known Totals (1)	Final Estimates of X variables (2)	Direct Estimates of the X variables (3)	(2)-(1)	(3)-(1)	Sampling units
1	83081	83081	62133	0	-20948	57
2	249069	249069	227430	0	-21639	17397
3	33099	33099	23961	0	-9138	426
4	27339	27339	22323	0	-5016	12
5	451	451	294	0	-157	63
....
182	4676	4676	3375	0	-1301	6

*Table 4B: Example of output of the weighting procedure for a domain of interest (2 digit NACE):
Final weights*

Unit identifier	Domain Nace2	q_k	Direct weight	Final weight
1	10	α_1	22	18,2
2	10	α_2	1,4	2
3	93	α_3	10,5	12
4	14	α_4	3	5
5	17	α_5	6,4	4,2
...	...	α_k
n_p	14	α_{np}	18	15

The estimator effect for the final weights has been calculated on the sample of responding enterprises with less than 100 persons employed at division of activity level (NACE Rev.1.1-2 digit), for the following subset of target variables:

1. Turnover (code 12 11 0)

2. Value added at factor cost (code 12 15 0)
3. Personnel costs (code 13 31 0)
4. Gross investment in tangible goods (code 15 11 0)
5. Number of employees (code 16 13 0)
6. Wages and salaries (code 13 32 0).

The estimator effect values confirm the higher efficiency gained by using the GREG estimator instead of the direct estimation for most of the considered divisions of activities and target variables; the main exception concerns the variable “Gross investment in tangible goods”, which is hardly predictable by a model. Moreover, the variables “Turnover” and “Value added at factor cost” have an estimator effect higher than 1 for some divisions, i.e., 73-“Research and development” and 74-“Other business activities”, that are characterised by specialised activities where the high amounts invoiced by the enterprises can be attained by a relatively small number of skilled employees.

In conclusion, apart from a small group of economic activity classes, the variable “average number of employees” has shown a good correlation with the following target variables of interest: “turnover”, “production value”, whereas it is not enough correlated with “Gross investment in tangible goods”.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bethel, J. (1989), Sample Allocation in Multivariate Surveys. *Survey Methodology* **15**, 47–57.
- Breidt, F. J. and Opsomer, J. D. (2000), Local Polynomial Regression Estimators in Survey Sampling. *The Annals of Statistics* **28**, 1026–1053.
- Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1976), Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika* **63**, 615–620.
- Deville, J. C. and Särndal, C.-E. (1992), Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* **87**, 367–382.
- Deville, J. C., Särndal, C.-E., and Sautory, O. (1993), Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association* **88**, 1013–1020.
- Hedlin, D., Falvey, H., Chambers, R., and Kokic, P. (2001), Does the Model Matter for GREG Estimation? A Business Survey Example. *Journal of Official Statistics* **17**, 527–544.
- Kish, L. (1995), Methods for Design Effects. *Journal of Official Statistics* **11**, 55–77.
- Lehtonen, R. and Veijanen, A. (1998), Logistic Generalized Regression Estimators. *Survey Methodology* **24**, 51–55.

- Montanari, G. E. and Ranalli, M. G. (2005), Nonparametric Model Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association* **100**, 1429–1442.
- Ohlsson, E. (1995), Coordination of Samples using Permanent Random Numbers. In: *Business Survey Methods* (eds. Cox, B. G., Binder, D. A., Chinnapa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S.), Wiley, New York, 153–169.
- Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer Verlag, New York.

Specific section

8. Purpose of the method

The method is used for estimation, when auxiliary information is available at unit or domain level. It can be used to reduce the variance of the estimates if a strong correlation between the target variable and the auxiliary variables exists. At the same time, GREG allows to calibrate to the known population totals of the auxiliary variables x . This means that GREG is a particular case of a calibration estimator when the distance function is linear, i.e., the final weights that satisfy the calibration equations w are chosen to minimise the following distance with the initial weights d :

$$\sum_s \frac{(w - d)^2}{2d}.$$

9. Recommended use of the method

1. GREG is recommended when a linear relationship between target y and covariate variables x is present, $y = \beta x + \varepsilon$.

10. Possible disadvantages of the method

1. GREG can introduce a large variation in weights that can cause an increase in variance, see formula (6) to quantify the impact.
2. Possibly correction weights g too far from unity or negative final weights as the correction factors (see formula (4)) can be in some cases a negative quantity.
3. Even being asymptotically unbiased, bias can be introduced if sample size is too small (see also section 14).
4. GREG can be very sensitive to presence of outliers (see “Weighting and Estimation – Outlier Treatment”); an illustrative example with discussion can be found in Hedlin et al. (2001). This issue is very relevant to business survey where target variables are typically non-normal and very skewed.

11. Variants of the method

1. Specific case: Ratio regression.
2. Non-linear GREG estimators. Expression (2) can be applied on general models. In fact, the prediction \hat{y}_i , that for GREG is based on linear model can be based on more complex models if the target variable for example is not normal. An example of non-linear GREG is logistic GREG which is based on logistic model when the target variable is a binary variable. The use of more complex models, however, requires more detailed information on the x variable w.r.t. the knowledge of population total that is needed by (linear) GREG.

12. Input data

1. Ds-input1 = elementary sample data containing covariates, direct weights and scale coefficients q_i , model group (i.e., level for which the model is specified).

2. Ds-input2 = known totals on the covariates for each model group.

13. Logical preconditions

1. Missing values
 1. GREG is calculated on sample values on DS-input1 after imputation – anyway, variance estimation is affected by the imputation.
 2. Ds-input2 cannot contain missing values.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. If the auxiliary variables are categorical, the known totals for different partitions should not be in conflict.

14. Tuning parameters

1. Choice of the auxiliary covariates in the model, a rule of thumb for the choice of categorical variable is to define categories so that the sample totals are greater than 30.
2. Choice of the model group level.
3. Choice of q_i .

15. Recommended use of the individual variants of the method

1. Non-linear GREG can be used when auxiliary variables are available for each unit in the population and the relationship with the target variable is markedly non-linear.

16. Output data

1. Ds-output1 = elementary sample data set containing the new final weights.

17. Properties of the output data

1. The final weights allows to satisfy the implicit constraints given by the known totals of the auxiliary variables.

18. Unit of input data suitable for the method

Sample units, also separately by model group.

19. User interaction - not tool specific

1. Choice of auxiliary covariates.
2. Choice of the group level.

3. Choice of q_i .

20. Logging indicators

1. The run time of the application.
2. Iterations to attain convergence in the estimation process.
3. Characteristics of the input data, for instance problem size.

21. Quality indicators of the output data

1. The coefficient of variation of the final weights in comparison with the basic weights.
2. Presence of negative weights, in this case it may be appropriate to consider a different underlying model or to use a calibration estimator with a function that allows to restrict the range of final weights (see the module “Weighting and Estimation – Calibration”).
3. Variance, coefficient of variation of produced estimates.
4. Check of equality of sample estimates of x and known population totals.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Main Module
2. Quality Aspects – Quality of Statistics

24. Related methods described in other modules

1. Weighting and Estimation – Calibration
2. Weighting and Estimation – Outlier Treatment

25. Mathematical techniques used by the method described in this module

1. Matrix algebra

26. GSBPM phases where the method described in this module is used

1. 5.6 “Calculate weights”
2. 5.7 “Calculate aggregates”

27. Tools that implement the method described in this module

1. CALMAR (Deville, Särndal and Sautory 1993)
2. CLAN (Statistics Sweden)
3. BASCULA (The Netherlands)

4. GES (StatCan)
5. GENESEES (ISTAT)
6. Survey, an R package downloadable from the CRAN
7. Sampling, an R package downloadable from the CRAN
8. REgenesees (ISTAT), an R package downloadable from the JoinUP:
<https://joinup.ec.europa.eu/software/regenesees/release/release150#download-links>

28. Process step performed by the method

Estimation

Administrative section

29. Module code

Weighting and Estimation-M-Generalised Regression Estimator

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	06-02-2012	first version	Loredana Di Consiglio, Claudia De Vitiis, Cristina Casciano	ISTAT
0.2	03-05-2012	second version	Loredana Di Consiglio, Claudia De Vitiis, Cristina Casciano	ISTAT
0.2.1	22-06-2012	second version – with corrections	Loredana Di Consiglio, Claudia De Vitiis, Cristina Casciano	ISTAT
0.2.2	11-11-2013	second version – after EB review	Loredana Di Consiglio, Claudia De Vitiis, Cristina Casciano	ISTAT
0.2.3	13-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:32



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Outlier Treatment

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Winsorisation.....	5
3. Preparatory phase	10
4. Examples – not tool specific.....	10
5. Examples – tool specific.....	10
6. Glossary.....	10
7. References	10
Specific section.....	13
Interconnections with other modules.....	15
Administrative section.....	17

General section

1. Summary

In business surveys, the distribution of variables is often highly skewed, resulting in sample observations that differ substantially from the majority of observations in the sample. The literature refers to these units as outliers.

Outliers can be *representative* (representing other population units similar in value to the observed outliers) or *non-representative* (unique in the population). Here we will consider only the case of representative outliers, i.e., correct values representing other units in the population. Since representative outliers affect the variability of the standard estimators (such as: Horvitz-Thompson or Generalised regression estimators (GREG)), an appropriate way of handling them is required.

The objective of outlier treatment is to make estimates for the population coherent with the real parameters for the population. This means that outlier treatment should be always a trade-off between variance and bias. For small samples, variance is usually the dominating factor in the MSE. On the other hand, bias dominates when the sample size is large.

The module describes one frequently applied estimation method used to reduce the impact of outlying units: Winsorisation. The general idea of Winsorisation involves modifying the outlying observation so that it has less impact on the estimate of a parameter. The effectiveness of the Winsor estimator in terms of its resistance to unusually large residuals depends on the choice of cut-off values, therefore the methods used to estimate the robust regression parameters and the bias parameters need to estimate cut-off values. The cut-offs are optimal only at the level at which estimates are being conducted. The Winsor estimator is easy to implement, but it performs best under models (used for estimating robust regression parameters) that are only moderately robust. Winsorisation can be applied to a large class of estimators (GREG estimators, model-based regression estimators, ratio estimators) and involves modifying their standard forms. This results in estimates with acceptable bias and a smaller variance than that of standard forms, non-Winsorised estimators. We can observe the bias-variance trade-off at the low level of estimation but aggregated Winsorised estimates have large biases, resulting in less precision compared to standard aggregated estimates.

2. General description of the method

In business surveys target variables tend to be highly skewed and populations may contain a number of extreme values, the so-called outliers. Although outliers are extreme, they need not necessarily be incorrect but are an integral part of each survey population and cannot be dismissed in the analysis.

According to Hawkins, “an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980). In the statistical literature outliers are observations that differ substantially from most of the observations in the sampled and the unsampled parts of the population. Outliers may be extreme big values or extreme small values. We can distinguish *large outliers*, if the values are extremely large than the other values of the “normal” units, or *small outliers*, if the values are extremely smaller than the other values of the “normal” units.

Two distinct types of outliers can be defined: “y-outliers” or “outliers in the y-direction” and “x-outliers” or “outliers in the x-direction” (Rousseeuw and Leroy, 2003) where *y-values* and *x-values* are the study variable and an auxiliary variable, respectively.

“Y-outliers” denote the *y* values of a few sample units that are very distant from the *y* values of other sample units. Another class of outliers comprises the *x* values of a few sample units that are very distant from the *x*-values of other sample units. These are “x-outliers”. They can have a substantial impact on the stability of the overall sample estimate because of their so-called “leverage” (Bergdahl et al., 1999).

Some authors (Chambers, 1986; Eltinge and Cantwell, 2006) classify outliers into three groups. The first are *representative outlier* values, which represent other population units similar in value to the observed outliers. These are correctly measured sample values that are outlying relative to the rest of the sample data and we cannot assume that similar values do not exist in the non-sampled part of the survey population. The second group consists of *non-representative outlier* values, which are unique in the population (in the sense that there is no other unit like them) (Chambers, 1986). The third group comprises gross measurement errors, which are outlying observations that are not true values.

Here we will not consider gross errors in the sampled data, caused by deficiencies in the survey processing (e.g., miscoding). Such errors are corrected during the data editing process (Eltinge and Cantwell, 2006).

Since outliers usually have a huge impact on estimates, outlier detection and their treatment are important elements of statistical analysis. This is true especially when estimation is carried out at a low level of aggregation. In the case of small sample sizes, outliers can affect variance. Even if the sample size is large, the influence of an outlier can significantly increase the variance resulting in a decreased efficiency of estimation. Dealing with outliers has two aspects: the first one involves identifying outlying observations in an objective way, and the second one focuses on ways of handling them to reduce their effect on survey estimates.

While non-representative outliers can be treated by post-stratification, representative outliers should be handled in the survey estimation process, by the use of outlier resistant or robust estimation procedures (Ren and Chambers, 2002).

In this module we consider only representative outliers, in other words, any extreme values that represent other true observations in the population.

There are three main methods of dealing with outliers in a finite population, apart from removing them from the dataset (Cox et al., 1995):

1. reducing the weights of outliers (trimming weight),
2. changing the values of outliers (Winsorisation, trimming),
3. using robust estimation techniques such as M-estimation.

Weight trimming reduces large weights to a fixed cut-off value and adjusts weights below this value to maintain the untrimmed weight sum, reducing variability at the cost of introducing some bias (Elliott, 2007). The literature mentions various approaches to determine cut-off points at which to trim weights. Most standard methods are ad hoc in that they do not use the data to optimise bias-variance trade-offs.

According to Potter (1990), a weight should be trimmed at the point where the loss of precision due to a large weight is larger than the bias introduced by trimming the weight. It can be said that the general approach involves reducing the survey weight associated with that observation (Chambers, 1996; Detlefsen, 1992; Elliott and Little, 2000; Potter, 1988, 1993; Theberge, 2000; Zaslavsky et al., 2001). In many business surveys it is a relatively common practice to set the survey weight equal to one. One could say that the identified outlier is a “non-representative” outlier (Eltinge and Cantwell, 2006). In some cases setting a weight equal to one may be viewed as the limiting case of more refined adjustment procedure like Winsorisation or M-estimation (Eltinge and Cantwell, 2006). Winsorisation is frequently used in business surveys, so it is presented below in more detail. The general idea of Winsorisation is that if an observation exceeds a pre-set cut-off value, then the observation is replaced by that cut-off value or by a modified value closer to the cut-off value.

2.1 Winsorisation

In business statistics, which are characterised by skewed distributions, GREG estimation procedures may provide unsatisfactory results. One of the most popular methods suggested in the literature consists in modifying values in the sample so that the estimator becomes robust and is not affected by large residuals (Kokic and Bell, 1994; Chambers, 1996; Chambers et al., 2000; Rivest and Hidirolou, 2004; Dehnel and Gołata, 2010). This approach is exemplified by Winsor estimation, which was applied for the first time in a survey conducted by Searls (1966). Winsorisation involves identifying cut-off (thresholds) values. Sample observations whose values lie outside certain pre-set cut-off values are transformed in order to make them closer to the cut-off value.

Winsorisation may be one-sided or two-sided. One-sided Winsorisation adjusts influential values deemed to be too large. Two-sided Winsorisation adjusts influential values deemed to be both too large and too small.

Cut-off values are derived in a way that approximately minimises the MSE of estimates. All sampled units are divided into two (or three) groups. One group contains typical observations, which are left unmodified, the other one(s) contain(s) observations regarded as (large or small) outliers. The classification is made on the basis of one (if outliers are not divided into large and small) or two (when, on the contrary, large and small outliers are distinguished) pre-set cut-off values. Then, values of the study variable outside the cut-off values are transformed so that they are no longer regarded as outliers. It should be stressed, however, that the modified values are artificial and may sometimes be unacceptable. As a result of Winsor estimation, we obtain a modified sample, in which untypical observations have been replaced with typical ones. Further calculations are conducted for the modified sample. Any kind of estimation can be used at this stage. Here, GREG estimation is illustrated.

The Winsorised estimator of the population total is defined $\hat{Y}_{win} = \sum_{i \in s} w_i y_i^*$ where y_i^* is the modified value of the study variable.

First, let us consider the case when we have only large outliers.

Two types of Winsorisation can be applied in the treatment of outliers. Winsorised Type I estimator is based on an arbitrary assumption whereby any outliers exceeding a pre-set cut-off value K are always replaced by that value K :

$$y_i^* = K \text{ if } y_i > K \text{ and } y_i^* = y_i \text{ otherwise.}$$

On the contrary, with a Type II estimator, as the GREG weight \tilde{w}_i decreases, the contribution of the observed values of the outliers increases – that is the modified value of the study variable “approaches” the value of the outlier, i.e., the real value of the variable.

Under Type II Winsorisation:

$$y_i^* = \left(\frac{1}{\tilde{w}_i} \right) y_i + \left(1 - \frac{1}{\tilde{w}_i} \right) K \text{ if } y_i > K \text{ and } y_i^* = y_i \text{ otherwise.}$$

The use of Winsor estimation reduces estimator variance, while, at the same time, it may introduce bias. However, if cut-off values are chosen appropriately, the decline in variance is big enough to offset the bias of MSE (Hedlin, 2004).

The main difficulty then lies in the choice of cut-off values for dividing observations in the sample. The optimum selection has a strong effect on estimation quality.

The Winsor estimator, with GREG estimation, can be expressed as:

$$\hat{Y}_{win} = \sum_{i \in s} \tilde{w}_i y_i^* = \sum_{i \in s} w_i g_i y_i^* \quad (1)$$

where, in the presence of outliers, modified values of the study variable y_i^* are calculated in the following manner (Gross et al., 1986):

$$y_i^* = \begin{cases} \left(\frac{1}{\tilde{w}_i} \right) y_i + \left(1 - \frac{1}{\tilde{w}_i} \right) K_{Ui} & \text{if } y_i > K_{Ui} \\ y_i & \text{if } K_{Li} \leq y_i \leq K_{Ui} \\ \left(\frac{1}{\tilde{w}_i} \right) y_i + \left(1 - \frac{1}{\tilde{w}_i} \right) K_{Li} & \text{if } y_i < K_{Li} \end{cases} \quad (2)$$

$$g_i = \left(1 + x_i' \left(\sum_{i \in s} w_i x_i x_i' \right)^{-1} \left(t_x - \sum_{i \in s} w_i x_i \right) \right)' \quad (3)$$

where:

$U = \{1, \dots, i, \dots, N\}$ - target population of size N ;

$s (s \subseteq U)$ - sample;

$\tilde{w}_i = w_i g_i$;

$w_i = 1/\pi_i$ - sampling weights;

g_i - weights dependent on the value of a vector of auxiliary variables for sampled units;

$x_i = (x_{1i}, \dots, x_{ki}, \dots, x_{Ki})'$ - vector of auxiliary variables;

$t_x = \sum_{i \in U} x_i$ - population total;

K_{Ui} - upper cut-off value; K_{Li} - lower cut-off value.

Based on formula (1) it can be assumed that a unit drawn into the sample is regarded as an element representing $(\tilde{w}_i - 1)$ non-sampled units. Hence, according to formula (2), an observation regarded as an outlier contributes its unweighted values, while the non-sampled units, represented by the remainder of the weight $(\tilde{w}_i - 1)$, contribute pre-set upper or lower cut-off values.

Cut-off values are calculated to minimise MSE of Winsorised estimator under the model (Preston and Mackin, 2002):

$$K_{Ui} = \mu_i^* - \frac{B_U}{(\tilde{w}_i - 1)} \quad (4)$$

$$K_{Li} = \mu_i^* - \frac{B_L}{(\tilde{w}_i - 1)} \quad (5)$$

where:

$\mu_i^* = E(Y_i^* | x_i)$ - conditional expectation under the assumed regression model;

$B_U = E[\hat{Y}_{winU} - \hat{Y}_{DIR}]$ - bias of \hat{Y}_{winU} ;

$B_L = E[\hat{Y}_{winL} - \hat{Y}_{DIR}]$ - bias of \hat{Y}_{winL} ;

\hat{Y}_{winU} - Winsor estimator of the population total when only upper Winsorisation is performed;

\hat{Y}_{winL} - Winsor estimator of the population total when only lower Winsorisation is performed.

When Winsorisation is mild and reasonably symmetric, being μ_i^* difficult to estimate, we can replace μ_i^* with μ_i . Then the approximately optimal cut-offs are (Preston and Mackin, 2002):

$$K_{Ui} = \mu_i - \frac{B_U}{(\tilde{w}_i - 1)} = \mu_i + \frac{G}{(\tilde{w}_i - 1)} \quad (6)$$

$$K_{Li} = \mu_i - \frac{B_L}{(\tilde{w}_i - 1)} = \mu_i + \frac{H}{(\tilde{w}_i - 1)} \quad (7)$$

Under the assumption $\mu_i = \hat{\mu}_i = \hat{\beta}x_i$ (Preston and Mackin, 2002), cut-off values are estimated based on the following formulas:

$$\hat{K}_{Ui} = \hat{\mu}_i - \frac{B_U}{(\tilde{w}_i - 1)} = \hat{\mu}_i + \frac{G}{(\tilde{w}_i - 1)} \quad \text{where } G = -B_U \quad (8)$$

$$\hat{K}_{Li} = \hat{\mu}_i - \frac{B_L}{(\tilde{w}_i - 1)} = \hat{\mu}_i + \frac{H}{(\tilde{w}_i - 1)} \quad \text{where } H = -B_L \quad (9)$$

where $\hat{\mu}_i = \hat{\beta}x_i$ - a robust estimate of regression parameter μ_i (see below).

In order to estimate the bias B_U under Winsorisation we can use the Kokic and Bell approach (1994). According to that approach, the value of B_U can be calculated by solving the equation:

$$G - E \left[\sum_{i \in s} \max\{D_i - G, 0\} \right] = 0 \quad (10)$$

where $D_i = (Y_i - \mu_i^*)(\tilde{w}_i - 1)$ are weighted residuals. Assuming $\hat{\mu}_i$ is a robust estimate of parameter μ_i , we obtain $\hat{D}_i = (Y_i - \hat{\mu}_i)(\tilde{w}_i - 1)$.

We can write the function $\psi_U(\hat{D}_{(k)})$ (Kokic and Bell, 1994):

$$\psi_U(\hat{D}_{(k)}) = \hat{D}_{(k)} - \sum_{i \in s} \max\{\hat{D}_i - \hat{D}_{(k)}, 0\} = (k+1)\hat{D}_{(k)} - \sum_{j=1}^k \hat{D}_{(j)} \quad (11)$$

where:

(k) - a number assigned to the unit drawn into the sample after ordering all units in the sample according to non-ascending estimated residuals $\hat{D}_i: \hat{D}_{(1)} \geq \hat{D}_{(2)} \geq \dots \geq 0 \geq \dots$.

By solving $\psi_U(G) = 0$ one can obtain the value of G . In practice, since it is difficult to find the right solution of the equation, two methods are proposed. According to the first one, G is estimated using the formula:

$$\hat{G} = \frac{1}{(k^* + 1)} \sum_{j=1}^{k^*} \hat{D}_{(j)} \quad (12)$$

where k^* is the last value of k for which the value of $\psi_U(\hat{D}_{(k)})$ is non-negative.

The second approach involves using linear interpolation between $\frac{1}{(k^* + 1)} \sum_{j=1}^{k^*} \hat{D}_{(j)}$ and $\frac{1}{(k^* + 2)} \sum_{j=1}^{k^*+1} \hat{D}_{(j)}$.

Then, \hat{G} can be expressed as (Preston and Mackin, 2002):

$$\hat{G} = \frac{\psi_U(\hat{D}_{(k^*+1)}) \left[\frac{1}{(k^* + 1)} \sum_{j=1}^{k^*} \hat{D}_{(j)} \right] - \psi_U(\hat{D}_{(k^*)}) \left[\frac{1}{(k^* + 2)} \sum_{j=1}^{k^*+1} \hat{D}_{(j)} \right]}{(\psi_U(\hat{D}_{(k^*+1)}) - \psi_U(\hat{D}_{(k^*)}))} \quad (13)$$

The value of H can be computed similarly. Estimates of weighted residuals $\hat{D}_i = (Y_i - \hat{\mu}_i)(\tilde{w}_i - 1)$ are arranged in ascending order $\hat{D}_{[1]} \leq \hat{D}_{[2]} \leq \dots \leq 0 \leq \dots$. Function $\psi_L(\hat{D}_{[m]})$ can be written as:

$$\psi_L(\hat{D}_{[m]}) = \hat{D}_{[m]} - \sum_{i \in s} \min\{\hat{D}_i - \hat{D}_{[m]}, 0\} = (m+1)\hat{D}_{[m]} - \sum_{l=1}^m \hat{D}_{[l]} \quad (14)$$

where:

$[l] = [1]..[m]$ - a number assigned to the unit drawn into the sample after ordering all units in the sample by estimated residuals \hat{D}_i .

The value \hat{H} can thus be evaluated as (cf. formula (15)) (Preston and Mackin, 2002):

$$\hat{H} = \frac{\psi_L(\hat{D}_{[m^{**}+1]}) \left[\frac{1}{[m^{**}+1]} \sum_{l=1}^{m^{**}} \hat{D}_{[l]} \right] - \psi_L(\hat{D}_{[m^{**}]}) \left[\frac{1}{[m^{**}+2]} \sum_{l=1}^{m^{**}+1} \hat{D}_{[l]} \right]}{(\psi_L(\hat{D}_{[m^{**}+1]}) - \psi_L(\hat{D}_{[m^{**}]})} \quad (15)$$

where m^{**} is the last value of m for which the value of $\psi_U(\hat{D}_{[m]})$ is non-positive.

In order to estimate cut-off values \hat{K}_{U_i} and \hat{K}_{L_i} , in addition to the above bias parameters $G = -B_U$ and $H = -B_L$ it is necessary to compute $\hat{\mu}_i = \hat{\beta}x_i$ which is an estimate of μ_i^* . For this purpose, robust regression methods can be used. Those recommended in the literature (Preston and Mackin, 2002) include: *Trimmed least squares (TLS)*, *Trimmed least absolute value (LAV)*, *Sample Splitting*, *Least median of squares (LMS)*.

The method of *Trimmed least squares (TLS)* involves first fitting an Ordinary Least Squares (OLS) regression model to minimise the function:

$$F = \sum_{i \in S} (y_i - \beta^T x_i)^2 \quad (16)$$

Then fitted values are calculated, and then residuals. In the second step, units with the largest positive and negative residuals are removed. As a rule, the sample is reduced by about 5%. Finally, a new regression model is fitted to the reduced sample in order to estimate the value of μ_i^* . One advantage of the TLS is that it is quick to run and simple.

Another method used in robust regression is *Trimmed least absolute value (LAV)*. It consists in fitting a regression model to minimise the function:

$$F = \sum_{i \in S} |y_i - \beta^T x_i| \quad (17)$$

After evaluating fitted values and residuals, as is the case in the TLS method, units with the largest positive and negative residuals are removed. A new regression model is fitted to the reduced sample. It is expected that the LAV method is a more robust regression model than the TLS technique because large residuals which are not squared have less influence on the regression parameters.

Another example of robust regression is *Sample Splitting Technique* based on Ordinary Least Squares (OLS). It is applied to a dataset that has been randomly split into two halves. A regression model is fitted to each half of the data while the residuals are calculated using the model applied to the half of the data that was not used to fit the model. Then, after merging the data, units with the largest positive and negative residuals are removed. The process is repeated until a certain percentage of data has been deleted. The SS technique is expected to be more robust than TLS because the residuals used to remove the ‘outlier’ units are not calculated from a regression model that has been generated using these ‘outlier’ units.

The list of robust regression techniques cannot be complete without the *Least median of squares (LMS)* technique. It was described by Rousseeuw and Leroy (2003). It resembles the bootstrap method. It involves drawing subsamples of size $n - 1$ from a sample of size n using simple random sampling with replacement. For each subsample trial regression model parameters are calculated and then their squared residuals, which are used to calculate the median. The model with the smallest median of squared residuals is selected. The LMS technique should be more robust than TLS because

an OLS regression model is fitted in the absence of "outlier" units, without totally removing these 'outlier' units.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bergdahl, M., Black, O., Bowater, R., Chambers, R., Davies, P., Draper, D., Elvers, E., Full, S., Holmes, D., Lundqvist, P., Lundström, S., Nordberg, L., Perry, J., Pont, M., Prestwood, M., Richardson, I., Skinner, C., Smith, P., Underwood, C., and Williams, M. (1999), *Model Quality Report in Business Statistics, Volume I, Theory and Methods for Quality Evaluation*. <http://users.soe.ucsc.edu/~draper/bergdahl-et al-1999-v1.pdf>
- Chambers, R. L. (1986), Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association* **81**, 1063–1069.
- Chambers, R. L. (1996), Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics* **12**, 3–32.
- Chambers, R., Dorfman, A. H., and Wehrly, T. E. (1993), Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. *Journal of the American Statistical Association* **88**, 268–277.
- Chambers, R., Kokic, P., Smith, P., and Cruddas, M. (2000), Winsorization for Identifying and Treating Outliers in Business Surveys. *Proceedings of the Second International Conference on Establishment Surveys (ICES II)*, 687–696.
- Chambers, R., Brown, G., Heady, P., and Heasman, D. (2001), Evaluation of Small Area Estimation Methods – an Application to Unemployment Estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: a Methodological Perspective*.
- Chambers, R. L., Falvey, H., Hedlin, D., and Kokic P. (2001a), Does the Model Matter for GREG Estimation? A Business Survey Example. *Journal of Official Statistics* **17**, 527–544.
- Cox, B. G., Binder, A., Chinnappa, N. B., Christianson, A., Colledge, M. J., and Kott P. S. (eds.) (1995), *Business Survey Methods*. John Wiley & Sons.

- Dehnel, G. and Gołata, E. (2010), On some robust estimators for Polish Business Survey. *Statistics in Transition* - new series **11**, number 2, Warszawa 2010. s. 287-312 (Central Statistical Office and Polish Statistical Association), 58–71, Summ. - Bibliogr. ISBN 978-83-7027-431-3.
- Detlefsen, R. (1992), A Weight Adjustment Technique. Internal Memorandum, Bureau of the Census.
- Elliott, M. (2007), Bayesian weight trimming for generalized linear regression models. *Survey Methodology* **33**, 23–34.
- Elliott, M. R. and Little, R. J. A. (2000), Model-Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics* **16**, 191–209.
- Eltinge, J. and Cantwell, P. (2006), Outliers and Influential Observations in Establishment Surveys. Paper prepared for presentation to the Federal Economic Statistics Advisory Committee (FESAC), 09-06-2006.
- Gross, W. F., Bode, G., Taylor, J. M., and Lloyd-Smith, C. W. (1986), Some finite population estimators which reduce the contribution of outliers. *Proceedings of the Pacific Statistical Conference, 20–24 May 1985, Auckland, New Zealand*.
- Hawkins, D. (1980), *Identification of Outliers*. Chapman and Hall.
- Hedlin, D. (2004), Business Survey Estimation. R&D, Sweden.
- Kokic, P. N. and Bell, P. A. (1994), Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics* **10**, 419–435.
- Potter, F. J. (1988), Survey of Procedures to Control Extreme Sampling Weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453–458.
- Potter, F. (1990), A Study of Procedures to Identify and Trim Extreme Sample Weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 225–230.
- Potter, F. J. (1993), The Effect of Weight Trimming on Nonlinear Survey Estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758–763.
- Preston, J. and Mackin, C. (2002), Winsorization for Generalised Regression Estimation. Paper for the Methodological Advisory Committee, November 2002, Australian Bureau of Statistics.
- Ren, R. and Chambers, R. L. (2002), Outlier robust imputation of survey data via reverse calibration. S3RI Methodology Working Paper M03/19.
- Ren, R. and Chambers, R. L. (2002a), Outlier Robust Methods: Outlier Robust Estimation and Outlier Robust Imputation By Reverse Calibration. Report for Euredit, <http://www.cs.york.ac.uk/euredit/>.
- Rivest, L.-P. and Hidiroglou, M. (2004), Outlier treatment for disaggregated estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- Rousseeuw, P. and Leroy, A. (2003), *Robust Regression and Outlier Detection*. John Wiley & Sons.
- Searls, D. T. (1966), An estimator which reduces large true observations. *Journal of the American Statistical Association* **61**, 1200–1204.
- Theberge, A. (2000), Calibration and Restricted Weights. *Survey Methodology* **26**, 99–107.

Zaslavsky, A. M., Schenker, N. and Belin, T. R. (2001), Downweighting Influential Clusters in Surveys: Application to the 1990 Post Enumeration Survey.

Specific section

8. Purpose of the method

The module describes one estimation method frequently applied in business surveys, used to identify and handle outliers: Winsorisation. The general idea of Winsorisation is that if an observation exceeds a pre-set cut-off value, then the observation is replaced by that cut-off value or by a modified value closer to the cut-off value. As a result of Winsor estimation, we obtain a modified sample, in which untypical observations have been replaced with typical ones. The impact of outlying units is reduced. Further calculations are conducted for the modified sample. Any kind of estimation can be used at this stage.

9. Recommended use of the method

1. The method presented in this module is recommended for use in the case when the study variable(s) are highly skewed and several auxiliary variables that can be used to improve estimation including outliers. Such a situation is common in business statistics. The growing use of auxiliary information from administrative registers and the need to substantially reduce sample sizes or to produce more effective estimates has increased the importance of recognising and dealing with the data problem.
2. It is particularly suited to sample survey estimation. It can be used for various estimators (here, GREG estimation is illustrated) and sampling schemes.
3. In the case of stratified random sampling, the use of Winsor estimator can reduce the impact of outliers on stratum estimates while the population estimates remain unchanged (Rivest and Hidioglou, 2004).
4. The method is flexible because the cut-offs can be chosen to suit the situation. It is simple to implement for applications with multiple variables and estimates.

10. Possible disadvantages of the method

1. The one-sided Winsorisation can introduce a negative bias, which can result in inconsistent estimates.
2. The values modified by Winsor estimator are artificial and may sometimes be unacceptable.

11. Variants of the method

1. There are three main methods of dealing with outliers in a finite population (Cox et al., 1995):
 - reducing the weights of outliers (trimming weight);
 - changing the values of outliers (Winsorisation, trimming);
 - using robust estimation techniques such as M-estimation.

Winsorisation is most frequently used in business surveys. Two types of Winsorisation can be distinguished. The difference between them consists in the treatment of outliers.

Winsorised Type I estimator is based on an arbitrary assumption whereby any outliers exceeding a preset cut-off value K are always replaced by that value K .

In the case of Type II estimator, as weight \tilde{w}_i decreases, the contribution of outliers increases – the modified value of the study variable “approaches” the value of the outlier, i.e., the real value of the variable.

Winsorisation cut-offs can be chosen on different levels, e.g.:

- specifying a cut-off value for observations by stratum;
- specifying an individual cut-off value for each observation.

12. Input data

1. The input data set has to contain individual information for all units in the sample. The input data set can contain information coming from auxiliary sources, e.g., administrative register. Specific software (e.g., SAS) may be based on different structures of the input data set in the procedure of robust estimation.

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. Cut-offs.

15. Recommended use of the individual variants of the method

1. Surveys of very skewed populations which contain a few extreme values: surveys of business, agriculture, personal income and fortune.

16. Output data

1. Estimates of desired levels (target variable values after Winsorisation), quality measures for the estimates (e.g., variances, MSE).

17. Properties of the output data

1. The user should check the quality of estimates based on their knowledge of the investigated phenomenon, MSE, variance, bias of estimates.

18. Unit of input data suitable for the method

Information about variable of interest and auxiliary variables should be available for all units in the sample (sample level).

19. User interaction - not tool specific

The countermeasures against outliers can be divided into:

1. The detection of outliers – quantitative judgment, which requires an indicator of the degree of divergence of each data. Various methods of computing such indicators have been developed.
2. Outlier treatment:
 - “weight modification,” under which the weight of the sample unit is modified;
 - “value modification,” under which the value reported by the sample unit is modified;
 - the combination of the two, under which both the weight and the value reported by the sample are modified;
 - robust estimation techniques.

20. Logging indicators

1. Run time of the application.
2. Characteristics of the input data.
3. The number of units for which Winsorisation changed the values.

21. Quality indicators of the output data

1. MSE
2. Variance
3. Bias

22. Actual use of the method

1. *Survey of Employment, Payrolls, and Hours (SEPH)*, Statistics Canada: weight modification.
2. *State and Metro Area Employment, Hours, and Earnings*, Bureau of Labor Statistics America: reduces the impact of outliers through “weight reduction”.
3. *Consumer Price Index*, Australian Bureau of Statistics: Modifies the value of the outlier to the value next in size to the outlier through Winsorisation.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Statistical Data Editing – Main Module
2. Weighting and Estimation – Main Module

24. Related methods described in other modules

1. Weighting and Estimation – Calibration
2. Weighting and Estimation – Generalised Regression Estimator

25. Mathematical techniques used by the method described in this module

1. Regression
2. Ordinary Least Squares (OLS)
3. Trimmed least squares (TLS)
4. Trimmed least absolute value (LAV)
5. Sample Splitting
6. Least median of squares (LMS)

26. GSBPM phases where the method described in this module is used

1. 5.3 Review, validate, edit
2. 5.4 Impute
3. 5.6 Calculate weights
4. 5.7 Calculate aggregates

27. Tools that implement the method described in this module

1. Several popular statistical packages – including SAS, R, STATA, S-PLUS, LIMDEP, and E-Views – have procedures for robust regression analysis.
2. Least absolute deviations – SAS users call this procedure with the *LAV* command within the IML library. In STATA, median regression is performed with the quantile regression (qreg) procedure.
3. Least median of squares – SAS users can call least median of squares with the *LMS* call in *PROC IML*, S-Plus users can execute this algorithm with *lmsreg*.
4. Weighted least squares – SAS users call this procedure with the *LTS* command within the IML library.

28. Process step performed by the method

Estimation

Administrative section

29. Module code

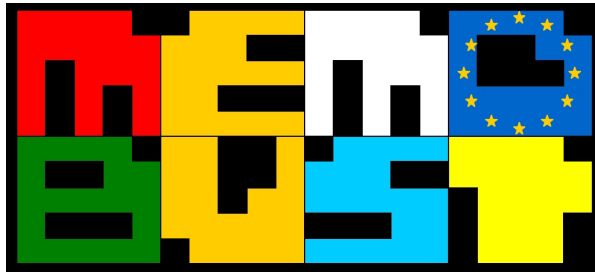
Weighting and Estimation-M-Outlier Treatment

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	29-02-2012	first version	Grazyna Dehnel	GUS
0.2	31-05-2012	second version	Grazyna Dehnel	GUS
0.3	10-12-2012	third version	Grazyna Dehnel	GUS
0.4	31-05-2013	fourth version	Grazyna Dehnel	GUS
0.5	09-09-2013	fifth version	Grazyna Dehnel	GUS
0.6	25-02-2014	sixth version	Grazyna Dehnel	GUS
0.6.1	11-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:32



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Preliminary Estimates with Design-Based Methods

Contents

General section.....	3
1. Summary	3
2. General description of the method	4
2.1 A design-based estimation method based on composite estimator	5
3. Preparatory phase	6
4. Examples – not tool specific.....	6
5. Examples – tool specific.....	6
6. Glossary.....	6
7. References	7
Specific section.....	8
Interconnections with other modules.....	9
Administrative section.....	10

General section

1. Summary

Timeliness is a particularly critical component of quality for producing short-term business statistics at the National Statistical Institutes (NSIs) of the European Community, as each Member State has to meet the standard quality requirements of the Regulation No 1165/98 – amended by the Regulation No 1158/2005 – about terms for transmission of the results and details of the information provided on statistical indicators, particularly on short-term statistics. The Amendment EU Regulation on Short Term Statistics requests all the statistical institutes of the EU Member States to transmit *preliminary short term* indicators to EUROSTAT with a reduced delay comparing to the timeliness set in the original 1998 Regulation (Eurostat, 2000, 2001, 2005). In OECD context, also, research projects were settled and useful documentation produced (Di Fonzo, 2005).

Frequently, in the NSIs short term statistics are based on fixed panel surveys of enterprises or rotating panels with a partial overlap from one year to another. More precisely, the amended regulation provides for a substantial improvement of timeliness for the production of the most important short-term indicators.

A common approach for dealing with *preliminary estimates* focuses essentially on the study and the definition of efficient estimators, exploiting almost exclusively auxiliary information in the estimation phase. Often preliminary estimation merely involves the use of the quick respondent units. In fact, in order to obtain “good” preliminary estimates, standard survey strategy often aims to achieve high quick response rate by means of a well-structured plan of follow-up. In some surveys the “largest” units are carefully supervised.

The main theoretical problem to be faced in a short-term preliminary estimation context concerns the possible self-selection of quick respondents that can lead to biased estimates of the unknown population mean and variances.

A useful documentation on preliminary estimation problems (even though not comprehensive) can be downloaded from the OECD web site¹.

This module focuses on estimation methods referring to the design-based approach. In particular describes a method proposed in Rao et al. (1989) which uses, to produce estimate referred to time t , data pertaining both to time t and $t-1$, with the aim to minimise the mean square error of the estimate.

Apart from this particular method, design-based (or model-assisted) estimation methods for preliminary estimates using quick respondents refer to the class of non-response weighting adjustment procedures, which are used in general when the Theoretical Sample (TS) is not achieved in practice in the Observed Sample (OS). In the case of preliminary estimates the observed sample coincides with the quick respondent set of units, available at the point of time when preliminary estimation has to be performed.

It is worth noting that in the context of preliminary estimate production the two most frequent situations at NSIs are: (i) using for the preliminary estimates the same estimator used for the final

¹ For the issue of the preliminary subsample the link is:

http://www.oecd.org/document/17/0,3746,en_2649_33715_30386193_1_1_1_1,00.html.

estimates computed on quick respondents, or (ii) referring to model-based estimators and ignoring the sampling design which generated data.

2. General description of the method

In general, the standard process going from data collection to elaboration of survey data needs to be accomplished within a fixed period of time, especially if the final estimates must be disseminated at a prefixed time point τ_f . In this context, direct estimators of the target parameters – based on the sampling units included in the TS and selected through a probabilistic sampling design – are design unbiased and consistent; the sampling error depends on the variability of the phenomenon under study, on the planned sample size and on the effectiveness of the selection procedure. Direct estimates based on the OS – that is a subset of quick respondents of TS with size depending on the nonresponse rate – can be biased in function of the random response process generating the OS.

We assign the term “preliminary” to the estimates computed using the statistical information available at a point of time τ_p preceding the time τ_f , on the basis of the OS denoted in this case as Preliminary Sample (PS), i.e., the sub-sample of *quick respondent units* that is available to be processed to produce the estimate at τ_p . The corresponding *final estimate* is based on a final sample, including both *quick* and *late respondents*, observed from τ_p and τ_f . The most straightforward practice in this situation is to apply the same estimation techniques utilised to produce the final estimates. Alternative estimation techniques (De Sandro and Gismondi, 2004; D’Alò et al., 2007) should take under control the *bias* and the *revision error*, given by the difference between final and preliminary estimates. In order to test the quality of the preliminary estimator, the revision error should be evaluated for different survey occasions.

Some indicators of the revision error can be defined and compared on the basis of the time series of provisional estimates and final ones. Among them, the following indicators can be evaluated.

- *Average total revision*, that is the average of the difference between the latest available value and the first release for each observation period. This measure indicates a possible bias of the first release.
- *Average absolute revision*, that is the average of the absolute difference between the latest available value and the first release for each observation period, regardless of their sign. This measure indicates the stability of the first release.
- *Range of total revisions*. Highest and lowest total revisions to the first release for all observation periods. This range indicates the volatility of the first release. The total range covers all the revisions and may include outliers.

Preliminary estimation methods may be classified in function of the stage on which specific preliminary methods are applied. In fact, it is possible to identify methods which act:

- at the sampling design stage, by selecting a preliminary subsample of TS (cf. the module “Sample Selection – Subsampling for Preliminary Estimates”);
- at the estimation stage, in the following ways:
 1. by means of imputation techniques of missing data, that are applied to the non-respondent units in TS but not in PS (cf. the topic “Imputation”);

2. by means of weighting adjustment, i.e., calculating nonresponse correction factor when early respondents are used in the standard estimator, the same adopted for the final estimate, modifying the sampling weights assigned to the units in PS in order to take into account non respondents in TS;
3. by applying direct and indirect estimators, using known population totals of auxiliary variables and/or time series of preliminary and final estimates of the variable of interest.

The estimation of individual response probabilities – useful to modify sampling weights of the ordinary Horvitz-Thompson estimator – is quite difficult because of randomness of some nonresponse and the lack of enough reliable auxiliary variables (Rizzo et al., 1996). Imputation techniques render easier the estimation process, but normally do not reduce bias because they are founded on data concerning respondent units only. These evidences stressed a wider recourse to a model-based approach, as remarked in Särndal et al. (1993), Valiant et al. (2000), Särndal and Lundström (2005).

In the model-assisted approach, weighting may be based on a *calibration approach*. A *calibrated weight* is obtained by the multiplication of the *direct or design weight* – defined as the reciprocal of the inclusion probability – with a *correction factor*. The correction factor is a nonresponse adjustment weight that attempt to compensate for unit nonresponse. A commonly used procedure for obtaining these weights is to divide the total sample into a set of weighting classes based on information known for both respondents and non-respondents and then to increase the base weights for the respondents in a weighting class to represent the non-respondents in that class (Kalton, 1983; Särndal and Lundström, 2005). Several methods to define adjustment cells are presented in literature (Rizzo et al., 1996; Eltinge and Yansaneh, 1997; Breiman et al., 1984; Little, 1986).

Depending on the informative context, the totals used for calibration may: (i) be known at population level; (ii) be estimated - using expansion weights – by the TS units or (iii) represent the final estimates of previous survey occasions. In order to reduce the bias, the auxiliary variables should explain both the main study variables and the inverse response probability.

In some survey there is an extensive amount of information available for the non-respondents. This information may derive from the sampling frame or by matching sampled elements with administrative records. Besides, in panel surveys and other surveys involving more than one wave of data collection, extensive information of non-respondents at later waves is available from their responses at early waves.

It is useful to underline, finally, that when the target variables are dependent on the provisional response mechanism, the preliminary estimates may be affected by some bias.

2.1 *A design-based estimation method based on composite estimator*

For this method the approach is based on a probabilistic design as both the theoretical sample and the observed quick respondent sample are considered as generated by a random design. The expected value $E(.)$ and the variance $V(.)$ of the estimators are considered with respect to these sampling designs. Furthermore, a random mechanism of nonresponse is supposed to generate the anticipated sample.

In this context, Rao et al. (1989) proposed the *composite estimators* that may represent an improvement of *the standard estimator*.

Generally speaking, the basic composite estimator is obtained as weighted average of the *preliminary estimate* for time t and the *final estimate* of time $t-1$ adjusted for the difference between preliminary estimates referred to t and $t-1$.

For the estimate of a population total y_t , let Y_t^p , Y_t and $Y_t^* = Y_t - Y_t^p$ be respectively the preliminary estimate, the final estimates and the measurement errors in preliminary estimates at time t , $t = 1, \dots, T$. The proposed composite estimator is:

$$Y_{t,\alpha} = \alpha Y_t^p + (1 - \alpha)[Y_{t-1} + (Y_t^p - Y_{t-1}^p)],$$

being α a weight varying between 0 and 1.

To determine the “optimal” α , i.e., that assuring minimum variance, some reasonable assumptions are made:

a1: $E(Y_t^p - Y_{t-1}^p) = E(Y_t - Y_{t-1})$, $E(.)$ denoting the expected value,

a2: $|B(Y_t^p)| \geq |B(Y_t)|$, $i=t, t-1$ and $B(.)$ denoting the bias; furthermore, it is assumed for simplicity that $B(Y_t) = 0$ and $B(Y_t^p) = \delta$.

Then we get

$$\alpha = [V(Y_{t-1} - Y_{t-1}^p) + Cov(Y_t^p, Y_{t-1}) - Cov(Y_t^p, Y_{t-1}^p)] [V(Y_{t-1} - Y_{t-1}^p) + \delta^2]^{-1}.$$

In a similar way the optimal α for the composite estimator for change $(y_t - y_{t-1})$ is shown in Rao et al. (1989), where the impact of the size of δ on the equivalence of using the optimal α for estimate level or change is discussed as well.

Variance and covariance terms can be estimated on survey data using usual formulas. The paper by Rao et al. (1989) introduces a further assumption about covariances which allows to simplify the expression for α ; this assumption, anyway, is valid when bias in preliminary estimates is due to undercoverage but not when it is due to nonresponse.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*. Chapman and Hall, New York.
- D'Alò, M., De Vitiis, C., Falorsi, S., Righi, P., and Gismondi, R. (2007), Sampling Strategies for Preliminary Estimates Production in Short-Term Business Surveys. *Proceedings of the 2007 intermediate conference Risk and prediction*, Società Italiana di Statistica.
- De Sandro, L. and Gismondi, R. (2004), Provisional Estimation of the Italian Monthly Retail Trade Index. *Contributi-Istat*, 24/2004.
- Deville, J.-C. and Tillé, Y. (2004), Efficient Balanced Sampling: the Cube Method. *Biometrika* **91**, 893–912.
- Di Fonzo, T. (2005). The OECD project on revisions analysis: First elements for discussion. Paper presented at OECD STESEG meeting, Paris, 27-28 June 2005.
<http://www.oecd.org/dataoecd/55/17/35010765.pdf>
- Eltinge, L. and Yansaneh, I. S. (1997), Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology* **23**, 33–40.
- EUROSTAT (2000), *Short-term Statistics Manual*. Eurostat, Luxembourg.
- EUROSTAT (2001), Conclusion of the First Meeting of the Expert Group Contro-Stratified European Sample for Retail Trade, Final Report, July 2001. Eurostat, Luxembourg.
- EUROSTAT (2005), *Council Regulation No 1165/98 Amended by the Regulation No 1158/2005 of the European Parliament and of the Council – Unofficial Consolidated Version*. Eurostat, Luxembourg.
- Little, R. J. A. (1986), Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review* **54**, 139–157.
- Kalton, G. (1983), Compensating for Missing Survey Data. Survey Research Center, University of Michigan, Ann Arbor, MI.
- OECD, Short-Term Economic Statistics (STES) Timeliness Framework.
<http://www.oecd.org/std/short-termeconomicstatisticsstestimelinessframework.htm>
- Rao, J. N. K., Srinath, K. P., and Quenneville, B. (1989), Estimation of Level and Change using Current Preliminary Data. In: Kasprzyk, Duncan, Kalton, and Singh (eds.), *Panel Surveys*, John Wiley & Sons, New York, 457–485.
- Rizzo, L., Kalton, G., and Brick, M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse. *Survey Methodology* **22**, 43–53.
- Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer Verlag.
- Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley & Sons, New York.

Specific section

8. Purpose of the method

The method is used for the preliminary estimation of the target variable, with the aim to obtain the estimates relying on statistical information available at time preceding the time t , i.e., on the basis of only a set of quick respondents which define the so-called preliminary sample.

9. Recommended use of the method

1. When model-based method cannot be used because auxiliary variables are not available or the time series is not long enough.

10. Possible disadvantages of the method

1. The improvement of the revision error can be weak.

11. Variants of the method

- 1.

12. Input data

1. Final estimates of preceding time $t-1$ and standard preliminary estimates at time t .

13. Logical preconditions

1. Missing values
 1. Not applicable.
2. Erroneous values
 1. Not applicable.
3. Other quality related preconditions
 1. Not applicable.
4. Other types of preconditions
 1. Not applicable.

14. Tuning parameters

1. Alpha (α) to be evaluated on the basis of variances and covariances.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. Ds-output1 = composite preliminary estimates of the target parameter.

17. Properties of the output data

1. The composite preliminary estimates should guarantee a lower revision error than the direct estimates.

18. Unit of input data suitable for the method

Preliminary and Final Estimates at previous time and preliminary estimates at time t .

19. User interaction - not tool specific

- 1.

20. Logging indicators

- 1.

21. Quality indicators of the output data

1. Revision errors.
2. Quality assessment of the result.

22. Actual use of the method

1. None.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Imputation – Main Module

24. Related methods described in other modules

1. Sample Selection – Subsampling for Preliminary Estimates
2. Weighting and Estimation – Preliminary Estimates with Model-Based Methods

25. Mathematical techniques used by the method described in this module

1. Variance-Covariance estimation

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate aggregates

27. Tools that implement the method described in this module

1. No software tools are available

28. Process step performed by the method

Estimation of target parameters on the basis of information collected on quick respondents.

Administrative section

29. Module code

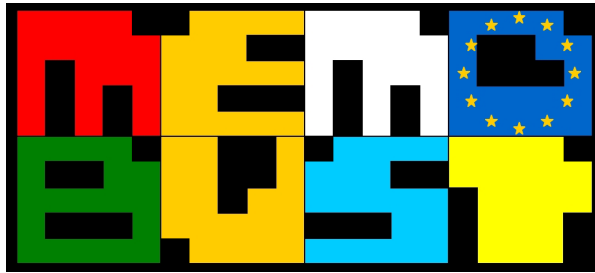
Weighting and Estimation-M-Preliminary Estimates Design-Based

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	31-12-2012	first version	Claudia De Vitiis	ISTAT
0.2	11-02-2013	first revisions	Claudia De Vitiis	ISTAT
0.3	30-09-2013	revised according to review by Norway	Claudia De Vitiis	ISTAT
0.3.1	07-11-2013	revised according to review by Editorial Board	Claudia De Vitiis	ISTAT
0.3.2	13-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:33



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Preliminary Estimates with Model-Based Methods

Contents

General section.....	3
1. Summary	3
2. General description of the method	4
2.1 Particular cases and extensions	5
3. Preparatory phase	5
4. Examples – not tool specific.....	5
5. Examples – tool specific.....	5
6. Glossary.....	5
7. References	5
Specific section.....	7
Interconnections with other modules.....	9
Administrative section.....	10

General section

1. Summary

For each survey, the standard process from collection to elaboration of survey data needs to be accomplished within a fixed period of time, i.e., the final estimates must be disseminated at the prefixed time t . In this context, direct estimators of the target parameters – based on the sampling units included in the Theoretical Sample (TS), selected by a probabilistic sampling design – are design unbiased and consistent; the sampling error depends on the variability of the phenomenon under study, on the planned sample size and on the effectiveness of the selection procedure. Direct estimators based on the Observed Sample (OS) – that is a subset of TS whose size depends on the total nonresponse rate – can be biased in function of the response process generating the OS.

We assign the term “preliminary” at the estimates computed using the statistical information available at time preceding the time t , on the basis of the OS denoted as Preliminary Sample (PS). The most straightforward practice in this situation is to apply the same estimation techniques utilised to produce the final estimates. Alternative estimation techniques should take under control the bias and the revision error, given by the difference between final and preliminary estimates. In order to test the quality of the preliminary estimator, the revision error should be evaluated for different survey occasions.

The main theoretical problem to be faced in a short-term preliminary estimation context concerns the possible self-selection of quick respondents, that can lead to biased estimates of the unknown population mean and variances. In the context of short-term business surveys – usually planned for estimating parameters such as indexes and their changes over time – one common method is based on the evaluation, for each design stratum, of the direct estimator of the index imputing the missing responses for the sampling units belonging to TS. Another type of procedure utilises the direct estimates of the design stratum indexes without imputation of the missing responses both in OS and in PS. These approaches can be based on imputation methods supposing no systematic differences between early and late respondents.

Preliminary estimation methods may be classified in function of the stage on which specific preliminary methods are applied. In fact, it is possible to identify methods that are acting:

- at the sampling design stage, by selecting a preliminary subsample of TS;
- at the estimation stage, in the following ways:
 1. by means of imputation techniques of missing data, that are applied to the non-respondent units in TS but not in PS;
 2. by means of weighting adjustment, i.e., modifying the sampling weights assigned to the units in PS in order to take into account non respondents in TS;
 3. by applying direct and indirect estimators, using known population totals of auxiliary variables and/or time series of preliminary and final estimates of the variable of interest.

The techniques based on the selection of a preliminary sample and the methods requiring imputation and weighting adjustment are generally based on unit level models. These models use disaggregated auxiliary information coming from survey data at previous times and/or administrative register data. For the methods in the last class the relation between the variable of interest and the auxiliary variables is usually formalised through domain level models in which the auxiliary information is expressed in terms of domain known totals or estimates. In the last class fall an estimation technique developed by Rao et al. (1989) in which preliminary estimates are computed assuming AR(1) models for final estimates and the revision error. This is the main specific model-based procedure used for the computation of preliminary estimation and it is described in this module.

2. General description of the method

In the context of a given sampling survey we mean as *preliminary estimate* the estimation of a parameter of interest obtained on the basis of a sub-sample of *quick respondent units* that is available within a time lag after the reference time point t (or end of the reference period) of the survey, while the correspondent *final estimate* is based on a final sample, including both *quick* and *late respondents*, observed within a time lag. The indicators measuring the statistical quality of a *preliminary estimation method* are based on the differences, evaluated at the different times (identifying the correspondent survey occasions) between preliminary estimates obtained by means of the method under study, and the corresponding final estimates. These differences are known as *revision errors*.

In this context, Rao et al. (1989) adopt a time series approach: let Y_t^P , Y_t and $Y_t^* = Y_t - Y_t^P$ be respectively the preliminary estimate at time t , the final estimates and the measurement errors in preliminary estimates at time t , $t = 1, \dots, T$. Furthermore, Y_t and Y_t^* are supposed to follow an AR(1) process:

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \quad (1)$$

$$Y_t^* = \psi Y_{t-1}^* + \zeta_t, \quad \zeta_t \sim N(0, \sigma_0^2) \quad (2)$$

ε_t and ζ_t assumed to be independent, while ϕ and ψ are autocorrelation coefficients ranging between -1 and 1.

Rewriting models (1) and (2) in state space form and ignoring sampling errors, they obtain the following final preliminary estimate for the period $t+1$ by means of Kalman filter (see for instance Harvey, 1984)

$$\hat{Y}_{t+1} = \alpha(\phi Y_t) + (1 - \alpha)(Y_{t+1}^P - \psi Y_t^*), \quad (3)$$

where $\alpha = \sigma^2 / \sigma_0^2$.

The preliminary estimate (3) can be viewed as a weighted average of the final estimate of the previous period t and the preliminary estimate for time $t+1$ adjusted for the previous measurement error.

Two alternative ways to obtain starting preliminary estimates Y_t^P can be used and both of them will be described in the next section.

2.1 Particular cases and extensions

Whenever any auxiliary information is available at current time, an extension of the basic previous method is possible by introducing in the model the auxiliary information correlated with the target variable of interest. In that case, the AR(1) model assumed for the final estimates (1) can be generalised in the following way:

$$Y_t = \phi Y_{t-1} + \sum_{k=1}^P \beta_k X_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \quad (4)$$

Another extension can be obtained by introducing other previous estimates that can be considered highly correlated with the estimates at current time. For example, in the case of monthly estimates it is reasonable to assume that the final estimates at time t is both correlated to the final estimate at time $t-1$ and with the final estimate at time $t-12$. In this case, the following model can be assumed:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-12} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \quad (5)$$

This model is known in the literature as a seasonal autoregressive model (see Choi and Varian, 2009).

A mixed extension of model (4) and (5) can be further considered, assuming the following model:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-12} + \sum_{k=1}^P \beta_k X_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \quad (6)$$

Moreover dummy variables can be introduced into the previous model whenever specific domains estimations are required.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Bolfarine, H. and Zacks, S. (1992), *Prediction Theory for Finite Populations*. Springer-Verlag, New York.

D’Alò, M., Gismondi, R., Solari, F., and Naccarato, A. (2006), Estimation in Repeated Business Surveys using Preliminary Sample Data. *Atti della XLIII Riunione Scientifica SIS, 14-16 giugno, Torino*.

Choi, H. and Varian, H. (2009), Predicting the Present with Google Trends. Draft report Google Inc.

- De Sandro, L. and Gismondi, R. (2004), Provisional Estimation of the Italian Monthly Retail Trade Index. *Contributi-Istat*, 24/2004.
- Harvey, A. C. (1984), Dynamic Models, the Prediction Error Decomposition and State-space. In: D. F. Hendry and K. F. Wallis (eds.), *Econometrics and Quantitative Economics*, Blackwell, Oxford, 37–59.
- Lamberti, A., Naccarato, A., and Pallara, A. (2004), Improving Timeliness of Short-term Business Statistics through State-space Modelling of Preliminary Survey Data. *Proceedings of the European Conference on Quality and Methodology in Official Statistics*.
http://q2004.destatis.de/download/Table-of-contents_Programme.pdf.
- Matei, A. and Ranalli, M. G. (2011), Adjusting for nonignorable nonresponse using a latent variable modeling approach. *Proceedings of the LVIII Meeting of the ISI, Dublin, 21-26 August 2011*.
http://www.stat.unipg.it/~giovanna/papers/AMatei_GRanalli_isi11.pdf
- OECD, Short-Term Economic Statistics (STES) Timeliness Framework.
<http://www.oecd.org/std/short-termeconomicstatisticsstestimelinessframework.htm>
- Papageorgiou, H. and Vardaki, M. (2002), Trade-off between Timeliness and Accuracy. Doc.Eurostat/A4/Quality/02/Timeliness/Timeliness and Accuracy.
- Rao, J. N. K., Srinath, K. P., and Quenneville, B. (1989), Estimation of Level and Change using Current Preliminary Data. In: Kasprzyk, Duncan, Kalton, and Singh (eds.), *Panel Surveys*, John Wiley & Sons, New York, 457–485.
- Royall, R. M. (1992), Robustness and Optimal Design Under Prediction Models for Finite Populations. *Survey Methodology* **18**, 179–185.
- Tam, S. M. (1987), Analysis of Repeated Surveys Using a Dynamic Linear Model. *International Statistical Review* **55**, 63–73.
- Valliant, R. (1999), Uses of Models in the Estimation of Price Indexes: a Review. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Specific section

8. Purpose of the method

The method is used for the preliminary estimation of the target variable, with the aim to obtain the estimates relying on statistical information available at time preceding the time t , i.e., on the basis of only a set of quick respondents which define the so-called preliminary sample.

9. Recommended use of the method

1. The time series of final and preliminary estimates should be long enough.

10. Possible disadvantages of the method

1. When the time series of preliminary and final estimates is short the estimation of model parameters can be very unstable.

11. Variants of the method

1. As an alternative to Rao et al. (1989) a time series methods based on the treatment of unit non-response may be applied. In this case, the late response is treated as nonresponse but in order to avoid biased estimates, the self-selection of quick respondents mechanism should not be considered as totally random. Difference between early and late respondents must be considered. In this framework, it could be interesting describing the process of self-selection of quick respondent by means of a latent variables, which can be interpreted as the ability to respond quickly. This predisposition can be used in order to estimates the quick response probabilities using a logistic model, allowing at the same time to deal with the non-ignorable auto-selection of preliminary respondent can be treated allowing . For more detail on the method see Matei and Ranalli (2010).
2. The preliminary estimation may be treated also using the dynamic linear model. For detailed information on these type of models see Harvey (1984) and Tam (1987). An application of this type of modelling in the context of preliminary estimation can be found in Lamberti et al. (2004).
3. Finally, some small area methods, for instance synthetic type estimator and modified GREG can be easily adapted also for the preliminary estimation, especially when the preliminary sample size is too small for computing reliable estimates.

12. Input data

1. Time series of final estimates.
2. Time series of final estimates of the revision errors.

13. Logical preconditions

1. Missing values
 1. Not applicable.
2. Erroneous values

- 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. Not applicable.

15. Recommended use of the individual variants of the method

1. The variants of the method can be used when additional auxiliary information or correlated estimates are available.

16. Output data

1. Ds-output1 = preliminary estimates of the target parameter.

17. Properties of the output data

1. The model-based preliminary estimates should guarantee a lower revision error than the direct estimates.

18. Unit of input data suitable for the method

Time series of Preliminary and Final Estimates at previous times.

19. User interaction - not tool specific

1. Choice of auxiliary covariates and/or estimates.

20. Logging indicators

1. Not applicable.

21. Quality indicators of the output data

1. Time series of revision errors.
2. Quality assessment of the result.
3. Model diagnostics to evaluate the model fitting when model-based estimators are applied.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

- 1.

24. Related methods described in other modules

1. Weighting and Estimation – Preliminary Estimates with Design-Based Methods

25. Mathematical techniques used by the method described in this module

1. Matrix algebra
2. Kalman filter

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate aggregates

27. Tools that implement the method described in this module

1. No software tools are still available.

28. Process step performed by the method

Estimation of target parameters on the basis of information collected on quick respondents.

Administrative section

29. Module code

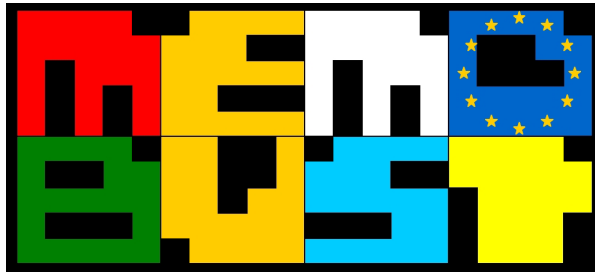
Weighting and Estimation-M-Preliminary Estimates Model-Based

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	01-03-2012	first version	Michele D'Alò, Claudia De Vitiis	ISTAT
0.2	26-06-2012	second version	Michele D'Alò, Claudia De Vitiis	ISTAT
0.3	30-09-2013	final version	Michele D'Alò, Claudia De Vitiis	ISTAT
0.3.1	13-11-2013	revisions based on review by Editorial Board	Michele D'Alò, Claudia De Vitiis	ISTAT
0.3.2	13-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:33



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Small Area Estimation

Contents

General section.....	3
1. Summary	3
2. General description.....	3
3. Design issues	9
4. Available software tools	9
5. Decision tree of methods	9
6. Glossary.....	11
7. References	11
Interconnections with other modules.....	14
Administrative section.....	15

General section

1. Summary

Business surveys carried out by National Statistical Institutes are usually aimed to obtain estimates of target parameters, e.g., the overall amount of industrial turnover for the whole population of business enterprises. Analogous parameters are usually defined with respect to relevant population sub-sets, i.e., sub-populations corresponding to geographical partitions (e.g., administrative areas) or sub-populations associated to economic cross-classification (e.g., enterprise size and sector of activity). An example is given by the estimation of the industrial turnover for each administrative region (e.g., NUTS2 level), or for each sector of activity (e.g., 2-digit NACE). An estimator of the parameter of interest for a given sub-population is said to be a *direct estimator* when it is based only on sample information from the sub-population itself. Unfortunately, for most of surveys the sample size is not large enough to guarantee reliable direct estimates for all the sub-populations. A ‘small area’ or ‘small domain’ is any sub-population for which a direct estimator with the required precision is not available. Even though the term ‘small domain’ may seem to be proper in the business survey context, ‘small area’ is intended in the literature as a general concept and it is used to indicate a general partition of the population according to geographical criteria or other structural characteristics (socio-demographic variables for household surveys or economic variables for business surveys). In the following we will utilise preferably the term small domain but the term small area will be used too in its wide and meaningful definition.

When direct estimates cannot be disseminated because of unsatisfactory quality, an ad hoc class of methods, called small area estimation (SAE) methods, is available to overcome the problem. These methods are usually referred as *indirect estimators* since they cope with poor information for each domain borrowing strength from the sample information belonging to other domains, resulting in increasing the effective sample size for each small area.

2. General description

Sampling designs for business surveys are usually stratified one stage designs, where strata are defined as the cross-classification of structural characteristics of the enterprises as geographical area, economic activity, size in terms of number of workers, etc. Planned domains of interest are usually given by the different sets of marginal strata. In this context small domains are defined as *planned domains* when they are obtained as strata or aggregation of strata. Furthermore small domains are defined as *unplanned domains* when they cut across strata.

This situation is showed in Figure 1, where domains of interest are geographical areas. The example is referred to a one stage stratified sampling design, in which h is the generic stratum ($h = 1, \dots, H$) and the dots are the sampling units. The figure shows, three different types of small areas that can be potentially encountered:

- the first type, denoted by d , is an example of unplanned small area, being the union of complete and incomplete strata. Then the corresponding sample size is a random variable;
- the second kind of small area, denoted by d' , is a special case of unplanned small area when no sample units are selected in the target small area;

- finally the third type, denoted by d'' , is an example of planned small area, being the union of complete strata.

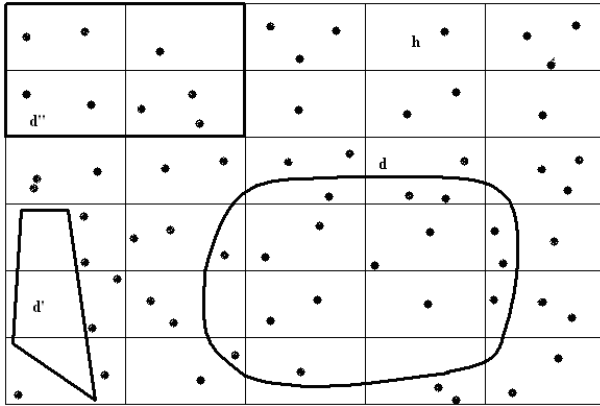


Figure 1. Different types of small areas.

Direct estimators, that are obtained within the design-based approach, may produce reliable domain estimates of the target parameters only when the domain sample sizes are sufficiently large. When the realised domain sample sizes are not large enough to guarantee reliable direct domain estimates, indirect methods provide tools to overcome the problem. The main idea underlying these techniques is to increase the effective sample size for each domain by means of the information from the units belonging to other domains considered “similar” (with respect to structural characteristics) to the small domain of interest. The set of all domains from which estimation methods borrow strength will be referred as *broad domain*. For instance in figure 1 the broad domain may be given by the union of all the H strata defining the largest rectangle. The more straightforward way to borrow strength is given by the *synthetic estimator*. According to Gonzalez (1973) an estimator is called a synthetic estimator if a reliable direct estimator for a large area (i.e., broad domain), covering several small areas, is used to provide small area estimates under the assumption that all the small areas have the same characteristics as the large area. Synthetic estimators increasing the effective sample size result in smaller variances than direct estimators. On the other hand, bias can seriously affect synthetic estimators, since they make too strong use of information from other areas allowing too little for local variation (overshrinkage). In order to balance the potential bias of a synthetic estimator against the instability of a direct estimator is to take a weighted average of the two estimators. The resulting estimators are known as *composite estimators* (Schaible, 1979). Synthetic and composite estimators are usually referred as *indirect methods*.

It is useful to consider the following classification of direct estimators according the use of auxiliary population information:

- no use of auxiliary population information, corresponding to Horvitz-Thompson (H-T) estimator (Horvitz and Thompson, 1952; Cochran, 1977);
- use of auxiliary population information, these methods improve the efficiency of H-T estimator by means of unit level auxiliary information (observed for each respondent unit) and the corresponding known population totals or means. This class of estimators may be further divided

in methods using: *Domain Specific level* (DS) auxiliary population information and *Aggregated Domain level* (AD) auxiliary population information. The former refers to auxiliary population information available for each small domain, the latter is related to the case of auxiliary population information for aggregations of two or more domains.

Almost all large scale business surveys use direct estimators exploiting auxiliary population information, such as Generalised regression estimator (GREG) or more in general Calibration estimator. Calibration estimator satisfies constraints entailing the equivalence between known auxiliary variables population totals, or means, and the corresponding calibrated estimates. Calibration weights are derived minimising a distance between survey and calibration weights. Deville and Särndal (1992) showed that GREG estimator is a particular case of Calibration estimator under the chi-square distance. For both DS and AD information, it is possible to obtain some well-known special cases of GREG estimator, e.g., Ratio, Post-stratified, Post-stratified Ratio and Ratio-raking estimators, that are broadly used in large scale surveys.

GREG estimator, obtained under the model-assisted framework, allows to define approximately unbiased, and in many cases consistent, direct estimators, exploiting the correlation between the target variable and a set of covariates. A linear fixed model is defined to obtain a reduction of design variance of H-T estimator.

In the case of AD auxiliary population information, Generalised Regression Estimator, $GREG_{AD}$, is approximately unbiased if the overall sample size is large enough, but consistency is obtained only under a large expected domain sample size. Note that, under AD auxiliary information, residuals are different to zero for all units belonging to the sample, then large negative residuals for all the sampled units not belonging to domain d can produce inefficiency.

When DS auxiliary population information is used, the corresponding Generalised Regression Estimator, denoted with $GREG_{DS}$, is approximately unbiased only if the domain sample size is sufficiently large. For $GREG_{DS}$, unlike $GREG_{AD}$, the sample residuals of the units outside domain d are null. Therefore $GREG_{DS}$ can be more efficient than $GREG_{AD}$.

An approximately unbiased direct estimator that may overcome the problems related to the above GREG estimators is known as Modified Direct (MD) estimator. It is equal to $GREG_{DS}$ estimator, but $GREG_{AD}$ regression coefficient vector is used. Then MD estimator borrows strength for estimating the regression coefficients but does not increase the effective sample size as indirect estimators. It is approximately unbiased as the overall sample size increases, also when the domain sample size is small. Note that, like $GREG_{DS}$, residuals are null outside the target domain. Then when the DS and AD regression coefficients are close each other, the MD estimator may results more efficient than $GREG_{DS}$. For more details on the above estimators, see Rao (2003).

SAE methods are characterised by the different ways to borrow strength from information other than the observed values of the target variable in each small domain.

Figure 2, taken from Elazar (2005), synthesises well the different approaches in borrowing strength:

- (a) cross-sectional way;
- (b) using auxiliary data;
- (c) exploiting spatial relationship;

(d) using over time relationship.

The simplest way to borrow strength is using the values assumed by the target variable in all the domains included in the broad domain. This implies assuming that all the domains have a common mean value of the target variable (see case (a) in figure 2). If it is possible to divide the population in sub-groups according to one or more auxiliary information, the following step is to assume common mean values for all the domains within each sub-group. This is a particular case of assuming linear relationship between the target variable and a set of covariates (see case (b) in figure 2). It must be underlined that in case (a) only small domain population counts are needed, while in case (b) users must know small domain population counts for each sub-group when dealing with categorical auxiliary variables or small domains population means when using quantitative variables. In both cases small domains play a symmetric role and have the same importance in the estimation process. Enhanced methods are involved when using spatial or temporal information. Case (c) in figure 2 is related to the case of using spatial information in the estimation process. The main idea is that units belonging to the closest geographical areas should be given more importance in the borrowing strength process. This implies the need of additional information such as distance or neighbourhood matrices among the domains. When small domains are not related to geographical areas, like frequently happen in business surveys, it may be difficult to identify appropriate distance or neighbourhood concepts for domains. The last way to borrow strength from other sources of data may be applied in case of repeated surveys, that is when several survey occasions are available. In this case it would be possible to use the information from the previous survey occasions or times.

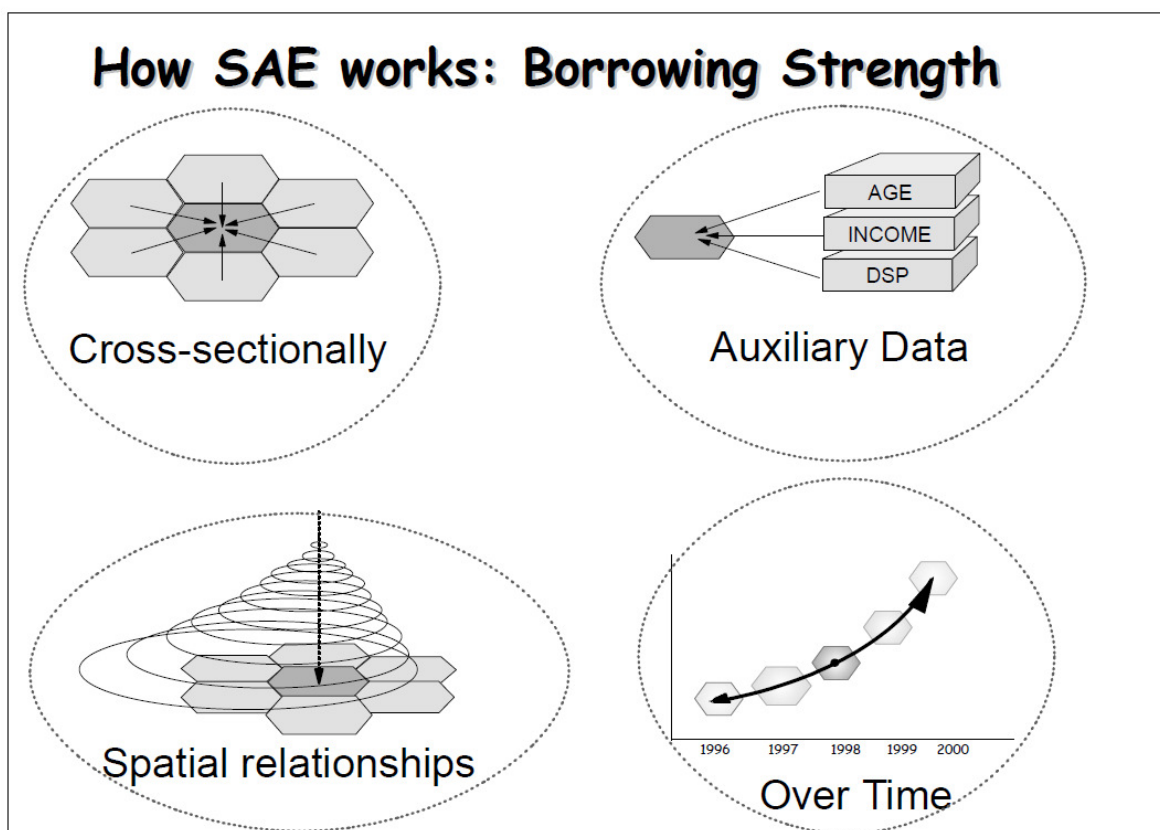


Figure 2. How to borrow strength: (a) Cross-sectionally, (b) using auxiliary data, (c) exploiting spatial relationship, (d) using over time relationship.

It is worthwhile to underline that the four approaches described above can be combined together defining in this way the complete set of SAE methods. In fact methods belonging to (a) or (d) can be used also in combination with (b) and/or (c). For example methods involving spatial correlation between the areas can also use auxiliary information, or methods to be applied when repeated survey data are available can also exploit spatial correlation and the information coming from auxiliary variables. Note that SAE methods in (a), (b) and (c) increase the effective domain sample size exploiting all the sampling information coming from the units belonging to the broad domain. SAE methods related to case (d) increase the domain sample size using the sampling information coming from the units observed from previous survey occasions, within the target domain. The joint use of cross-sectional and temporal information is possible too, e.g., using SAE methods related to cases (a) and (d). These techniques lead to a further increase of the effective sample size.

On the basis of the above description, it is useful to propose a classification of SAE methods. The small area estimators are divided into three groups according to the way they use the sampling and population information:

- (1) methods involving *spatial smoothing*, using data of all the small domains for only one survey time;
- (2) methods involving *temporal smoothing*, using data for only the small domain of interest for several survey occasions;
- (3) methods involving *spatial and temporal smoothing*, using data collected for all the domains at different survey times.

The three classes of methods can be further divided according to the inferential approach: *design-based* (d) or *model-based* (m) approach. In the first approach the target parameters are considered as unknown but fixed quantities while in the second one they are dealt as random variables and inference is based on the definition of an explicit model. The model formalises the relationship between data from several small domains within a broad domain, and/or the link between different survey occasions. Model specification involves extra auxiliary information correlated with the target variable, from census or administrative registers. In order to take into account simultaneously the previous classifications, the notation (d) and (m) will be combined with the indexes (1), (2), (3) denoting three different classes of smoothing, e.g., (d.1) will denote design-based methods involving spatial smoothing, (d.2) design-based methods with temporal smoothing, and finally (m.3) will indicate model-based methods using spatial and temporal smoothing.

As far as the case (d) is concerned we have:

- (d.1) the so called traditional methods, that is synthetic and composite estimators (see respectively Gonzalez, 1973 and Schaible, 1979). Particular cases of design-based composite estimators are the Sample-size dependent estimator (Drew et al., 1982), the James-Stein estimator (see Rao, 2003).
- (d.2) the methods for which it is possible to assume some time dependent correlation among direct estimators. For repeated surveys based on rotated samples, direct estimators can be suitably combined with a gross change estimator based on the common units in two consecutive samples. This provides additional information allowing to improve the efficiency of the estimator at each time. The original idea by Jessen (1942) and Patterson (1950), was improved using a multivariate framework by Gurney and Daly (1965). They introduced the concept of elementary estimator

related to each rotation group. The elementary estimators have been utilised for linear models, which make use of the correlation structure among the estimators to produce Minimum Variance Linear Unbiased Estimators (MVLUE). In practice, the specification and the inversion of the error correlation matrix may result in unstable estimates. One possible way to overcome this problem was suggested by Gurney and Daly (1965), who defined the class of composite estimators which combine the results of two consecutive samples in order to obtain actual estimates.

- (d.3) this class includes either the Gurney and Daly estimator for the case with more than one small area and the estimator proposed by Purcell and Kish (1980), known as SPREE (Structure Preserving Relation Estimator), for categorical data. This is based on the definition of two structures of data. The first is given by the complete population data related to a previous time. This is used to draw for each small area the associative structure information about the link between the target variable and a set of auxiliary variables (complete contingency table). The second source of information is the allocative structure, that is a set of current estimates for some marginal tables. Estimates preserve the observed relationships in the original associative structure except those specified in the allocative structure.

In the model-based approach models are explicitly defined and inference is drawn not anymore from the sample space but from a model on the population values (super-population model). Depending on the level at which the information is specified, area level or unit level models can be specified. In the former the link between target and auxiliary variables is defined for each area, while in the latter the relationship is specified for each unit. The more common methods are Empirical Best Linear Unbiased Predictor (EBLUP), Empirical Bayesian (EB) predictor and Hierarchical Bayesian (HB) predictor. Almost all these methods are based on multilevel models in which one level of the hierarchy is specified at area level. In details these models reduce to linear and generalised linear mixed models in the frequentist approach, and to hierarchical models in the Bayesian framework. Bayesian modelling implies the specification of priors distributions for all the parameters in the model. A regression function with respect to a set of auxiliary variables is introduced. This is usually referred in the frequentist framework as the fixed part of the model and the regression coefficients are indicated as the fixed effects. Moreover to consider the extra-variability not explained by the fixed part of the model, random effects related to each domain are added to the model. On the contrary if area random effects are not included into the model, synthetic model-based estimates are obtained instead of composite model-based estimates. For an extensive overview readers can refer to Rao (2003).

Three classes of model-based SAE methods can be defined:

- (m.1) methods using spatial smoothing. Seminar papers in this context are Fay and Herriot (1979) for area level models, Battese et al. (1988) for unit level models, Morris (1983) for the EB approach, and Datta and Ghosh (1991) and Ghosh (1992) for the HB modelling. Model specifications taking into account spatial correlation of the area random effects are proposed by Cressie (1991), Saei and Chambers (2003), and Pratesi and Salvati (2008);
- (m.2) methods utilising temporal smoothing. Not considering the pioneering works by Scott (1974) and Smith and Jones (1980), worth mentioning are the works by Bell and Hillmer (1987) and Binder and Dick (1989). The proposed methods are based on time series analysis. Milestones of this approach are: (i) considering the observed data over time as a finite subset of a realisation

of a stochastic process; (ii) the definition of state space models (and the application of the Kalman filter to obtain parameter estimates and the correspondent standard errors);

- (m.3) methods using both spatial and temporal smoothing. Some methods are proposed by Pfeiffermann and Burck (1990) on the basis of state space modelling. Important results are presented in Singh, Mantel and Thomas (1994), where four generalisations of the Fay – Herriot predictor are proposed. Saei and Chambers (2003) describe models and algorithms to obtain SAE estimates when linear mixed models involving both area and time random effects are defined.

3. Design issues

4. Available software tools

In the last decade the availability of software tools for SAE is increased significantly. Several routines for multilevel model estimation released by developer teams of R, SAS, SPSS, STATA, MLwiN and WinBUGS or OpenBUGS can be used for small area estimation. Besides, ad hoc software for SAE has been developed by some international projects on the small area estimation topic. It is worth mentioning the SAS macros produced by the EURAREA project, the R functions or libraries released by the projects BIAS, SAMPLE, AMELI and ESSnet-SAE.

Furthermore an extensive review of the SAE software tools is provided in the WP4 of the ESSnet-SAE project.

5. Decision tree of methods

In this section we report the step by step process of the activities related to the production of small area estimates as defined in the WP6 of the ESSnet-SAE project. This process is displayed in figure 3, where three separate stages are defined:

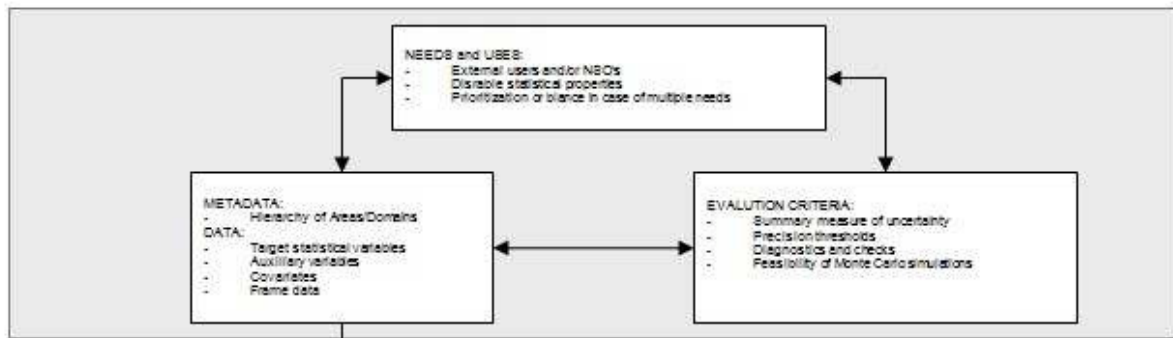
(I) clarification: identification of needs and purposes of small area estimation (e.g., estimation of key parameters or ranks for funding allocation);

(II) basic smoothing: direct, and design-based synthetic and composite estimates (triplet) are computed. No change of the inferential framework is needed compared to the direct estimates produced for the survey;

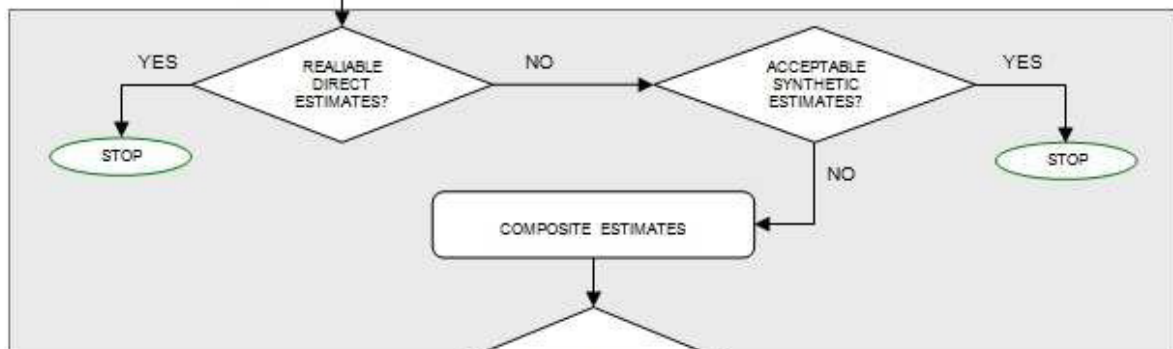
(III) enhancement: it is needed if the basic design-based smoothing is not effective. Quality assessment of the triplet of design-based small area estimates should identify weaknesses in order to work properly for improvements.

Therefore, according to point (III) when design-based methods cannot guarantee the requested precision of small area estimates, enhanced methods based on explicit modelling should be used. After computing model-based estimates, users must verify the validity of the hypotheses underlying the models. Furthermore should also check for the possible bias introduced for misspecification of the model by means a set of bias diagnostics. These diagnostics are based on the comparison between the whole set of direct estimates and model-based estimates. (see Brown et al., 2001). For an overview of model diagnostics for SAE see also the report on WP6 of the ESSnet SAE.

(I) Clarification



(II) Basic smoothing



(III) Enhancement

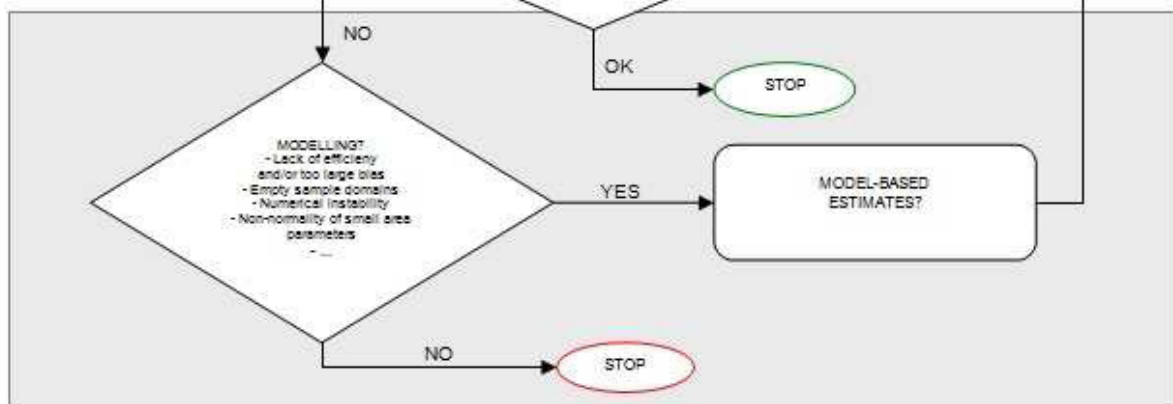


Figure 3. Flow chart of a SAE process.

Figure 4 shows the options available when dealing with model-based SAE. For each target variables the model selection phase should include issues as the choice of the more proper set of auxiliary variables and the definition of the small areas to be included in the broad domain. The following step concerns the choice between fixed and mixed effects models. As stated in the previous section the relationship between fixed effects (regression models) and random effects (mixed models) is analogous to that between synthetic and composite estimators. Users are expected to answer a couple of questions: (i) is the regression model good enough? (ii) does the extra computational effort of the mixed model pay off?

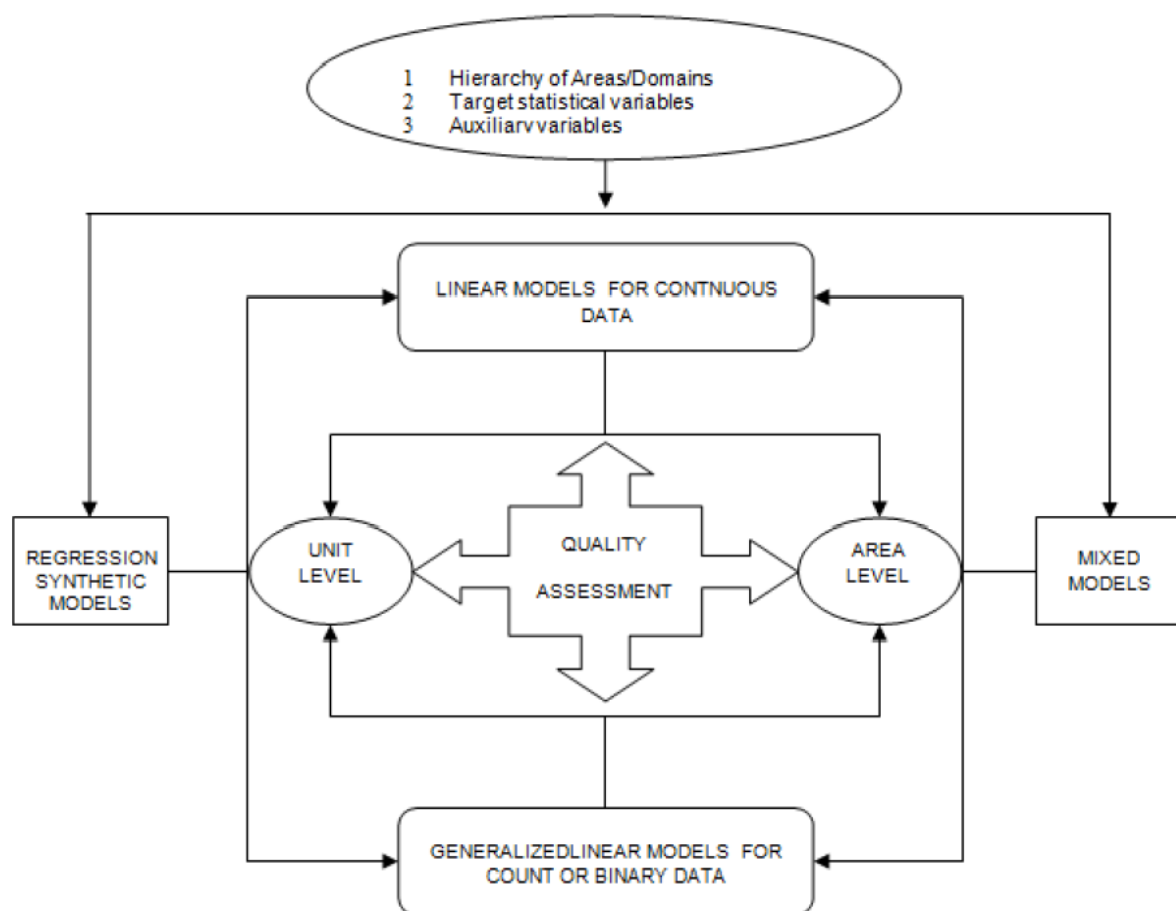


Figure 4. Flow chart of model-based SAE.

The following choice depends on the nature and the availability of the data at the two levels. The sampling design also plays a role in the choice. There may design features having a strong impact on the final estimates, e.g., stratification, multistage sample selection and/or clustering. Area level models take into account straightforwardly sampling weights since direct estimates are involved. If unit level models are used, design effects need to be considered as non-informative, given the auxiliary information.

Next step concerns the choice between linear and generalised linear (or nonlinear) models. From theoretical point of view, generalised linear models should be preferred for categorical data. In practice, however, linear models are computationally much easier, and often yield similar results.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28–36.

- Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001), Evaluation of small area estimation methods: an application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium Achieving Data Quality in a Statistical Agency: A Methodological perspective*, Statistics Canada.
- Cochran, W. G. (1977), *Sampling Techniques*. John Wiley & Sons, Hoboken, New Jersey.
- Cressie, N. A. (1991), *Statistics for spatial data*. John Wiley & Sons, Hoboken, New Jersey.
- Datta, G. S. and Ghosh, M. (1991), Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics* **19**, 1748–1770.
- Deville, J. C. and Särndal, C.-E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.
- Drew, J. D., Singh, M. P., and Choudhry, G. H. (1982), Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology* **8**, 17–47.
- ESSnet SAE (2012), Report on Workpackage 6 – Guidelines (contributors Istat (Italy), CBS (Netherlands), SSB (Norway), GUS (Poland), INE (Spain), ONS (United Kingdom). <http://www.essnet-portal.eu/sae-2>
- Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small place: an application of James-Stein procedures to Census Data. *Journal of the American Statistical Association* **74**, 398–409.
- Ghosh, M. (1992), Constrained Bayes estimation with applications. *Journal of the American Statistical Association* **87**, 533–540.
- Gonzalez, M. E. (1973), Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section*, American Statistical Association, 33–36.
- Gurney, M. and Daly, J. F. (1965), A multivariate approach to estimation in periodic sample surveys. *Proceedings of The Social Statistics Section*, American Statistical Association, 242–257.
- Jessen, R. J. (1942), Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agriculture Station Results Bulletin* **304**.
- Horvitz, D. G. and Thompson, D. J. (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.
- Morris, C. N. (1983), Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**, 47–59.
- Patterson, H. D. (1950), Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B* **12**, 241–255.
- Pfeffermann, D. and Burck, L. (1990), Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16**, 217–237.
- Pratesi, M. and Salvati, N. (2008), Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications* **17**, 113–141.
- Purcell, N. J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review* **48**, 3–18.

- Rao, J. N. K. (2003), *Small Area Estimation*. John Wiley & Sons, Hoboken, New Jersey.
- Saei, A. and Chambers, R. (2003), Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects. *S3RI Methodology Working Papers*, Southampton Statistical Sciences Research Institute, M03/15.
- Schaible, W. L. (1979), A composite estimator for small areas. *National Institute on Drug Abuse, Research monograph*, U.S. Government Printing Office, 24.
- Singh, A. C., Mantel, H. J., and Thomas, B. W. (1994), Time series EBLUPs for small areas using survey data. *Survey Methodology* **20**, 33–43.

Interconnections with other modules

8. Related themes described in other modules

1. Sample Selection – Main Module

9. Methods explicitly referred to in this module

1. Weighting and Estimation – Synthetic Estimators for Small Area Estimation
2. Weighting and Estimation – Composite Estimators for Small Area Estimation
3. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
4. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation
5. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

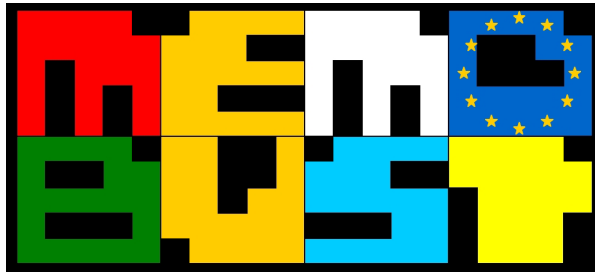
Weighting and Estimation-T-Small Area Estimation

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-04-2012	first version	Stefano Falorsi, Fabrizio Solari	ISTAT
0.2	14-08-2012	changes after review	Stefano Falorsi, Fabrizio Solari	ISTAT
0.2.1	25-02-2013	changes after review	Stefano Falorsi, Fabrizio Solari	ISTAT
0.2.2	10-09-2013	preliminary release		
0.3	24-10-2013	changes after EB review	Stefano Falorsi, Fabrizio Solari	ISTAT
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:34



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Synthetic Estimators for Small Area Estimation

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	7
4. Examples – not tool specific.....	7
5. Examples – tool specific.....	7
6. Glossary.....	7
7. References	7
Specific section.....	9
Interconnections with other modules.....	13
Administrative section.....	14

General section

1. Summary

In surveys conducted by statistical offices one of the main problem is to have reliable estimates for domains for which the sample size is too small or even equal to zero. It is the consequence of the fact that many institutions need more detailed information not only for the whole country but also for some specific subdomains such as geographic areas or other cross-sections. It also concerns business statistics where increasing demand exists for information for different classification of activities (e.g., trade, manufacturing, transport, construction, etc.) including small, medium and large enterprises and many variables (e.g., revenue, operating costs, taxes, etc.). In such situations direct estimates based only on specific domain sample data were insufficient because of high variability and small precision. The remedy could be the methodology of small area estimation (SAE) which plays an important role in the field of modern information provision, which aims to cut survey costs while lowering the respondent burden. Thanks to their properties, SAE methods enable reliable estimation at lower levels of spatial aggregation and with more specific domains, where direct estimation techniques display too much estimator variance. Another advantage over direct estimators is that small area estimation can be used to handle cases with few or no observations for a given domain in the sample. Therefore it is necessary in many situations to use indirect estimates that borrow strength by taking into account values of the variables of interest from related areas and from that point of view increasing the “effective” sample size. Generally speaking there are basically two types of indirect estimators: the synthetic and the composite estimators which can be derived under a design-based approach or taking into account the fact that an explicit area level or unit level model exists. The main aim of module is to provide a set of principles for synthetic estimators. Information about the first group of estimators can be found in the module “Weighting and Estimation – Composite Estimators for Small Area Estimation”.

2. General description of the method

One of indirect estimators is the synthetic estimator, which relies on a properly chosen model. Such a model takes into account auxiliary information from different sources, such as sample survey data, census data or administrative records, in other words it “borrows strength” to improve the process of estimation. Modeling in this area involves making use of implicit or explicit statistical models to indirectly estimate small area parameters of interest. The traditional synthetic estimators rely on an implicit linking model. In this case synthetic estimators for small areas are derived from direct estimators for a large area that covers some small areas under the assumption that the small areas have the same characteristics as the large area. In other words, an estimator is called a synthetic estimator if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area, see Rao (2003).

Recently explicit linking models have come to play a more important role in the literature on small area estimation and have brought significant improvements in techniques of indirect estimation. Drawing on mixed model methodology, these techniques incorporate random effects into the model. Random effects account for the between-area variation that cannot be explained by including auxiliary variables, see Mukhopadhyay and McDowell (2011). A broader discussion of synthetic estimators

derived only from linear mixed models can be found later in this module. It is worth noting that model-based estimators can also be derived from linear models without taking into account specific area effects. For more details, see Rao (2003).

It should be mentioned that there is a compromise between direct estimators and synthetic estimators. It relies on the fact that when the sample size for a specific domain is small, direct estimators have large variance and small precision but low or no bias. On the other hand, for the same specific domain, a synthetic estimator is often biased, especially if the above assumption is not fulfilled, but is better than a direct estimator from the point of view of precision, i.e., the variance of this estimator is smaller. This compromise is used by the so-called composite estimators which will be discussed in detail in the module “Weighting and Estimation – Composite Estimators for Small Area Estimation”.

As it was stated above synthetic estimators can be considered from design and model-based perspectives. Synthetic design-based estimators make use of survey weights $d_i = 1/\pi_i$, which are based on the probability distribution and depend on the specific sample design, i.e., first order inclusion probabilities are used $\pi_i = P(i \in s)$ but they also take into account information about the domain under study (for instance information about the population area size for domain d or a known total value for the variable X for the d -th small area/domain can be used). Synthetic model-based estimators make use of a properly chosen statistical model that “borrows strength” in making an estimate for one small area from sample survey data collected in other small areas.

Later in this document the most common synthetic estimators (both design and model-based) for the total value of study variable Y in the d -th domain will be introduced. Various formulas for different synthetic estimators will be shown. All these formulas are based on taking into account the more reliable Horvitz-Thompson direct estimator for the broad area and domain and use it to construct an estimator for the small area/domain. The review of the estimators is based on ESSnet Project on Small Area Estimation (2012b) and Rao (2003).

One of the simplest synthetic estimators is the so-called BARE (Broad Area Ratio Estimator) which takes into account only additional information about the population area size N_d . The formula for the BARE estimator is as follows:

$$\hat{Y}_{d,BARE} = N_d \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i} = N_d \frac{\hat{Y}_{BA}}{\hat{N}_{BA}}, \quad (1)$$

where N_d is the population area size for domain d , s denotes the sample, d_i is the initial weight associated with the i -th unit in the sample and y_i is the value of the target variable for this unit, $d = 1, \dots, D$. This formula states that the total value for the variable under study y for the large area is proportionally allocated in all small areas according to the population area sizes N_d .

If domain-specific auxiliary information is available in the form of k -vector of known totals \mathbf{X}_d^T , then as estimator for the domain total Y_d the so-called the regression-synthetic estimator can be taken into account. The formula below is appropriate for any model that has been used to derive the parameter $\hat{\boldsymbol{\beta}}$ which is the regression coefficient based on data from a broad area (the whole country or the entire region):

$$\hat{Y}_{d,GRS} = \mathbf{X}_d^T \hat{\boldsymbol{\beta}}. \quad (2)$$

The $\hat{\beta}$ regression coefficient is the solution of the sample weighted least squares equations and its formula can be found in Rao (2003).

The next simple estimator in the class of synthetic estimators is the so-called ratio-synthetic estimator. It takes into account a broad area survey estimate $\hat{Y}_{BA} = \sum_{i \in S} d_i y_i$ and can be used when the value of a single auxiliary variable X is available in the form of a total value for each small area from another source. The formula for this estimator is given by:

$$\hat{Y}_{d,RSE} = X_d \frac{\sum_{i \in S} d_i y_i}{\sum_{i \in S} d_i x_i} = X_d \frac{\hat{Y}_{BA}}{\hat{X}_{BA}}, \quad (3)$$

where X_d is the known total value for the variable X for the d -th small area/domain and $\hat{X}_{BA} = \sum_{i \in S} d_i x_i$ is the direct survey estimate of the total of the only one auxiliary variable at the broad area.

In some situations “good” direct estimates (acceptable precision) for the broad area are also known for cross-classification of respondents, e.g., sex, age groups, place of residence for social surveys or ownership form in business statistics. In such cases, if population sizes or auxiliary variable totals are known for all cross-classifications for specific small areas, it is possible to construct an appropriate synthetic estimator which is called a post-stratified estimator. In this approach classification counts play the role of poststrata.

When population sizes N_{dg} for cross-classification g in small area d are known it is possible to construct a so-called count-synthetic estimator $\hat{Y}_{d,CSE}$ given by the formula:

$$\hat{Y}_{d,CSE} = \sum_g N_{dg} \frac{\hat{Y}_g}{\hat{N}_g}, \quad (4)$$

where \hat{Y}_g is the direct survey national estimate of the variable under study for cross-classification cell g , \hat{N}_g is the direct survey national estimate of the national population size for cross-classification cell g , N_{dg} is the known population size for cross-classification g in small area d and g denotes the cross-classifications of poststrata, e.g., $g=1$ to 16 can represent firms according to size of the firm (four variants: micro, small, medium and large size) and section (also four variants: trade, manufacturing, transport, construction).

In the case when the total value of a single auxiliary variable X is known at cross-classifications for each small area and is measured in the survey, a so-called combined ratio-synthetic estimator $\hat{Y}_{d,CRSE}$ can be constructed. Its formula is given by:

$$\hat{Y}_{d,CRSE} = \sum_g X_{dg} \frac{\hat{Y}_g}{\hat{X}_g}, \quad (5)$$

where \hat{X}_g is the direct survey national estimate of the auxiliary variable for cross-classification cell g and X_{dg} is the known value of the auxiliary variable for cross classification cell g of the small area d .

Estimators discussed above were derived under a design-based approach and make use of design weights during the process of estimation. What follows below is a discussion of synthetic estimators obtained assuming that an explicit area level or unit level model exists.

One very important class of synthetic estimators are those which are based on linear mixed models, see Inglese, Russo and Russo (2008), Eurarea (2004). The first synthetic estimator, called Synth A in the EURAREA project, is given by Eurarea (2004):

$$\hat{Y}_d^{Synth A} = \mathbf{X}_d^T \hat{\boldsymbol{\beta}}^{unit}, \quad (6)$$

where \mathbf{X}_d^T is a vector of area level k covariates of known population totals and which is based on a unit level mixed model:

$$y_{di} = \mathbf{x}_{di}^T \boldsymbol{\beta} + u_d + e_{di}, \quad (7)$$

where $u_d \sim iid N(0, \sigma_u^2)$, $e_{di} \sim iid N(0, \sigma_e^2)$, \mathbf{x}_{di} is the vector of k covariates relates to the i -th unit within area d , $\boldsymbol{\beta}$ is the $(k \times 1)$ vector of the model coefficients, u_d is the random area effect associated with small area d , and e_{di} is the unit level random error. In most cases, the variances σ_u^2 and σ_e^2 are unknown so they have to be estimated using for example the restricted maximum likelihood method (REML). Other possibilities of estimation also exist. For example, Proc Mixed in SAS can be used to calculate ML or REML estimates of σ_u^2 and σ_e^2 . For details, see Rao (2003). The weighted least squares estimator for the model coefficients $\boldsymbol{\beta}$ of size $(k \times 1)$ is given by:

$$\hat{\boldsymbol{\beta}}^{unit} = (\sum_{d=1}^D \mathbf{x}_d^T \hat{\mathbf{V}}_d^{-1} \mathbf{x}_d)^{-1} (\sum_{d=1}^D \mathbf{x}_d^T \hat{\mathbf{V}}_d^{-1} \mathbf{y}_d) \quad (8)$$

where \mathbf{x}_d is an $(n_d \times k)$ matrix of values of the k covariates related to area d , \mathbf{y}_d is the $(n_d \times 1)$ vector of the target variable and $\hat{\mathbf{V}}_d$ is the estimated variance-covariance matrix of the vector \mathbf{y}_d given by the formula:

$$\hat{\mathbf{V}}_d = \hat{\sigma}_e^2 \mathbf{I}_{n_d} + \hat{\sigma}_u^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T, \quad (9)$$

where \mathbf{I}_{n_d} and $\mathbf{1}_{n_d}$ denote an identity matrix of dimension n_d and an n_d -dimensional vector of 1s respectively and $\hat{\sigma}_e^2$, $\hat{\sigma}_u^2$ are estimates of the variance components σ_e^2 and σ_u^2 respectively.

The second synthetic estimator, called Synth B in EURAREA project, is given by Eurarea (2004):

$$\hat{Y}_d^{Synth B} = \mathbf{X}_d^T \hat{\boldsymbol{\beta}}^{area} \quad (10)$$

and it is based on an area level mixed model with auxiliary variables available at area level and random area-specific effects and errors independently normally distributed:

$$y_d = \mathbf{X}_d^T \boldsymbol{\beta} + u_d + e_d, \quad (11)$$

where $u_d \sim iid N(0, \sigma_u^2)$, $e_d \sim iid N(0, \sigma_e^2)$, y_d is the value for the target variable in d -th area and e_d is the area level random error. The weighted least squares estimator for the model coefficients $\boldsymbol{\beta}$ of size $(k \times 1)$ is given by:

$$\hat{\boldsymbol{\beta}}^{area} = (\sum_{d=1}^D \mathbf{X}_d \mathbf{X}_d^T / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2))^{-1} (\sum_{d=1}^D \mathbf{X}_d \hat{Y}_d / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2)). \quad (12)$$

Estimation of variance and MSE of synthetic estimators is described in details in Rao (2003). Generally speaking, the estimation process is different for variance and MSE in the design-based and model-based situation. The variance of the design-based synthetic estimators will be small compared with the variance of a direct estimator because of the fact that it depends only on the precision of direct estimators at a larger area level. From this point of view the variance of design-based synthetic estimators is estimated using standard design-based methods. For example the variance of the ratio-synthetic estimator or of the count-synthetic estimator can be estimated using the Taylor linearisation method. Similarly the variance of the regression-synthetic estimator can be estimated using resampling methods such as the jackknife. A broad discussion of the variance and the MSE of estimators under the design-based approach can be found in Särndal et al. (1992) and Rao (2003).

For synthetic model-based estimators the problem of constructing an MSE is more complicated than in the case of design-based estimators because it depends on the underlying model which can be very complex. A discussion devoted to MSE estimation in the model-based approach can be found in a report prepared during Essnet Project on Small Area Estimation (2012c). More details about MSE estimation in the model-based approach can also be found in the monograph by Rao (2003) and in Jiang and Lahiri (2006). Formulas for the MSE for Synth A and Synth B estimators can also be found in Inglese, Russo and Russo (2008) and Eurarea (2004).

3. Preparatory phase

4. Examples – not tool specific

We refer to Rao (2003) for many examples of applying synthetic estimators in real surveys including health variables, county estimates of wheat production in the state of Kansas with evaluation, etc. The article written by Marker (1999) contains also a broad discussion of synthetic estimators with many examples presented in detail.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Boonstra, H. J. and Buelens, B. (2011), *Model-based estimation*. Contribution to the Methods Series, Statistics Netherlands, The Hague.
- Essnet Project on Small Area Estimation (2012a), *Report on Workpackage 3 – Quality Assessment*, Final Version, March 2012.
- Essnet Project on Small Area Estimation (2012b), *Report on Workpackage 4 – Software Tools*, Final Version.
- Essnet Project on Small Area Estimation (2012c), *Report on Workpackage 6 – Guidelines*, Final Version, March 2012.
- EURAREA (2004), *Project Reference Volume*. <http://www.statistics.gov.uk/eurarea>
- Fabrizi, E., Rosaria, M., and Pacei, S. (2007), Small Area Estimation of Average Household Income Based on Unit Level Models for Panel Data. *Survey Methodology* **33**, 187–198.
- Ghosh, M. and Rao, J.N.K. (1994), Small Area Estimation: An Appraisal. *Statistical Science* **9**, 55–76.
- Gonzalez, J.F., Placek, P.J., and Scott, C. (1996), Synthetic Estimation in Followback Surveys at the National Center for Health Statistics. In: W.L. Schaible (ed.), *Indirect Estimators in U.S. Federal Programs*.

- Inglese, F., Russo, M., and Russo, A. (2008), Different approaches for evaluation precision Small Area Model-Based Estimators. *Proceedings of Q2008, European Conference on Quality in Official Statistics*.
- Jiang, J. and Lahiri, P. (2006), Mixed model prediction and small area estimation. *TEST* **15**, 1–96.
- Marker, D.A. (1999), Organization of Small Area Estimators Using a Generalized Linear Regression Framework. *Journal of Official Statistics* **15**, 1–24.
- MEETS (2011), *Use of Administrative Data for Business Statistics*, Final Report, Poznań (Poland) 2011.
- Mukhopadhyay, P. K. and McDowell, A. (2011), *Small Area Estimation for Survey Data Analysis Using SAS Software*. SAS Institute Inc., Cary, NC.
- Rao, J.N. K. (2003), *Small area estimation*. John Wiley & Sons, Hoboken, New Jersey.
- SAE package developers (2007), *Introduction to Small Area Estimation*.
www.bias-project.org.uk/software/SAE.pdf
- Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992), *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Stasny, E., Goel, P.K., and Rumsey, D.J. (1991), County Estimates of Wheat Production. *Survey Methodology* **17**, 211–225.

Specific section

8. Purpose of the method

The method is used for small area estimation and comprises some techniques (including the case with no and auxiliary information) used for estimation when the sampling size in the domain of interest is too small to obtain reliable estimates using a direct estimator. The purpose of this method is to provide acceptable estimates for small areas when using direct estimators is impossible (no units in the sample for specific domain) or there are only some units in particular small areas. Synthetic estimation is based on the concept of “borrowing strength” and uses both survey and auxiliary data from outside as well as within the domain/small area of interest. As a consequence using additional sources of information in synthetic estimators generally leads to higher precision and, if the key assumption of homogeneity within the larger domain is fulfilled, to reduction of bias.

9. Recommended use of the method

1. These estimators can be applied for estimation when sample data are not available for the domain of interest, since the only required information is local covariate means or totals and the value of $\hat{\beta}$, which is based on data from the entire region, or country, covered by the survey.
2. Synthetic estimators can be used even when sampling was not involved. It is especially important in business statistics where units are taken into a sample not always according to an appropriate sampling scheme. For example, we can use the synthetic approach when information about the mean or total value of y is known from some administrative source and means or totals of covariates are also known for the domain of interest and at the level of the population.
3. This class of indirect estimators should be recommended in all surveys when information for small areas/domains is needed mainly because of its simplicity, applicability to general sampling designs or surveys where a sample design is not present, and potential of increased accuracy in estimation by borrowing “strength” from similar small areas.

10. Possible disadvantages of the method

1. If the assumption that small areas have the same characteristics as the large area is not fulfilled, then estimates may not be appropriate. Such an assumption is quite strong, and in fact for some areas or domains, synthetic estimators can be heavily biased in the design-based framework, see Ghosh and Rao (1994).
2. When one wants to use synthetic estimators for small areas, it is very important that good auxiliary information is available.
3. When one wants to use synthetic estimators for population totals or means in small areas, it is very important to take possible selection effects into consideration as far as possible. Selection effects may cause systematic differences in the target variable between sample and population. In synthetic estimators based on a model and used for the entire population it may be less

useful to predict the non-observed part of the population and this may lead to huge bias, see Boonstra and Buelens (2011).

4. For some synthetic estimators, the estimates \hat{Y}_d for small areas do not add up to the direct large area estimate \hat{Y} . In such cases adjustment is needed in order to ensure coherence of estimates at different levels. A potential solution is to use the following formula:

$$\hat{Y}_{d,adj} = \frac{\hat{Y}_d}{\sum_d \hat{Y}_d} \hat{Y}. \quad (13)$$

A detailed discussion of several adjusting methods can also be found in the topic “Macro-Integration”.

11. Variants of the method

1. Variants of the method depend on the availability (or not) of auxiliary information. For example in the situation where the only available additional information is the population area sizes, the broad area ratio estimator can be used. If domain-specific auxiliary information is available in the form of known totals then the regression-synthetic estimator is a good solution.
2. Variants of the method depend also on how synthetic estimators were derived: under a design-based approach or taking into account the fact that an explicit area or a unit level model exists.

12. Input data

1. Input data set can be classified according to the type of synthetic estimator. For example, for the BARE, ratio-synthetic estimator, count-synthetic estimator and combined ratio-synthetic estimator, information about the design weights d_i for all units in the sample s is required as well as about the values of the target variable y_i and the auxiliary variable x_i (e.g., for the ratio-synthetic estimator). Depending on the synthetic estimator, as mentioned above, information is required about known population sizes N_d in the domain d , population sizes N_{dg} for cross-classification g in small area d , known total value X_d for the variable X for d -th small area or known value X_{dg} of the auxiliary variable for cross classification cell g of the small area d . This information can come from a census or administrative registers. In the case of the model-based synthetic estimators, information about known totals \mathbf{X}_d^T of auxiliary variables for all small areas is needed.

13. Logical preconditions

1. Missing values
 1. When an area contains no data in the sample, synthetic estimators may be used. This is one very important advantage of synthetic estimators compared especially to direct estimators.
2. Erroneous values
 1. Standard small area methods do not take into consideration errors in auxiliary variables. A possible misspecification of the area level variables or correction in the variables is not taken into account.

3. Other quality related preconditions

1.

4. Other types of preconditions

1.

14. Tuning parameters

1. The tuning parameters of synthetic estimators should be specified only for synthetic model-based estimators. Parameters for the convergence of the iterative method for such estimators may be: the maximum number of iterations, convergence criterion. Details of macros in SAS for unit and area level synthetic estimators are described in the Eurarea documentation (2004). Some functions are also available in R and put on the SAE page at the Cross portal, see Essnet Project on Small Area Estimation (2012b).

15. Recommended use of the individual variants of the method

1. When a domain in the sample is not represented at all or there are only a few sampled units in specific domains (using direct estimators is doubtful due to the large variance), synthetic design-based estimators should be taken into consideration at first because they are easy to implement and easy to understand by the recipient of statistical information.
2. In some cases basic synthetic estimators based on the design-based approach may give unacceptable results (e.g., when there are too many empty domains). In such situations, the model-based approach may be taken into account especially when additional covariates and/or correlations can be included in a model. This approach can be used both when the target variable y is quantitative (linear regression can be used) or categorical (logistic regression can be used).
3. If the auxiliary information used for synthetic estimators is not very predictive for the target variable, then predicted area means are pulled too much towards the general sample average. In this situation small area methods based on models with random area effects are more suitable, see Boonstra and Buelens (2011).

16. Output data

1. In many examples devoted to synthetic estimators, which can be found in the literature when the true value is known (simulation studies), an output dataset usually contains a table with the following information: estimates for a small area, variance of the estimator, MSE, confidence intervals and bias. In real applications, when the true value is not known, the output data set usually is poorer and consists of estimates for all small areas, the variance of the estimator or model-based MSE.

17. Properties of the output data

1. The user should check the quality of estimates based on their knowledge of the investigated phenomenon and the variance of the estimators. In simulation studies also MSE, bias of estimates and confidence intervals may be checked, see ESSnet Project on Small Area Estimation (2012a).

18. Unit of input data suitable for the method

Processing unit level data and domain level variables for computations of the synthetic estimator (area level Synth B) and its variance.

19. User interaction - not tool specific

1. Select the model (no models, unit-level model, area-level model), choose auxiliary variables to be included into the model.
2. Establish the level of aggregation.
3. In the case of synth A and synth B establish tuning parameters (convergence criteria, starting point, stopping point).
4. After the use of synthetic estimators quality indicators should be checked and verified in order to evaluate the final results (variance, MSE, interval confidence).

20. Logging indicators

1. The specific logging indicators depend on the type of synthetic estimator. It can include run time of the application and /or number of iterations to reach convergence in the estimation process and characteristics of the input data.

21. Quality indicators of the output data

1. Variance of the estimator (both in real and simulation studies).
2. MSE, bias and confidence intervals – usually in simulation studies when the real value of the parameter is known, see ESSnet Project on Small Area Estimation (2012a).

22. Actual use of the method

1. The method is applied in a wide range in the U.S. Federal Statistical System. It should be mentioned that in fact some of these applications had an experimental nature and were not disseminated in official statistics. For example synthetic estimators are used by The National Center for Health Statistics in United States and in agricultural surveys, see Gonzalez, Placek and Scott (1996), Stasny, Goel and Rumsey (1991). Synthetic estimators are also used by Statistics Canada to estimate some characteristics of the labour market. This technique is also used to estimate average household income, see Fabrizi, Rosaria and Pacei (2007). In business statistics synthetic estimators are rather used in a very limited range. Some applications can be found in the MEETS project, of which the main goal was to highlight possibilities of using administrative data resources for purposes of estimating enterprise indicators and the resulting benefits. In this project, small area estimators, including synthetic, were implemented to estimate some characteristics (revenue, number of employees, wages) according to short-term and annual statistics of medium and large sized enterprises. For details, see MEETS (2011). In the literature, however, it was pointed out that synthetic estimators may be applied to replace design-based methods to estimate population totals when a known random sample design is not present. It may, for instance, concern estimation based on incomplete registers, of which the VAT turnover register is an example, see Boonstra and Buelens (2011).

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Small Area Estimation
2. Macro-Integration – Main Module

24. Related methods described in other modules

1. Weighting and Estimation – Composite Estimators for Small Area Estimation
2. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
3. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation
4. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

25. Mathematical techniques used by the method described in this module

1. For design-based synthetic estimators basic knowledge of linear algebra is needed. For model-based synthetic estimators the knowledge of iterative methods is required.

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate weights
2. 5.7 Calculate aggregates

27. Tools that implement the method described in this module

1. A review of the available small area estimators routines and software can be found in Essnet Project on Small Area Estimation (2012b, c). It covers such statistical programs as R (packages sae2, arm, mass lme4, MCMCglmm, INLA and many others are mentioned), SAS (including proc MIXED and proc IML), STATA, SPSS, MLwiN and WinBUGS. Some specific information devoted to synthetic estimators, their application and implementation can also be found in some other documents.
 - Eurarea project <http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html>
 - R package sae2 which is not available on CRAN and can be downloaded from the website of the BIAS project: <http://www.bias-project.org.uk/>, see SAE package developers (2007). Information about R packages can also be found in Essnet Project on Small Area Estimation (2012b).

28. Process step performed by the method

Estimation of parameters in disaggregated domains

Administrative section

29. Module code

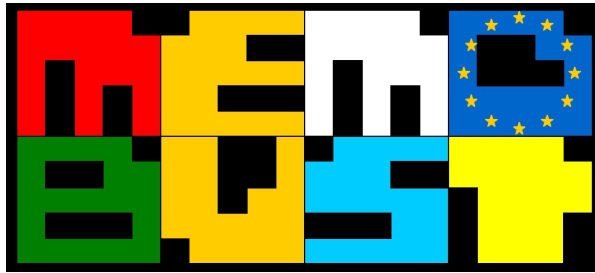
Weighting and Estimation-M-Synthetic Estimators for SAE

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	10-02-2012	first version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.2	14-01-2013	second version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.3	31-01-2014	third version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.4	14-03-2014	fourth version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.4.1	19-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:34



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Composite Estimators for Small Area Estimation

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	5
4. Examples – not tool specific.....	5
5. Examples – tool specific.....	5
6. Glossary.....	5
7. References	5
Specific section.....	7
Interconnections with other modules.....	10
Administrative section.....	12

General section

1. Summary

In surveys conducted by statistical offices one of the main problems is to have reliable estimates for domains for which the sample size is too small or even equal to zero. It is the consequence of the fact that many institutions need more detailed information not only for the whole country but also for some specific subdomains such as geographic areas or other cross-sections. It also concerns business statistics where increasing demand exists for information for different classification of activities (e.g., trade, manufacturing, transport, construction, etc.) including small, medium and large enterprises and many variables (e.g., revenue, operating costs, taxes, etc.). In such situations direct estimates based only on specific domain sample data are insufficient because of high variability and small precision. The remedy could be the methodology of small area estimation (SAE) which plays an important role in the field of modern information provision, which aims to cut survey costs while lowering the respondent burden.

Thanks to their properties, SAE methods enable reliable estimation at lower level of spatial aggregation and with more specific domains, where direct estimation techniques display too much variance. Another advantage over direct estimators is that small area estimation can be used to handle cases with few or no observations for a given domain in the sample. Therefore it is necessary in many situations to use indirect estimates that borrow strength by taking into account values of the variables of interest from related areas and from that point of view increasing the “effective” sample size.

Generally speaking there are basically two types of indirect estimators: the synthetic and the composite estimators which can be derived under a design-based approach or taking into account the fact that an explicit area level or unit level model exists. In this part of the handbook only design-based composite estimators are described. For details on model-based composite estimators see Rao (2003) or the modules mentioned in section 24 below. The main aim of this module is to provide a set of principles for composite estimators. Information about the first group of estimators can be found in the module “Weighting and Estimation – Synthetic Estimators for Small Area Estimation”.

2. General description of the method

Composite estimators provide a broad class of indirect estimators and are used in situations when the direct estimator is not taken into account because of its large variance and the synthetic estimators give unacceptable results because of bias. Composite estimators can be seen as estimators which give a compromise between the large variance of direct estimators and the bias of synthetic estimators and from that point of view they are built for balancing the properties of the direct and the synthetic estimator. When the sample size is quite large the direct estimator is valuable. On the other hand when the sample size is small or even equal to zero synthetic estimators are more valuable. From that point of view a composite estimator can be considered as an estimator that usually takes into account a direct and an indirect estimate and is better in the sense of having smaller bias and variance.

One common type of the composite estimator is a weighted average of two estimators – direct ($\hat{Y}_{dir,d}$) and synthetic ($\hat{Y}_{synth,d}$). Generally speaking, this class of estimators is a very easy solution to the problem of large bias of synthetic estimators and large variance of direct estimators. Composite estimators can be defined as follows:

$$\hat{Y}_{com,d} = \gamma_d \hat{Y}_{dir,d} + (1 - \gamma_d) \hat{Y}_{synth,d} \quad (1)$$

where γ_d is a weight from the interval $[0,1]$ in the small area d . The above expression is a convex combination of the direct and synthetic estimators and, in general, the choice of a proper weight γ_d depends on the size of the sample in the small area d . If the sample size in the small area is large enough, then the direct estimator should receive a bigger weight. Otherwise if the sample size gets smaller than the synthetic part receives a bigger weight.

Finding the right value of the weight γ_d constitutes the main problem in the use of composite estimators. This is very important from the point of view of balancing the potential bias of the synthetic estimator against the instability of the direct estimator. The way of selecting this weight is very controversial. One of the most common solution is to take $\gamma_d = n_d/N_d$, where n_d is the sample area size for domain d and N_d is the population area size for domain d . Alternatively γ_d can be obtained by minimising the mean square error (MSE) of the composite estimator, see Rao (2003). In this second approach the weights can be obtained by minimising the MSE of the composite estimator $\hat{Y}_{com,d}$, with respect to γ_d , under the assumption that the covariance between direct and synthetic estimator is small compared to the MSE of $\hat{Y}_{synth,d}$. In this approach it can be shown that the optimal weight is given by the formula:

$$\gamma_d = \frac{MSE(\hat{Y}_{synth,d})}{MSE(\hat{Y}_{dir,d}) + MSE(\hat{Y}_{synth,d})}. \quad (2)$$

Some other ways of finding γ_d are discussed in Ghosh and Rao (1994), Holmoy and Thomsen (1998) and Singh, Gambino and Mantel (1993). Here, our attention will be focused only on the so-called sample size dependent estimator(SSD) which is a special case of the composite estimator with weights γ_d which depend on the domain counts \hat{N}_d and N_d where \hat{N}_d is the sum of all design weights in domain d , i.e., $\hat{N}_d = \sum_{i=1}^{n_d} d_i$, and N_d is the population size in domain d . In Drew, Singh and Choudhry (1982) the proposition for γ_d is as follows:

$$\gamma_d = \begin{cases} 1 & \text{if } \hat{N}_d \geq \alpha N_d, \\ \frac{\hat{N}_d}{\alpha N_d} & \text{otherwise,} \end{cases} \quad (3)$$

where α is subjectively chosen parameter. Generally speaking when the sample size in domain d increases, γ_d is close to 1 and the composite estimator $\hat{Y}_{com,d}$ is very similar to direct estimator. Otherwise the synthetic estimator has a bigger contribution.

Another proposition can be found in Särndal and Hidiroglou (1989):

$$\gamma_d = \begin{cases} 1 & \text{if } \hat{N}_d \geq N_d, \\ \left(\frac{\hat{N}_d}{N_d}\right)^{h-1} & \text{otherwise,} \end{cases} \quad (4)$$

where h is subjectively chosen. When $\alpha = 1$ and $h = 2$, the weight γ_d is the same in the first and the second approach.

A discussion devoted to different types of composite estimators derived under design-based approach can also be found in Rao (2003).

Estimation of the MSE of the composite estimators, even when a weight γ_d is fixed, runs into difficulties similar to those for synthetic estimators. For details, see the module on synthetic estimators

and Rao (2003) where a broad discussion devoted to the problem of MSE estimation of composite estimators can be found.

3. Preparatory phase

4. Examples – not tool specific

In the literature one can find many examples of composite estimators both in real surveys and simulation studies. Eklund (1998) used composite estimators to estimate the net coverage error for the 1997 U.S. Census of Agriculture at the state level. Falorsi, Falorsi and Russo (1994) used the composite estimator of the number of unemployed in Health Service Areas of the Friuli region in Italy. The method was also applied in the Labour Force Survey by Griffiths (1996). An example of the use of the sample size dependent estimator can be found in Farver (2002) where this estimator was used in the estimation of food-animal productivity parameters. A broad discussion devoted to examples of applications of composite estimators can also be found in Rao (2003).

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Costa, A., Sattora, A., and Ventura, E. (2009), Using composite estimators to improve both domain and total area estimation. <http://www.econ.upf.edu/docs/papers/downloads/731.pdf>
- Drew, D., Singh, M. P., and Choudhry, G. H. (1982), Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology***8**, 17–47.
- Essnet Project on Small Area Estimation (2012a), *Report on Workpackage 3 – Quality Assessment*. Final Version, March 2012.
- Essnet Project on Small Area Estimation (2012b), *Report on Workpackage 6 – Guidelines*. Final Version, March 2012.
- Eklund, B. (1998), Small area estimation of coverage error for the 1997 Census of Agriculture. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 335–338.
- Falorsi, P. D., Falorsi, S., and Russo, A. (1994), Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey. *Survey Methodology***20**, 171–176.
- Farver, T. B. (2002), Comparison of ratio-synthetic, sample-size dependent and EBLUP estimators as estimators of food-animal productivity parameters. *Preventive Veterinary Medicine***52**, 313–332.
- Ghosh, M. and Rao, J.N.K. (1994), Small Area Estimation: An Appraisal. *Statistical Science***9**, 55–76.

- Griffiths, R. (1996), Current Population Survey Small Area Estimations for Congressional Districts. *Proceedings of the Section on Survey Research Method*, American Statistical Association, 314–319.
- Holmoy, A.M. K. and Thomsen, I. (1998), Combining Data from Surveys and Administrative Record Systems. The Norwegian Experience. *International Statistical Review* **66**, 201–221.
- Longford, N.T. (2005), *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer.
- MEETS (2011), *Use of Administrative Data for Business Statistics*. Final Report, Poznań (Poland).
- Molina, I. and Marhuenda, Y. (2013), *Package SAE*.
<http://cran.r-project.org/web/packages/sae/sae.pdf>
- Opsomer, J.D., Botts, C. and Kim, J.Y. (2003), Small area estimation in a watershed erosion assessment survey. *Journal of Agricultural, Biological, and Environmental Statistics* **8**, 139–152.
- Rao, J.N.K. (2003), *Small area estimation*. John Wiley & Sons, Hoboken, New Jersey.
- Särndal, C.-E. and Hidiroglou, M.A. (1989), Small domain estimation: a conditional analysis. *Journal of the American Statistical Association* **84**, 266–275.
- Singh, M.P., Gambino, J.G., and Mantel, H. (1993), Issues and options in the provision of small area statistics. In: G. Kalton, J. Kordos, and R. Platek (eds.), *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Designs*, vol. 1, Central Statistical Office, Warsaw, 37–75.
- Ugarte, M.D., Goicoa, T., Militino, A.F., and Sagaseta-Lopez, M. (2009), Estimating unemployment in very small areas. *SORT* **33**, 49–70.

Specific section

8. Purpose of the method

The method is used for small area estimation and involves some variants of combining two estimators into one by taking a weighted average of these estimators. Even though many small area estimators, both design- and model-based, have the basic form of a linear weighted combination of two estimators, the most common approach is to take the direct and synthetic estimator in the formula for the composite estimator. The aim of this intervention is to balance the potential bias of a synthetic estimator and the high variance of a direct one.

9. Recommended use of the method

1. This estimator can be useful in domains in which a direct estimator has a large variance.
2. This estimator can be useful in surveys when analysed domains vary very much in terms of sample size. To avoid the inconvenience related to switching from a direct estimator to a synthetic one, the composite approach can be used, balancing the influence of the used estimators.
3. Because of the simplicity of composite estimators they should be recommended in all surveys when methods of small area estimation are used. They are easy to implement and not difficult to understand by the users. With direct and synthetic estimators they form the so-called triplet of small area estimates and can always be produced using existing data, see Essnet Project on Small Area Estimation (2012b).

10. Possible disadvantages of the method

1. How to establish the value of the weight γ_d is a matter of discussion.
2. Another problem is how to provide measures of error for a given small area – for example, for bias. It should be mentioned that the bias, even if smaller than for synthetic estimators, is also present for composite estimators.
3. Composite estimators are sometimes called shrinkage estimators because of the fact that all the direct estimates are pulled towards the corresponding synthetic estimate of a broader area. As a consequence composite estimators generally display less between-area variation than they should. In the literature this inconvenience is known as the over-shrinkage problem. For details, see Essnet Project on Small Area Estimation (2012b).
4. For some composite estimators, the estimates \hat{Y}_d for small areas do not add up to the direct large area estimate \hat{Y} . In such cases adjustment is needed in order to ensure coherence of estimates at different levels. Potential solution is to use following formula:

$$\hat{Y}_{d,adj} = \frac{\hat{Y}_d}{\sum_d \hat{Y}_d} \hat{Y}. \quad (5)$$

11. Variants of the method

1. Variants of the method depend on which estimators are taken into account in the formula of the composite estimator. In the basic approach, the composite estimator is a weighted average

of a direct and a synthetic estimator. However the expression of composite estimators can be considered as a convex combination of two different estimators than a direct and a synthetic estimator. In the literature devoted to small area estimation many estimators, both design and model-based, have the composite form. Rao (2003) provides many composite estimators including the sample size dependent estimator and the James-Stein method and many examples of their applications.

2. Variants of the method depend also on the way how the weight γ_d is established.

12. Input data

1. The input data set depends on which estimators are taken into account in the formula for composite estimators and the source of information. The input data set can contain individual information for all units in the sample. In this situation the direct and synthetic estimator can be calculated and, as a consequence, the composite estimator is directly established as a weighted sum of these two estimators. The input data set can also contain information coming from auxiliary sources. Specific software may be based on different structures of the input data set in the procedure of estimation using the composite approach.

13. Logical preconditions

1. Missing values
 1. When an area contains no data in the sample, synthetic estimators may be used. In this situation the composite estimator reduces to the synthetic one, i.e., $\gamma_d = 0$.
2. Erroneous values
 1. Standard small area methods do not take into consideration errors in auxiliary variables. A possible misspecification of the area level variables or correction in the variables is not taken into account.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. Because of the fact that a composite estimator consists of a direct and a synthetic estimator, parameters for the convergence of the iterative method may be the same as for the model-based synthetic estimator: the maximum number of iterations, and the convergence criterion. One of the tuning parameter could also be the weight γ_d .

15. Recommended use of the individual variants of the method

1. In some situations where small areas vary strongly in terms of sample size a direct estimator can be good for areas with the largest sample sizes. On the other hand, a direct estimator is very poor when the representation in the sample is very small or equal to zero. In this case a

synthetic estimator may be more effective. Switching from one estimator to the other is inconvenient. The problem can be solved by using composite estimation, which balances inconveniences of these two estimators, see Longford (2005).

2. Because of the fact that composite estimators are easy to implement compared to explicit model-based estimators, they are recommended to use as basic smoothing approach in all surveys when small area estimation methods are taken into account.
3. When the composite weights depend only on the sub-sample sizes, it is possible to derive composite estimates for a large number of target variables at the same time. For comparison at the same time a model applies only to one or very few variables so it is impractical to build models for all variables in the sample. It is usually impractical to build models for all the statistical variables that are collected in the sample, neither at the national level nor at the small-area level, see Essnet Project on Small Area Estimation (2012b). Summing up, composite estimators (especially SSD) are useful when dealing with many variables comparison with fitting appropriate models for different variables.
4. Some recommendations devoted to how establish some parameters in composite estimators can be found in the literature. For example, it is recommended, with regard to sample size dependent estimators, that in formula (4) h should be equal to 2, see Särndal and Hidiroglou (1989). For the weight γ_d in formula (3) it is recommended that $\alpha = 1$. However in the Canadian Labour Force Survey α is equal to $2/3$.

16. Output data

1. An output dataset usually contains a table with estimates for all small areas. The following measures may also be included in an output data set: MSE, variance, confidence intervals or bias especially in simulation studies when the true value of parameters are known and it is very easy to calculate them.

17. Properties of the output data

1. The user should check the quality of estimates based on their knowledge of the investigated phenomenon and MSE, variance, bias of estimates or confidence intervals if possible, see Essnet Project on Small Area Estimation (2012a).

18. Unit of input data suitable for the method

For the purpose of computations using composite estimators both unit level data and domain level variables can be used.

19. User interaction - not tool specific

1. Select estimators as components of the composite estimator.
2. Establish the weight γ_d as a weighting factor in the formula for the composite estimator.
3. Choose auxiliary variables to be included into the synthetic part of the composite estimator.
4. Establish the level of aggregation.
5. Establish tuning parameters (convergence criteria, starting point, stopping point) if necessary.

6. After the use of the composite estimator quality indicators, if possible, should be checked and verified in order to evaluate the final results (MSE, confidence interval).

20. Logging indicators

1. The logging indicators generally speaking depend on the two estimators taken into account in the formula for the composite estimators and may cover: run time of the application, number of iterations to reach convergence in the estimation process, characteristics of the input data, see also the item “logging indicators” in the module “Weighting and Estimation – Synthetic Estimators for Small Area Estimation”.

21. Quality indicators of the output data

1. Compare with quality indicators of the output data for synthetic estimators mentioned in the module “Weighting and Estimation – Synthetic Estimators for Small Area Estimation”.

22. Actual use of the method

1. Applications of composite estimators can be found in different areas of statistics. Composite estimators are in use in environmental statistics in a survey conducted in the Rathbun Lake Watershed in Iowa, see Opsomer, Botts and Kim (2003). Other examples of using composite estimators can be found in Costa, Sattora, Ventura(2009). In their article, which was based on a cooperation between The Institute of Statistics of Catalonia(IDESCAT) and the UniversitatPompeuFabra, composite estimators and their application to several areas of interest are described. Sample size dependent estimators are in use in surveys devoted to the labour market. For example, The Canadian Labour Force Survey, uses a sample size dependent estimator to produce Census Division level estimates. Another application of sample size dependent estimators in labour market statistics can be found in Ugarte et al.(2009). Some actual applications of composite estimators in business surveys can be found in documentation of the MEETS project, where composite estimators were implemented to estimate some characteristics (revenue, number of employees, wages) according to short-term and annual statistics of medium-sized and large enterprises. For details, see MEETS (2011). See and compare it with the information devoted to the actual use of the method in the module “Weighting and Estimation – Synthetic Estimators for Small Area Estimation”.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Small Area Estimation

24. Related methods described in other modules

1. Weighting and Estimation – Synthetic Estimators for Small Area Estimation
2. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
3. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation
4. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

25. Mathematical techniques used by the method described in this module

1. Basic knowledge of linear algebra is needed. When composite estimators are built using the model-based approach the knowledge of iterative methods is required.

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate weights
2. 5.7 Calculate aggregates

27. Tools that implement the method described in this module

1. In many cases own codes are required to implement the above mentioned composite estimators. However there are some functions in R which help to obtain composite estimates. For example, in the SAE package written by Isabel Molina and Yolanda Marhuenda one can find the `ssd` function which calculates sample size dependent estimators as a composition of direct and synthetic estimators. For details, see Molina and Marhuenda (2013).

28. Process step performed by the method

Estimation of parameters in disaggregated domains.

Administrative section

29. Module code

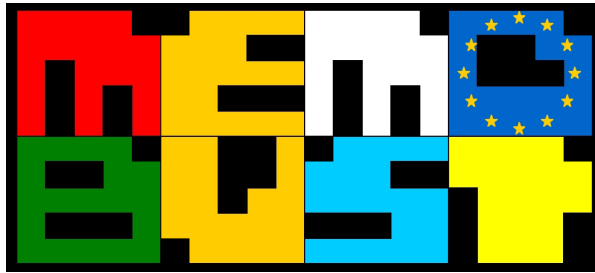
Weighting and Estimation-M-Composite Estimators for SAE

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	10-02-2012	first version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.2	14-01-2013	second version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.3	31-01-2014	third version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.4	14-03-2014	fourth version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.4.1	17-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:35



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	4
4. Examples – not tool specific.....	4
5. Examples – tool specific.....	5
6. Glossary.....	5
7. References	5
Specific section.....	7
Interconnections with other modules.....	10
Administrative section.....	12

General section

1. Summary

Small area (or small domain) estimation methods are a set of techniques allowing the estimation of parameters of interest for domains where the direct estimators (e.g., HT or GREG; see the theme module “Weighting and Estimation – Main Module” and the method module “Weighting and Estimation – Generalised Regression Estimator”, respectively) cannot be considered reliable enough, i.e., their variance is too high to be released. National Statistical Office surveys are usually planned at a higher level, hence, whenever more detailed information is required, the sample size may be not large enough to guarantee release of direct estimates and in some cases, smaller domains may happen to be without sample units. Small area methods increase the reliability of estimation by “borrowing strength” from a set of areas in a larger domain for which the direct estimator is reliable. This means that information from other areas is used and/or additional information from different sources is exploited (see the theme module “Weighting and Estimation – Small Area Estimation”).

The area level EBLUP, which is described in this module, is a linear combination of the area (domain) direct estimator and a predicted component based on a linear mixed model. The model relates the parameter of interest to known auxiliary variables for each of the domains that constitute the partition of the whole population. An effect to account for (within) domain homogeneity is included in the model.

2. General description of the method

The EBLUP area level is a small area estimation method (see the theme module “Weighting and Estimation – Small Area Estimation”). It is based on a linear mixed model which formulates the relationship between the parameter of interest and auxiliary area level information.

Let θ_d be the parameter to be estimated for each domain d. A linear relationship between θ_d and a set of covariates whose values are known for each domain of interest is assumed. In details

$$\theta_d = \mathbf{X}_d^T \boldsymbol{\beta} + u_d, \quad (1)$$

where \mathbf{X}_d is the vector of covariates for domain d and the u_d s ($d=1, \dots, D$) are domain effects assumed to be distributed with mean zero and variance σ_u^2 . The random effects account for the extra variability not explained by the auxiliary variables in the model.

Beside the model on the parameters, let us specify the sampling model. A design unbiased direct estimators $\hat{\theta}_d$ is supposed to be available (but not necessarily for all the domains), that is

$$\hat{\theta}_d = \theta_d + e_d, \quad (2)$$

where the e_d s are the sampling errors associated with the direct estimators, for which $E(e_d | \theta_d) = 0$, i.e., the direct estimator is assumed to be unbiased, and $V(e_d | \theta_d) = \varphi_d$, where the variances φ_d are supposed to be known.

Combining equations (1) and (2) a linear mixed model is obtained. The model is formulated as follows:

$$\hat{\theta}_d = \mathbf{X}_d^T \boldsymbol{\beta} + u_d + e_d . \quad (3)$$

Normality for e and u is commonly assumed for estimation of the Mean Square Error (MSE), but this assumption is not necessary for estimating the parameter. On the basis of model (3) the empirical best linear unbiased estimator (EBLUP) is

$$\hat{\theta}_d^{\text{EBLUP_AREA}} = \gamma_d \hat{\theta}_d + (1 - \gamma_d) \mathbf{X}_d^T \hat{\boldsymbol{\beta}} , \quad (4)$$

where

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \varphi_d}$$

is the weight of the direct estimator and $\hat{\boldsymbol{\beta}}$ is the weighted least square (WLS) estimator of the regression coefficient vector $\boldsymbol{\beta}$, where the weights for estimating $\boldsymbol{\beta}$ are provided by a diagonal matrix whose generic element is given by $\hat{\sigma}_u^2 + \varphi_d$. The estimation for the parameters σ_u^2 and $\boldsymbol{\beta}$ has to be obtained recursively. Moreover, as already mentioned above, in order to avoid identifiability problems for the variance components, the sampling variances $V(e_d | \theta_d) = \varphi_d$ ($d=1, \dots, D$) are assumed to be known.

Nevertheless, if information at unit level is available, then under the hypothesis of homoscedasticity of the sampling errors, the variance φ_d can be estimated from a unit level model (see the method module “Weighting and Estimation – EBLUP Unit Level for Small Area Estimation”) or a generalised variance function (see Wolter, 2007, or Eurostat, 2013, p. 95). Anyway, this would affect the MSE of the predicted domain values (Bell, 2008).

For more details on model specification, methods for estimation of $\hat{\sigma}_u^2$ see Rao (2003, pp. 115-120). Details on the estimation of the MSE are given in Rao (2003, pp. 103 and 128-130).

For the application of the method the user can use several specific software in SAS or R. A review is available in ESSnet SAE Work Package 4 “Software Tools” downloadable from <http://www.cros-portal.eu/content/sae>.

3. Preparatory phase

4. Examples – not tool specific

We refer to the example reported in Fuller (2009, section 5.5, table 5.13) dealing with the prediction of wind erosion in Iowa for the year 2002. These data are taken from the U.S. National Resources Inventory. The same data have been used in Mukhopadhyay and McDowell (2011) and ESSnet SAE (2012) to display the use of SAS PROC MIXED and the R function mixed.area.sae respectively when area level model is applied for small area estimation. The data report for the 44 Iowa counties the direct estimates of each county of the cube root of the wind erosion measure, the total number of segments (population size), the sample number of segments (sample size). Auxiliary information is given by the erodibility index. There are 44 counties in Iowa, so all the counties are sampled, but for illustrative purposes 4 additional empty counties are created.

For the computation of area level EBLUP sampling errors of direct estimates are needed. Segments are supposed to be drawn by means of simple random sample. Preliminary analysis supported the hypothesis of a common within area variance. Hence sampling errors can be computed as σ_e^2/n_d , where σ_e^2 is obtained from the data and n_d is the sample size in county d .

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bell, W. R. (1999), Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute*.
<http://www.census.gov/did/www/saipe/publications/files/Bell99.pdf>
- Bell, W. R. (2008), Examining sensitivity of small area inferences to uncertainty about sampling error variances. U.S. Census Bureau, Small Area Income and Poverty Estimates.
<http://www.census.gov/did/www/saipe/publications/files/Bell2008asa.pdf>
- Bell, W. R. (2009), The U.S. Census Bureau’s small area income and poverty estimates program: a statistical review. <http://cio.umh.es/data2/T1A%20William.R.Bell@census.gov.pdf>
- Buelens, B., van den Brakel, J., Boonstra, H. J., Smeets, M., and Blaess, V. (2012), Case study, report Statistics Netherlands. *Essnet SAE WP5 report*, 62–81.
- Chandra, H. and Chambers, R. (2007), Small area estimation for skewed data. Small Area Estimation Conference, Pisa, Italy.
- Cressie, N. (1992), REML Estimation in empirical Bayes smoothing of census undercount. *Survey Methodology* **18**, 75–94.
- Datta, G. S., Ghosh, M., Nangia, N., and Natarajan, K. (1996), Estimation of median income of four person families: A Bayesian approach. In: D. A. Berry, K. M. Chaloner, and J. K. Geweke (eds.), *Bayesian Analysis in Statistics and Econometrics*, Wiley, New York, 129–140.
- Datta, G. S., Ghosh, M., Steorts, R., and Maples, J. (2009), Bayesian Benchmarking with Applications to Small Area Estimation property. Small Area Estimation Conference, Elche, Spain.
- Dick, J. P. (1995), Modelling net undercoverage in the 1991 Canadian Census. *Survey Methodology* **21**, 44–55.
- Ericksen, E. P. and Kadane, J. B. (1985), Estimating the Population in Census Year: 1980 and Beyond (with discussion). *Journal of the American Statistical Association* **80**, 927–943.
- ESSnet SAE (2012), *WP4 Final Report Deliverables of the project*.
http://www.cros-portal.eu/sites/default/files/WP4report_0.pdf
- EURAREA Consortium (2004), *PROJECT REFERENCE VOLUME*, Vol. 1.

<http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-andmodelling/eurarea/index.html>

- Eurostat (2013), *Handbook on precision requirements and variance estimation for ESS household surveys*. Methodologies and Working papers.
- Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.
- Fuller, W. A. (2009), *Sampling Statistics*. John Wiley & Sons, Hoboken, New Jersey.
- Montanari, G. E., Ranalli, M. G., and Vicarelli, C. (2009), Estimation of small area counts with the benchmarking property. Small Area Estimation Conference, Elche, Spain.
- Moura, E. A. S. and Holt, D. (1999), Small area estimation using multilevel models. *Survey Methodology* **25**, 73–80.
- Mukhopadhyay, P. K. and McDowell, A. (2011), *Small Area Estimation for Survey Data Analysis Using SAS Software*. <http://support.sas.com/resources/papers/proceedings11/336-2011.pdf>
- Pfeffermann, D. and Tiller, R. (2006), Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association* **101**, 1387–1397.
- Prasad, N. G. N. and Rao, J. N. K. (1990), The Estimation of the Mean Squared Error of Small Area Estimation. *Journal of the American Statistical Association* **85**, 163–171.
- Rao, J. N. K. (2003), *Small area estimation*. John Wiley & Sons, Hoboken, New Jersey.
- Torabi, M., Datta, G. S., and Rao, J. N. K. (2009), Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scandinavian Journal of Statistics* **38**, 598–608.
- Wang, Y., Fuller, W. A., and Qu, Y. (2008), Small area estimation under restriction. *Survey Methodology* **34**, 29–36.
- Wolter, K. M. (2007), *Introduction to Variance Estimation*. Springer, London.

Specific section

8. Purpose of the method

The method is used for small area estimation, which is a specific class of methods used for estimation when sampling size in the domain of interest is too small to attain efficient direct estimation. The method increases the reliability of the estimates by introducing a linear relationship between the direct estimates and known area level auxiliary variables.

9. Recommended use of the method

1. The method can be applied for estimation when few or even no sample data are available for one or more domains of interest.
2. The method can be applied on macrodata referred to domain level.
3. The method is useful to improve direct estimators if a set of covariates with a strong relationship with the variable of interest is available.
4. The variances of the small area direct estimates has to be known. Usually a smoothed model for variance estimation is applied and variances are assumed to be known. This affects the MSE (see Bell, 1999).
5. Covariates are needed only at domain level.

10. Possible disadvantages of the method

1. If the model is not correctly specified the estimator can be affected from bias.
2. When adding up small domains estimates to a larger domain, it is not ensured that direct estimates at larger level are obtained. A simple way to ensure consistency is to ratio adjust the EBLUP area level estimator. Benchmarking can be also set as a constraint to obtain small area estimates. This would produce different methods that will not be reported in the present handbook. (Wang et al., 2008; Pfeiffermann and Tiller, 2006; Montanari et al., 2009; Datta et al., 2009).
3. Symmetry of the distribution is required while in business survey skewness may be present. If transformation of variables does not suffice to reduce skewness advanced methods may be employed (Chandra and Chambers, 2007).
4. Assumptions of normality with known variance might be untenable at small sample sizes.
5. Model variance σ_u^2 can be estimated to be zero. This is an undesirable result. Hierarchical Bayesian methods are good alternatives and they always result in strictly positive variances, see, e.g., Bell (1999) and Buelens et al. (2012).

11. Variants of the method

1. Variants of the method are given by the different estimation methods for the variance component of model (3), e.g., Maximum Likelihood (ML) or Restricted (or Residual) Maximum Likelihood (REML) (Cressie, 1992), or the method of moments.

2. On the basis of model (3), an estimator making use of only the regression component is given by the area level synthetic estimator:

$$\hat{\theta}_d^{\text{Synth_arealevel}} = \mathbf{X}_d^T \hat{\boldsymbol{\beta}}.$$

This estimator uses only the relationship with the covariates and does not exploit the information on the variable of interest in the direct estimator. This estimator can be applied when a domain has no sample data.

12. Input data

Input data sets can be classified according to the source of information needed to apply the method. A first data set contains information calculated on sample data whereas a second one contains information provided from auxiliary sources. Specific software tools may need various structures for the input to produce estimation. We refer to the links in Section 27 below for software tools that make possible the application of the EBLUP area level.

1. Data set input 1 = a data set (macrodata) with direct estimates of the indicators for each domain and their variances.
2. Data set input2 = a data set (macrodata) containing population size and covariates for each domain.

13. Logical preconditions

1. Missing values
 1. Direct estimates in one or more domain can be missing. The EBLUP area level estimator does not account explicitly for missing values in the sample observations.
2. Erroneous values
 1. Standard small area methods do not take into consideration errors in the target variables. Possible misspecification of the area level auxiliary variables or correction in the variables are not taken into account by the EBLUP area level (but see Torabi et al., 2009).
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. Normality is often assumed for the estimation of the MSE.
 2. Sampling variance of the direct estimator has to be known or estimated aside from the area level model.

14. Tuning parameters

1. Parameters for the convergence of the iterative method: number of iterations and/or stopping rule, starting value for the variance of the random effects.

15. Recommended use of the individual variants of the method

1. Synthetic area level estimator is needed whenever no sample occurs in a specific domain.

2. For the estimation of the random component of the variance, software tools apply ML or REML. The method of moments is more robust with respect to non-normality.

16. Output data

1. Data set output1 = a dataset with predicted (macrodata) values for each domain and possibly MSE.

17. Properties of the output data

1. User should check MSE and bias diagnostic of the resulting estimates (see the ESSnet/sae site <http://www.cros-portal.eu/content/sae>).

18. Unit of input data suitable for the method

1. Processing domain level variables for the fitting of the model and the computations of the estimator.
2. Processing unit level data to compute variance estimation of the direct estimator (input for the method).

19. User interaction - not tool specific

1. Select the model, auxiliary variables to be included in the model, e.g., by means of AIC, BIC and cAIC.
2. Determine the aggregate level to which the model is defined, i.e., different models can be assumed for different large domains (aggregation of small domains).
3. Transformation of variable may be needed to satisfy model assumptions (symmetry and homogeneity).
4. Tuning parameters for convergence and specification of starting value for the variance of the random effects.
5. Choice of the method to be used for the estimation of the variance component.
6. After using the method, the quality indicators and logging should be inspected to assess possible presence of bias or inconsistency at different level of aggregation of estimates. Finally MSE for assessing reliability of estimates has to be monitored (see guidelines on the ESSnet/sae site <http://www.cros-portal.eu/content/sae>).

20. Logging indicators

1. Run time of the application.
2. Number of iterations needed to attain convergence in the estimation process.
3. When estimating the variance of the random effects zero or negative values can be attained. This may suggest problems in the variance estimation of the direct estimator. Otherwise hierarchical Bayes to fit model (3) may be applied (Datta et al., 1996).
4. Features of the input data set, e.g., size as it may affect computer time. Anyway problem size does not usually occur with EBLUP area method.

21. Quality indicators of the output data

1. MSE
2. Model Bias diagnostic
3. Benchmarking
4. Model selection diagnostic: AIC, BIC, cAIC

22. Actual use of the method

1. The method is applied by U.S. Census for poverty estimation since 1993, and by Statistics Canada for census undercount estimation.
2. Fay and Herriot (1979)
3. Bell (2009)
4. Dick (1995)

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Main Module
2. Weighting and Estimation – Small Area Estimation

24. Related methods described in other modules

1. Weighting and Estimation – Generalised Regression Estimator
2. Weighting and Estimation – Synthetic Estimators for Small Area Estimation
3. Weighting and Estimation – Composite Estimators for Small Area Estimation
4. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation
5. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

25. Mathematical techniques used by the method described in this module

1. ML or REML by means of Newton-Raphson algorithm

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate aggregates

27. Tools that implement the method described in this module

1. The collection of SAS macros included in the zip file [The EURAREA 'Standard' estimators and performance criteria](http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html) of the EURAREA project (<http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html>)

2. mixed.area.sae an R function produced by ESSnet SAE (ESSnet/sae site, <http://www.cros-portal.eu/content/sae>)
3. R package sae2 (BIAS project website: <http://www.bias-project.org.uk/>)
4. SAMPLE project codes in <http://www.sample-project.eu/it/the-project/deliverables-docs.html>

28. Process step performed by the method

Estimation of parameters in disaggregated domains

Administrative section

29. Module code

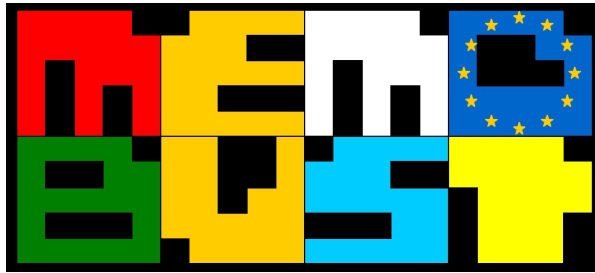
Weighting and Estimation-M-EBLUP Area Level for SAE

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	03-06-2011	first version	Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.2	25-11-2011	second version	Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.3	09-01-2012	third version	Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.3.1	17-10-2013		Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.3.2	21-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:35



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: EBLUP Unit Level for Small Area Estimation

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	5
4. Examples – not tool specific.....	5
5. Examples – tool specific.....	5
6. Glossary.....	5
7. References	6
Specific section.....	8
Interconnections with other modules.....	11
Administrative section.....	13

General section

1. Summary

The aim of Small Area Estimation (SAE) is to compute a set of reliable estimates for each small area for the target variable(s) of interest, whenever the direct estimates (see “Weighting and Estimation – Main Module” and “Weighting and Estimation – Generalised Regression Estimator”) cannot be considered enough reliable, i.e., the correspondent variances (see the module “Quality Aspects – Quality of Statistics”) are too high to make those estimates releasable.

Small area methods provide a set of techniques to obtain the estimates of interest in the National Statistical Institutes (NSIs) large scale survey, where more detailed information is required, and the sample size is not large enough to guarantee release of direct estimation. SAE methods which increase the reliability of estimates ‘borrowing strength’ from a larger area.

The unit level EBLUP estimator, which is described in this module, is a linear combination of the direct information and a regression synthetic prediction of non-sampled units. The fixed part of the model links the target values to some known auxiliary variables, for each units belonging to the larger area to which the small areas of interest belong to. The area specific random effects is instead introduced in order to take into account the correlation among the units with each small area (between area variation).

2. General description of the method

The unit level mixed model can be used when unit-specific auxiliary variables are available in each small area. The area-specific random effect terms are considered in order to take into account the between area variation through the correlation among units within a small area. The basic unit level linear mixed model is the nested error regression model formulated by Battese et al. (1988). It can be expressed as follows:

$$y_{di} = x_{di}^T \boldsymbol{\beta} + u_d + e_{di} \quad (1)$$

where

$$\begin{aligned} u_d &\sim iid N(0, \sigma_u^2) \\ e_{di} &\sim iid N(0, \sigma_e^2) \\ \forall i &= 1, \dots, N_d \\ d &= 1, \dots, D \end{aligned}$$

and y_{di} is the variable of interest for the i -th population unit in the d -th small area. Assuming non informative sampling designs, like simple random sampling, has been used at the sampling stage, the same model assumed for the population values can be applied for the sample units. Therefore, using a matrix notation, the following model can be formalised

$$\mathbf{y}_s = \mathbf{x}_s \boldsymbol{\beta} + \mathbf{z}_s \mathbf{u} + \mathbf{e}_s \quad (2)$$

where \mathbf{y}_s is n -dimensional vector of the observed values for the variable y , \mathbf{x}_s is the $(n \times p)$ -dimensional matrix of the covariate values observed in the sampling units, \mathbf{e}_s is the n -dimensional error vector, \mathbf{z}_s is the $(n \times D)$ -dimensional incidence matrix of the sampling units in the small areas, and \mathbf{u} is the D -dimensional vector of area random effects.

In order to obtain the small area estimates based on the above model, either a predictive or a Bayesian approach can be employed (see Rao, 2003, for more details). Following the predictive framework, the Best Linear Unbiased Predictor (BLUP) is obtained by minimising the quadratic loss in the linear unbiased estimator class. The BLUP depends on the variance components and that are usually unknown, so their estimates need to be computed. Both variance components and fixed effects parameters can be estimated in different ways, for example by means of Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) (Cressie, 1992) methods.

Once the parameters of the model have been estimated, the Empirical Best Linear Unbiased Predictor (EBLUP) based on unit level linear mixed model is a composite-type estimator. Letting aside the finite population correction factor, it is given by

$$\hat{\theta}_d^{\text{EBLUP_UNIT}} = \gamma_d \left[\bar{y}_d + \left(\bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}} \right) \right] + (1 - \gamma_d) \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}} \quad (3)$$

where

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d}$$

and $\bar{\mathbf{X}}_d$ is the vector of known population means of the auxiliary variables in the d -th area and $\bar{\mathbf{x}}_d$ is the correspondent vector of sample means. Given the model, the fixed effects parameter are estimates using the whole available larger area sample information and of course, when the between area variation is small, the EBLUP estimator tends towards the synthetic estimator (being the variance of random effects small). More weight is instead given to direct information when the variance of random effects is big respect the total variance.

There are several extensions of the above described basic unit level model. Since the basic model does not take into account for sample data collected with a complex sample design, some methodological development have been directed to specify more complex models that take into account the features of the sampling design. For instance, Stukel and Rao (1999) proposed a two-fold nested error regression model sample data for data collected from a stratified two-stage sampling.

The issue is that, when an informative design is used the inclusion probabilities of sampling units depend on the values of the target variable the model which holds for the sample data is different from the model assumed for the population data, so that it would be the cause of severe bias into the prediction. A possible approach with this regard is to explicitly include all the design variables used for the sample selection as covariates or the sampling weights in the specification of the model. These two options can be untenable when too many design variables are involved and when the sample weights are not available for non-sampled areas or non-sampled units. A Pseudo EBLUP estimator was proposed by Prasad and Rao (1999) starting from unit linear mixed model.

Moreover, multivariate nested error regression model has been proposed in order to estimate more than one small area parameters of interest simultaneously. This type of model, applied in Datta et al. (1999), allows to take into account the correlation among the characteristics under study observed in the sample units.

Finally, the linear unit level mixed models should be applicable only for continuous observations, then some enhancement models has been considered in order to deal with categorical dependent variables. In that case, Generalised Linear Mixed Models (GLMM) can be considered. Within this logistic regression models with mixed effects are commonly used for estimating small-area proportions (Malec et al., 1997).

3. Preparatory phase

Model selection is crucial preparatory phase since the objective is to lessen the chances of introducing design-bias into the small area estimates due to poor model specification. Model selection for each target variable was carried out considering diagnostic criteria such as maximum likelihood, AIC, BIC, Conditional AIC (cAIC) , and Cross Validation (CV) such as in Vaida and Blanchard (2005), Boonstra et al. (2008), and Boonstra, Buelens and Smeets (2009). Once one or several models has been selected, it is necessary to assess the fitting quality of the model(s). The study of model residuals by graphical representations, like Histograms, Q-Q plots, box-plots and mapping the residual, allows to check if the model assumptions are fulfilled.

4. Examples – not tool specific

We refer to Battese, Harter, and Fuller (1988) for an example of data for application of EBLUP Unit level model. These data are taken from a sample survey that have been designed to estimate crop areas for large regions. The predictions of the crop area for small areas such as counties has generally not been done for the lacking of available data directed collected from these areas. In order to apply the method, satellite data in association with farm-level survey observations has been used. They considered the estimation of mean hectares of corn and soybeans per segment and the auxiliary variables are the number of pixels classified as corn and soybeans in each county. In the example were considered data for 12 Iowa countries and data obtained from land observatory satellites.

Their example relates to application of SAS macros for computing the predictors under the model.

The same data is used as an example in <http://www.cros-portal.eu/content/sae> for explaining the use of R function **mixed.unit.sae.R**.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **80**, 28–36.
- BIAS project website: <http://www.bias-project.org.uk/>
- Boonstra, H. J., Buelens, B., and Smeets, M. (2009), Estimation of unemployment for Dutch municipalities. Small Area Estimation 2009 Conference, Elche, Spain, June 29-July 1.
- Boonstra, H. J., van den Brakel, J., Buelens, B., Krieg, S., and Smeets, M. (2008), Towards small area estimation at Statistics Netherlands. *Metron* **LIV**, 21–49.
- Brown, J., Chambers, R., Heady, P., and Heasman, D. (2003), Evaluation of small area estimation methods: an application to the unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001*.
- Chandra, H. and Chambers, R. (2007), Small area estimation for skewed data. Small Area Estimation Conference, Pisa, Italy.
- Cressie, N. (1992), REML Estimation in Empirical Bayes Smoothing of Census Undercount. *Survey Methodology* **18**, 75–94.
- D’Alò, M., Di Consiglio, L., Falorsi, S., and Solari, F. (2008), The Use of Sample Design Features in Small Area Estimation. ISI 2009 Conference, Durban (South Africa), 16-22 August.
- Datta, G. S., Day, B., and Basawa, I. (1999), Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference* **75**, 269–279.
- Datta, G. S., Ghosh, M., Steorts, R., and Maples, J. (2009), Bayesian Benchmarking with Applications to Small Area Estimation property. Small Area Estimation Conference, Elche, Spain.
- Dick, J. P. (1995), Modelling net undercoverage in the 1991 Canadian Census. *Survey Methodology* **21**, 44–55.
- EURAREA Consortium (2004), *PROJECT REFERENCE VOLUME*, Vol. 1.
<http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html>.
- Ghosh, M. and Rao, J. N. K. (1994), Small area estimation: an appraisal. *Statistical Science* **9**, 55–93.
- Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997), *Small area inference for binary variables in National Health Interview Survey*. *Journal of the American Statistical Association* **92**, 815–826.
- Molina, I., Saei, A., and Lombardia, M. J. (2007), Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society, Series A* **170**, 975–1000.
- Montanari, G. E., Ranalli, M. G., and Vicarelli, C. (2009), Estimation of small area counts with the benchmarking property. Small Area Estimation Conference, Elche, Spain.

- Pfeffermann, D. (2002), Small area estimation – New developments and directions. *International Statistical Review* **70**, 125–143.
- Pfeffermann, D. and Tiller, R. (2006), Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association* **101**, 1387–1397.
- Prasad, N. G. N. and Rao, J. N. K. (1999), On robust small area estimation using a simple random effects model. *Survey Methodology* **25**, 67–72.
- Pushpal, K, Mukhopadhyay, P. K., and McDowell, A. (2011), *Small Area Estimation for Survey Data Analysis Using SAS Software*. SAS Institute Inc., Cary.
<http://support.sas.com/resources/papers/proceedings11/336-2011.pdf>
- Rao, J. N. K. (2003), *Small area estimation*. John Wiley & Sons, Hoboken, New Jersey.
- SAE ESSnet (2012), Deliverables of the project. <http://www.cros-portal.eu/content/sae>
- Saei, A. and Chambers, R. (2003), Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. Methodology Working Paper- M03/15, University of Southampton, United Kingdom.
- Stukel, D. M. and Rao, J. N. K. (1999), Small area estimation under two-fold nested errors regression models. *Journal of Statistical Planning and Inference* **78**, 131–147.
- Torabi, M., Datta, G. S., and Rao, J. N. K. (2009), Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scandinavian Journal of Statistics* **38**, 598–608.
- Vaida, F. and Blanchard, S. (2005), Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–370.

Specific section

8. Purpose of the method

The method is used for small area estimation, when direct estimates usually applied for official statistics are too unreliable and unit level auxiliary information are available.

9. Recommended use of the method

1. The method can be applied for estimation when auxiliary information/covariates are available for each sample unit. The mean or total population values need to be known at area level.
2. A linear model can be used when the data are continuous and normally distributed. A transformation of the data may be required before modelling to make the data normally distributed.
3. The method is useful to improve direct estimator if a set of covariates with a strong relationship with the target variable is available.
4. If the target variable is not continuous or normally distributed a generalised linear model might be applied. For instance, the variable of interest at unit level is often binary, so that the logistic or probit model should be more appropriate.
5. Both unit and area level auxiliary information can be considered.

10. Possible disadvantages of the method

1. If the model is not correctly specified the estimator can be affected from severe bias.
2. The basic method do not consider the sampling strategy to select the units.
3. When adding up small domains estimates to a larger domain, it is not ensured that direct estimator at larger level is obtained. A simple way to guarantee this type of consistency is by means of ratio adjustment of the EBLUP unit level estimator. Benchmarking can be also set as a constraint to obtain small area estimates. This would produce different methods that will not be reported in the present handbook (Wang, Fuller, and Qu, 2008; Pfeiffermann and Tiller, 2006; Montanari, Ranalli, and Vicarelli, 2009; Datta et al., 2009).
4. The model assumes symmetry of the distribution, while in some cases, like in business survey, skewness may be present. If transformation of variables do not suffice to reduce skewness, advanced method may be considered. For instance by employing M-quantile models (Chandra and Chambers, 2007).
5. Standard small area models generally consider only i.i.d. area random effects, whereas more realistic and efficient models might include further structured random effects, such as time for repeated surveys and spatial autocorrelated random effects.

11. Variants of the method

1. Variants of the method are given by the different estimation methods for the variance component of model (3), e.g., Maximum Likelihood ML or Restricted Maximum Likelihood (REML) (see Cressie, 1992), or Method of moments.

2. On the basis of the assumed model, an estimator which uses only the regression component is given by the unit level synthetic estimator:

$$\hat{\theta}_d^{\text{Synth_unitlevel}} = \mathbf{X}_d^T \hat{\beta}$$

This estimator is always applied for no sampled domain.

3. For repeated sample surveys, extensions aimed to introduce time random effects can be also considered.
4. In order to consider the spatial autocorrelation among areas a unit level model with spatially correlated area effect can be considered. The spatial correlation can be introduced through the variance-covariance matrix of the random effects in function of the distance between areas or by modelling directly the random effects by means of SAR-type model.
5. Multinomial models are considered in Molina et al. (2007).

12. Input data

Input data set can be classified according to the source of information needed to apply the method. The first data set contains sample information whereas the second one contains information provided from auxiliary source at area level. Specific software tools may need various structure for the input to produce estimation. We refer to links in section 27 for software tools that make possible the application of EBLUP unit level.

1. Ds-input1 = a sample data set contains the target variable and auxiliary variables observed for each sampling unit.
2. Ds-input2 = a data set (macrodata) with mean or total values of covariates for each domain, and population size of the domain.

13. Logical preconditions

1. Missing values
 1. EBLUP unit level estimator does not account explicitly for missing values in the sample observations.
2. Erroneous values
 1. Standard small area methods do not take in consideration errors in the target variables and covariates. Possible misspecification of the auxiliary variables or correction in the variables are not taken into account by EBLUP unit level model (see Torabi et al., 2009).
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. Normality is often assumed for the estimation of the MSE.
 2. Sampling design is usually not considered in the inference.

14. Tuning parameters

1. Parameters for the convergence of the iterative method: number of iterations and/or stopping rule, starting value for the variance components of the models.

15. Recommended use of the individual variants of the method

1. For non-sampled area only synthetic type estimates can be computed.
2. For estimation of random component of the variance, software tools applies ML or REML.

16. Output data

1. Ds-output1 = the target values estimates for each domain and corresponding MSE.

17. Properties of the output data

1. User may check MSE bias diagnostic (see SAE ESSnet site <http://www.cros-portal.eu/content/sae>) of the resulting estimates.

18. Unit of input data suitable for the method

Sample unit level information for target variable and covariates to fit the model and to estimates the model parameters included the area random effects. Population area level means or totals for each domain to compute the estimator.

19. User interaction - not tool specific

1. Model selection, the choice of which auxiliary variables to include in the model, e.g., by means of AIC and BIC, cAIC
2. Satisfy the model hypotheses, like symmetry and homogeneity. A transformation of the variable may be needed.
3. Specification of starting value for the variance of the random effects and tuning parameters for convergence
4. Choice of method for variance component estimation
5. The use of the quality method should provide some evidence regarding spatial bias/autocorrelation at different level of aggregation of estimates. Finally MSE for assessing reliability of estimates has to monitored (see guidelines on <http://www.cros-portal.eu/content/sae>).

20. Logging indicators

1. Run time of the application
2. Number of iteration to attain convergence in the estimation process
3. Out of the range estimation of the target parameter can be attained when linear mixed model is assumed, in this case the normal assumption should be relaxed.

4. Underestimate of MSE can be possible under normality assumption and predictive approach to inference. Generalised linear mixed models and Hierarchical Bayes approach to inference may alternatively be applied.
5. Characteristics of the input data, for instance problems size.

21. Quality indicators of the output data

1. MSE
2. Model Bias diagnostic
3. Benchmark
4. Model selection diagnostic: AIC, BIC, cAIC
5. Analysis of the residual
6. Spatial distribution of area level residual (Maps)

22. Actual use of the method

The method is applied in:

1. Netherlands, for the production of the yearly estimates of unemployment fractions for all Dutch municipalities.
2. Spain, to produce reliable quarterly estimates of consumption expenditures of household and for the survey of the information Society-Families.
3. United Kingdom, to produce 2007/08 middle layer super output area MSOA-level estimates of the proportion of households in poverty for England and Wales, calculated based on equivalised household income after housing costs and produced using the same methodology that was used to produce mean income estimates.
4. Brazil, to generate estimates of poverty and inequality for 5500 Brazilian municipalities.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Main Module
2. Weighting and Estimation – Small Area Estimation
3. Quality Aspects – Quality of Statistics

24. Related methods described in other modules

1. Weighting and Estimation – Generalised Regression Estimator
2. Weighting and Estimation – Synthetic Estimators for Small Area Estimation
3. Weighting and Estimation – Composite Estimators for Small Area Estimation
4. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)

5. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

25. Mathematical techniques used by the method described in this module

1. ML or REML by means of Newton-Raphson algorithm

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate aggregates

27. Tools that implement the method described in this module

1. Eurarea SAS macro and function (<http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html>)
2. R functions produced by ESSnet SAE (<http://www.cros-portal.eu/content/sae>)
3. R package sae2 (BIAS project website: <http://www.bias-project.org.uk/>)
4. SAMPLE project codes in <http://www.sample-project.eu/it/the-project/deliverables-docs.html>

28. Process step performed by the method

Estimation of parameters in disaggregated domains

Administrative section

29. Module code

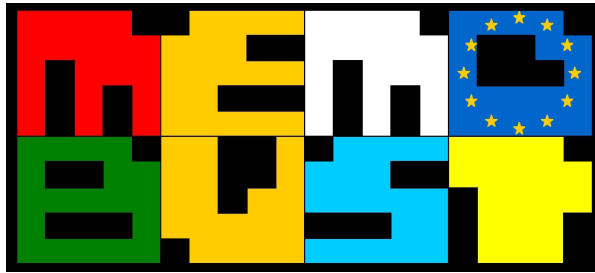
Weighting and Estimation-M-EBLUP Unit Level for SAE

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	30-12-2011	first version	Michele D'Alò, Andrea Fasulo	ISTAT
0.2	08-03-2012	second version	Michele D'Alò, Andrea Fasulo	ISTAT
0.2.1	26-03-2012	second version	Michele D'Alò, Andrea Fasulo	ISTAT
0.3	10-09-2013	third version	Michele D'Alò, Andrea Fasulo	ISTAT
0.3.1	12-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:35



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Small Area Estimation Methods for Time Series Data

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Time series area level models.....	3
2.2 Time series unit level models	4
3. Preparatory phase	5
4. Examples – not tool specific.....	5
5. Examples – tool specific.....	5
6. Glossary.....	5
7. References	5
Specific section.....	7
Interconnections with other modules.....	9
Administrative section.....	10

General section

1. Summary

The aim of small area estimation (SAE) is to produce reliable estimates for each small area for the target variables of interest, whenever the direct estimates cannot be considered enough reliable, i.e., the correspondent variances are too high.

SAE estimators borrow strength from neighbouring areas and auxiliary information deriving from administrative data. Another relevant source of information derives from data measured on previous occasions. In this case specific models can be defined in order to take into account the augmented amount of information with respect to cross-sectional data. Furthermore it is possible to exploit potential correlations between data from the same area on different times. In fact, most repeated survey samples usually include only partial replacement of sample units therefore gain in efficiency can be achieved by borrowing strength from other areas and other time occasions.

Two alternative model specifications are described in the literature. The former is based on linear mixed models in which an additional time depending random effect is added both in unit and area level framework, while the latter refers to state space models specifications.

2. General description of the method

This section describes small area estimation methods using time information. Section 2.1 describes methods based on area level models while section 2.2 illustrates techniques involving unit level model specifications. All the sections are devoted to the description of small area models involving time series data. For the sake of simplicity expressions of predictors are not given but users can easily find them in the references. Once the model parameters are estimated and population means or totals for the auxiliary variables are available, predicted values for each area and time can be straightforwardly computed computing standard expressions.

2.1 *Time series area level models*

Linear mixed models (LMMs) are one of the more common tools used for model-based small area estimation. LMMs are used for both area and unit models (see the modules “Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)” and “Weighting and Estimation – EBLUP Unit Level for Small Area Estimation”, respectively). In the first case records refer to units while in case of area level models each record is related to each small area. Basic LMM specifications formulate the relationship between the variable of interest and a set of auxiliary information. Furthermore in order to take into account extra-variability an additional random term is added in the model. In detail a random intercept term is added for each small area. When data for different times are available, this class of models can be straightforward improved introducing an additional random term related to time.

In case of area level models Rao and Yu (1992, 1994) proposed an extension of the basic Fay-Herriot model (Fay and Herriot, 1979) to handle time series and cross-sectional data. They define a time random component nested in the area random component. In details the following combination of sampling and linking models is proposed:

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta} + u_d + v_{dt}, \quad d = 1, \dots, D, \quad t = 1, \dots, T, \quad (1)$$

where \mathbf{X}_{dt} is the vector of covariates for the small area d at time t , the e_{dt} -s are the sampling errors related to the direct estimates $\hat{\theta}_{dt}$, they are uncorrelated over area and time and the variances ϕ_{dt} are supposed to be known, the u_d -s are domain effects assumed to be distributed with mean zero and common variance, and the v_{dt} -s are time random effects nested into the area effects u_d -s.

Rao and Yu (1992, 1994) suggest a first order autoregressive AR(1) specification to model the time random component v of the model. Hence, the model they propose depends on both area-specific effects and area-by-time specific effects which are correlated across time. More complex ARMA modelling for the time random effect is possible, although it is not clear if such complex modelling will result in efficiency gains (for more details see Rao, 2003). Datta et al. (2002) and You (1999) use the Rao-Yu model but replace the AR(1) model specification by a random walk model.

Alternative model specification to (1) is given in EURAREA Consortium (2004) and Saei and Chambers (2003). More specifically additional independent area and time random effects, and time depending regression coefficients are assumed, that is

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta}_t + u_d + v_t, \quad d = 1, \dots, D, \quad t = 1, \dots, T. \quad (2)$$

For the time random effects v_t both uncorrelation and first order autoregressive assumptions are made. Furthermore, similarly to model (1), a model with time varying area effects is also specified:

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta}_t + v_{dt}, \quad d = 1, \dots, D, \quad t = 1, \dots, T. \quad (3)$$

Here the v_{dt} -s are random effects following independent AR(1) processes for $d = 1, 2, \dots, D$.

The algorithms to obtain the Best Linear Unbiased Estimators (BLUE) of the regression coefficients $\boldsymbol{\beta}$, Best Linear Unbiased Predictor (BLUP) of the small area parameters, and the correspondent Empirical Best Linear Unbiased Predictor (EBLUP), when the variance components are unknown, are described in details in Saei and Chambers (2003). Two estimation methods for variance components are given: Maximum Likelihood (ML) and Residual Maximum Likelihood (REML).

Pfeffermann and Burck (1990) propose a general model involving area-by-time specific random effects. Their model can be written as

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta}_{dt}, \quad d = 1, \dots, D, \quad t = 1, \dots, T, \quad (4)$$

where the coefficients $\boldsymbol{\beta}_{dt}$ are allowed to vary cross-sectionally and over time. The variation of $\boldsymbol{\beta}_{dt}$ over time is modelled by a state-space model.

2.2 Time series unit level models

The unit level mixed model can be used when unit-specific auxiliary variables are available in each small area. Linear mixed model plays an important role in SAE context. Random effects are intended to reduce the extra-variability not explained by fixed effects. Standard small area models generally consider only i.i.d. area random effects. As reported in section 1 more realistic and efficient models

should take into account additional random effects related to meaningful components, such as time in case of repeated surveys.

Analogously to what described for area level specifications, LMMs are the basic tool to perform small area estimation using time series data. Saei and Chambers (2003) and EURAREA Consortium (2004) adapt the model specifications (2) and (3) to the unit level model framework, obtaining respectively:

$$y_{diti} = x_{diti}^T \beta + u_d + v_t + e_{diti}, \quad d = 1, \dots, D, \quad t = 1, \dots, T, \quad i = 1, \dots, N_{dt} \quad (5)$$

and

$$y_{diti} = x_{diti}^T \beta + v_{dt} + e_{diti}, \quad d = 1, \dots, D, \quad t = 1, \dots, T, \quad i = 1, \dots, N_{dt}. \quad (6)$$

As before several assumptions can be made for the random effects v_{dt} . For instance independent area and time random effects can be defined. Saei and Chambers (2003) and EURAREA Consortium (2004) specify models for which the time random effects are modelled according to a first order autoregressive AR(1) process.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Datta, G. S., Lahiri, P., and Maiti, T. (2002), Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference* **102**, 83–97.

EURAREA Consortium (2004), *PROJECT REFERENCE VOLUME*, Vol. 1.

<http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/downloads/index.html>.

Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Ghosh, M., Nangia, N., and Kim, D. (1996), Estimation of median income of four person families: a Bayesian time series approach. *Journal of the American Statistical Association* **91**, 1423–1431.

Pfefferman, D. and Burck, L. (1990), Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16**, 217–237.

- Pfeffermann, D. and Tiller, R. (2006), Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association* **101**, 1387–1397.
- Rao, J. N. K. (2003), *Small area estimation*. John Wiley & Sons, Hoboken, New Jersey.
- Rao, J. N. K. and Yu, M. (1992), Small area estimation by combining time series and cross-sectional data. *Proceedings of the Survey Research Section*, American Statistical Association, 1–9.
- Rao, J. N. K. and Yu, M. (1994), Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics* **22**, 511–528.
- Saei, A. and Chambers, R. (2003), Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. Methodology Working Paper- M03/15, University of Southampton, United Kingdom.
- You, Y. (1999), *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation*. Unpublished Ph.D. dissertation, School of Mathematics and Statistics, Carleton University, Canada.
- You, Y., Rao, J. N. K., and Dick, P. (2004), Benchmarking hierarchical Bayes small area estimators in the Canadian Census undercoverage estimation. *Statistics in Transition* **6**, 631–640.

Specific section

8. Purpose of the method

This collection of methods can be used for small area estimation, when time series data are available, i.e., when survey data are collected for several survey occasions. Quality of the estimates is improved by introducing linear relationship between target and auxiliary variables, and explicitly introducing in the models time dependent parameters.

9. Recommended use of the method

1. The methods can be applied when auxiliary information is available either for each sample unit (unit level modelling) or for each small area (area level modelling). Mean or total population values need to be known at area level.
2. Normality assumption is requested. A transformation of the data may be required before applying the methods.
3. The methods can be applied even if no sample data is available for one or more areas.

10. Possible disadvantages of the method

1. If the model is not correctly specified bias can seriously affect small area predicted values.
2. Sampling strategy is indirectly taken into account only when applying area level models.
3. When summing up small area estimates over a larger domain, benchmarking with direct estimator is not guaranteed. Benchmarking can be obtained with a posteriori adjustment of small area predicted values. Elsewhere benchmarking constraints can be included in the model specification (see Pfeiffermann and Tiller, 2006, for the frequentist approach, and You et al., 2002, for the Bayesian framework).

11. Variants of the method

1. Bayesian approach; see for instance Ghosh et al. (1996), You (1999), and Rao (2003, pp. 258-262).

12. Input data

1. Ds-input1 = data set containing sample information for each time. Information could refer to unit level data or area level data.
2. Ds-input2 = data set with population size and covariate mean values or totals for each time and for each area.

13. Logical preconditions

1. Missing values
 1. Sampling information in one or more small area can be missing. The methods previously described do not account for missing values in the sample observations.
2. Erroneous values

1. Errors in the target variables are not taken into account.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 1. When applying area level models, some model specifications require sampling variance of the direct estimator to be known (or estimated outside the area level model).

14. Tuning parameters

- 1.

15. Recommended use of the individual variants of the method

- 1.

16. Output data

1. Data set output1 = a dataset with predicted small area values for each domain and for each time, error evaluation.
2. Data set output2 = model parameter estimates.

17. Properties of the output data

1. Users should check MSE of the resulting estimates and model bias diagnostics.

18. Unit of input data suitable for the method

Processing domain level variables for the fitting of the model and the computations of the estimator.
Processing unit level data to compute variance estimation of the direct estimator (input for the method).

19. User interaction - not tool specific

1. Selection of the model, auxiliary variables to be included in the model, e.g., by means of AIC, BIC in the frequentist framework, or DIC in the Bayesian context.
2. Transformation of variable may be needed to satisfy model assumptions (symmetry and homogeneity).
3. Tuning parameters for convergence and specification of the starting values for the model parameters in the frequentist approach, choice of the starting values for the parameters in the model and the number of chains in case of Bayesian modelling.

20. Logging indicators

1. Number of iterations needed to attain convergence in the estimation process.
2. Diagnostics criteria to evaluate convergence of MCMC and evaluation of mixing in case of multiple chains.

21. Quality indicators of the output data

1. MSE or Posterior variance.
2. Model Bias diagnostics.
3. Benchmarking.
4. Model selection diagnostic: AIC, BIC, or DIC.

22. Actual use of the method

- 1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Small Area Estimation

24. Related methods described in other modules

1. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
2. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation

25. Mathematical techniques used by the method described in this module

1. ML or REML by means of Newton-Raphson or scoring algorithms.
2. MCMC algorithms.

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate aggregates

27. Tools that implement the method described in this module

1. The collection of SAS macros included in the zip file The EURAREA ‘Standard’ estimators and performance criteria of the EURAREA project (<http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html>).

28. Process step performed by the method

Prediction of totals or mean values for disaggregated domains.

Administrative section

29. Module code

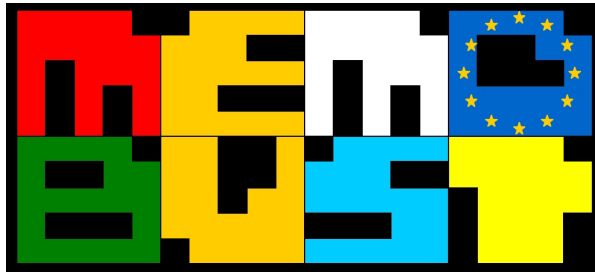
Weighting and Estimation-M-SAE Time Series Data

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	06-04-2012	first version	Michele D'Alò, Fabrizio Solari	ISTAT
0.2	10-05-2012	second version	Michele D'Alò, Fabrizio Solari	ISTAT
0.3	14-11-2013	third version	Michele D'Alò, Fabrizio Solari	ISTAT
0.3.1	18-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:36



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Estimation with Administrative Data

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction	3
2.2 Factors determining whether administrative data can be used to replace surveys	4
2.3 Using administrative data to replace surveys: general considerations	4
2.4 The statistical process.....	6
2.5 Findings per process step.....	7
2.6 Determining the active population	10
2.7 Estimation: available administrative data when the estimates have to be made	11
2.8 Estimation in case of almost complete coverage of administrative data	13
2.9 Estimation in case of few administrative data available.....	14
3. Design issues	18
4. Available software tools.....	18
5. Decision tree of methods	18
6. Glossary.....	19
7. References	19
Interconnections with other modules.....	21
Administrative section.....	22

General section

1. Summary

Official statistics produced by national statistical institutes (NSIs) can be based on primary or secondary data. Primary data are collected by the organisation also responsible for the statistical estimates, i.e., in this case the NSI. Secondary data are collected by another organisation or individual other than those responsible for the collection and aggregation of data from their initial source. An important secondary data source are administrative data. Administrative data are defined as data collected by another organisation for implementing an administrative regulation (or group of regulations). This module describes estimation techniques in case administrative data are used as replacement for a survey when estimating statistical variables. To keep the paper concise and illustrate the challenges with concrete examples, it is focussed on the use of Value Added Tax (VAT) data for estimating turnover.

2. General description

2.1 Introduction

Official statistics produced by national statistical institutes (NSIs) can be based on primary or secondary data. Primary data are collected by the organisation also responsible for the statistical estimates, i.e., in this case the NSI. Secondary data are collected by another organisation or individual other than those responsible for the collection and aggregation of data from their initial source. An important secondary data source are administrative data. Administrative data are defined as data collected by another organisation for implementing an administrative regulation (or group of regulations). Some of these administrative data can be used for statistical purposes. Tax data, i.e., data collected by tax authorities, can be considered as administrative data source. VATdata (Value Added Tax) of the tax office are the most widely used administrative data for enterprise statistics. For a complete overview of existing administrative data sources and their use for statistical purposes in Europe, we refer to Constanzo (2013) and the general results of the ESSnet project on the use of Administrative and Accounting Data (“ESSnet AdminData”).

The use of administrative data for statistical purposes has increased considerably during the last decade. Administrative data can be used in two ways:

- as auxiliary information in the statistical process.
- to replace survey data in the statistical process.

Examples of using administrative data as auxiliary information are:

- checking the validity of outlying survey values with administrative data,
- benchmarking the validity of survey estimates with administrative data,
- weighting survey results with GREGtype estimators (Kavaliauskiene et al., 2013) and administrative data as auxiliary information.

In this module, we focus on methodological issues arising when administrative data are used to replace survey data. Examples of using administrative data as replacement for survey data are:

- The use of the VATdata of the tax office for turnover estimates.
- The use of social security data from the tax office or social security agencies for employment estimates.
- The use of corporate tax data or building permits to estimate specific variables for annual statistics.
- The use of accountancy data to estimate specific variables for annual statistics.

2.2 *Factors determining whether administrative data can be used to replace surveys*

The first question which needs to be addressed is whether the used administrative data are suitable to replace variables from surveys. The answer on this question depends on several factors and affects the estimation technique. The most important factors to decide whether administrative data are suitable to replace survey variables are:

1. Does the NSI have legal access to these admin data?
2. Is the data transfer to the NSI guaranteed?
3. Is the NSI able to process large amounts of (administrative) data in a short time?
4. Can the administrative data be linked to the population frame derived from the statistical business register (SBR)?
5. Do the administrative data provide information about almost all enterprises within the target population when complete (i.e., completeness)?
6. Are the administrative data timely available (i.e., timeliness)?
7. Do differences in definition between administrative variables and statistical variables exist? Are these differences substantial and do they lead to biases in level and/or growth rate estimates. Are definition differences constant in time or not. Can the impact of differences in definition be monitored and corrected if required? (I.e., accuracy.)

If the answer on these seven questions do not reveal insuperable barriers – which is the case in several northern and north-western European countries – one may consider to use admin data for thereplacement of a survey. However, a methodology and a process needs to developed if the use of administrative data for replacing (variables in) surveys is considered. Guidelines for such a methodology, with is of course related to the statistical processes, are provided in the remaining part of this module.

VAT is the most commonly used administrative datasource in business statistics. Therefore, the term VAT instead of administrative data is used in concrete examples in the next modules of this theme. This choice has been made for sake of readability and concreteness. Methodologically the guidelines for VAT are also valid when using other admin data for estimating other statistical variables.

2.3 *Using administrative data to replace surveys: general considerations*

The general set-up when utilising VATdata for producing turnover estimates is that a combination of a survey and VATdata is used. In the survey, the large enterprises are generally completely enumerated. Since large enterprises often have a complex structure and their impact on the estimates is high,

correct surveyed observations from those large enterprises are considered crucial for producing reliable turnover estimates.

For the remaining small and medium enterprises VAT data are used instead of direct observations by the NSI. In other words, the general system of admin data based STS estimates consists of two parts:

1. use of a survey for the large enterprises (LEsurvey);
2. use of administrative data, for the remaining smaller enterprises.

The coverage of the large enterprise survey is a matter of debate. In order to define a more objective method to determine the coverage of the LEsurvey, Langford and Teneva (2012) have developed a method to calculate the impact of the ‘incompleteness’ factor on the boundary of the LEsurvey and the VAT part in an administrative data based STS system. This method is based on calculating revisions between the first estimates (with incomplete administrative data) and final estimates (with complete administrative data), by experimenting with different boundaries between the LEsurvey and the VAT parts. More information about revisions in official statistics can be found in the theme module “Quality Aspects – Revisions of Economic Official Statistics”.

Two fundamental issues arise for the population part which is estimated with administrative data:

1. Is the aim to produce estimates for population totals (and implicitly also for growth rates) or is the aim to produce growth rates only?
2. Are estimates produced at the microlevel, i.e., for individual enterprises, or at the macrolevel, i.e., using combinations of activities and size classes?

At first sight, there may not seem to be a big difference between estimates for population totals or for growth rates only. After all, by comparing the population total of the current period to that of a previous period, one can estimate the growth rate between these two periods. Conversely, given the growth rate between two periods and the population total in the first period, one can estimate the population total in the second period. However, there is a big difference between the two choices which affects the methodology. Estimates for population totals require better information about population characteristics and suspicious values than estimates of growth alone.

Concerning micro- versus macro-level estimates, both approaches can be used. The main advantage of micro-level estimates is that further processing is then easy: one simply has to aggregate over all enterprises in a publication cell to obtain an estimate for that publication cell. On the other hand, the choice of macro-level estimates can also be justified, because weighting with the Horvitz-Thompson estimator provides the same results as imputing missing values at microlevel using stratum averages.

The project ESSnet Admin Data has observed that most European NSIs produce:

- both levels and growth rates, and
- data at the microlevel.

The main reason for producing a) both levels and growth rates and b) micro-level estimates is that the great majority of the data are already available, allowing NSIs to construct an enriched dataset with (VAT-)turnover data of almost all enterprises, to which other survey data may be linked. Publishing growth rates only may provide slightly more stable results, as the link with the population frame is less critical. Moreover, outliers or non-predictable enterprises can be excluded more easily. However, level

estimates for year, quarter and (if possible) month can be used as auxiliary information and reference for (early) STS estimates with no or few admin data available (see section 2.9).

2.4 The statistical process

When administrative data are used for economic and STS statistics, a number of steps in the statistical process can be distinguished:

1. Data transfer from the tax office to the NSI;
2. Checking the data when entering the NSI (completeness, obvious errors, etc.);
3. Linking administrative units to statistical units;
4. Combining the result of the large-enterprise (LE) survey with the administrative data used for small and medium sized enterprises (SME);
5. Dealing with differences in definitions;
6. Editing of influential errors and outlier detection;
7. Determining the active population;
8. Estimation/imputation methods.

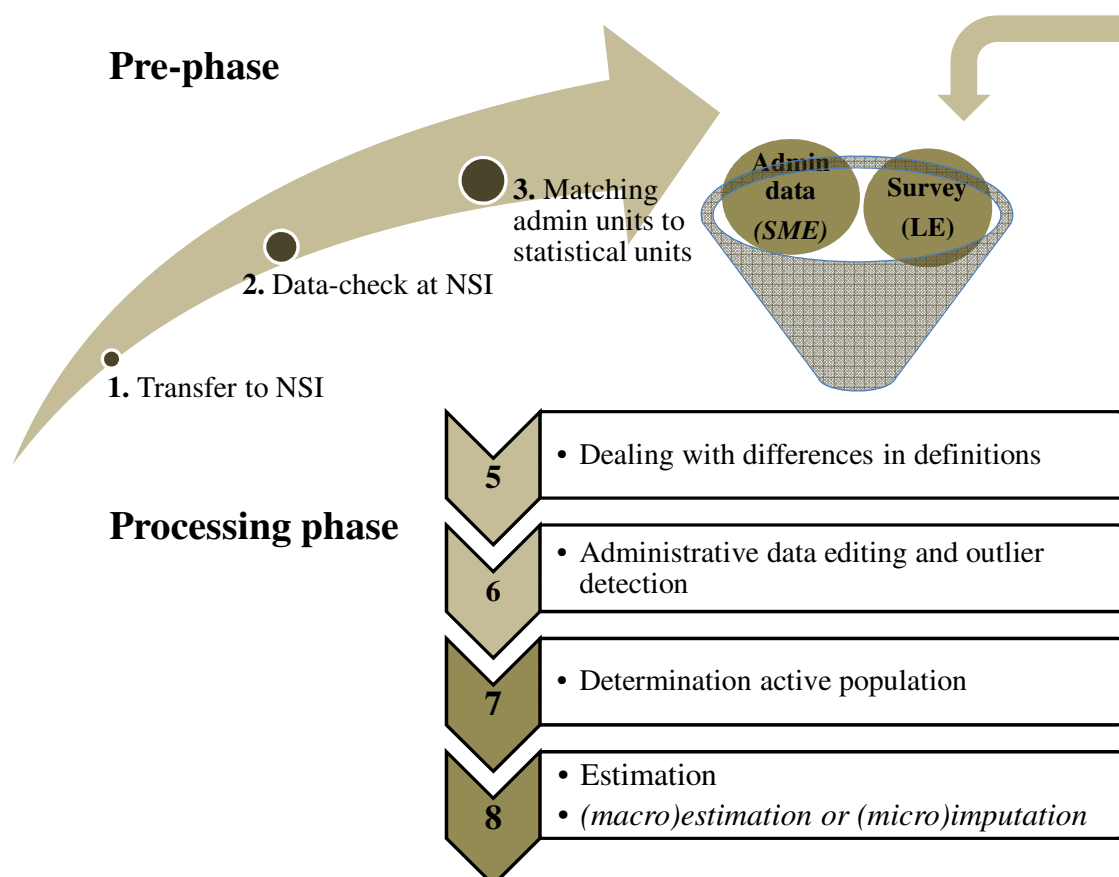


Figure 1. Overview of the statistical process of STS statistics using admin data.

This process is visualised in Figure 1. Note that admin data have to be matched with the business register first (step 3), before combining the results of the LEsurvey with the administrative data used for SME (step 4). Steps 1 until 6 are summarised in this theme. The theme module “Data Collection – Collection and Use of Secondary Data” provides more information about these steps. The remainder of this document is focussed on the determination of the active enterprises and estimation procedures, because both are closely related.

2.5 Findings per process step

Step 1: Data transfer from the tax office to the NSI

To produce annual and quarterly and monthly short-term statistics (STS) with administrative data, the transfer of admin data to a NSI should be guaranteed. Furthermore the NSI must decide whether it opts for only one transfer per month or quarter, or several data transfers per month. The latter allows more flexibility, especially for STS estimates used for internal use (e.g., for National Accounts).

Step 2: Checking the data (completeness, obvious errors etc.)

It is common practice for NSIs to perform elementary checks on the administrative data as soon as they arrive at the NSI. This is in order to check whether there is anything wrong with a specific admin data delivery (e.g., less/more admin data than usual, different distribution than usual, large errors).

Step 3: Matching administrative units to statistical units

This step consists of linking the administrative data to the population frame which is generally defined by the Statistical Business Register (SBR). Theoretically, a 100% match between two frequently used administrative sources (VAT and social security data) and the population for enterprise statistics should exist. In practice, this 100% match is not achieved due to:

- different enterprise units in the SBR and the admin data;
- time-lags, which may cause different timings of starting, stopping, merging and splitting enterprises in admin data and the SBR;
- maintenance peculiarities, which result in differences between administrative data and the SBR; and
- a (slightly) different population coverage because the smallest enterprises are exempted from VAT reporting in some countries.

However, the large majority of enterprises in the SBR should be matched with administrative data on an annual base. If this is not the case, it is recommended to improve this before proceeding further, because the added value of using VAT for turnover estimates is that these tax data are available for all enterprises when the data are complete. Furthermore, if not all administrative data can be matched with the SBR, it may be well possible that the non-linkable units represent specific parts of the populations. When this issue of non-matchable units is not resolved at this processing step, it may lead to estimation problems at a later stage.

When linking admin data to the SBR for STS another important issue is added; the incompleteness of the administrative data. Due to time-lags between the SBR and the administrative data source, late reporting starting enterprises are missed in the first estimates (because they are not yet included in the

SBR). In the case of apparently missing admin data, it is difficult to determine whether this is due to a) late reporting or b) because the enterprise has stopped. This issue is particularly important if the SBR population corresponds with a previous period (e.g., all active enterprises at the end of the previous year). However, the problem also arises if administrative data for the current month or quarter are linked to an up-to-date SBR. This situation is sketched in Figure 2. It will be described in more details in the next section of this module.

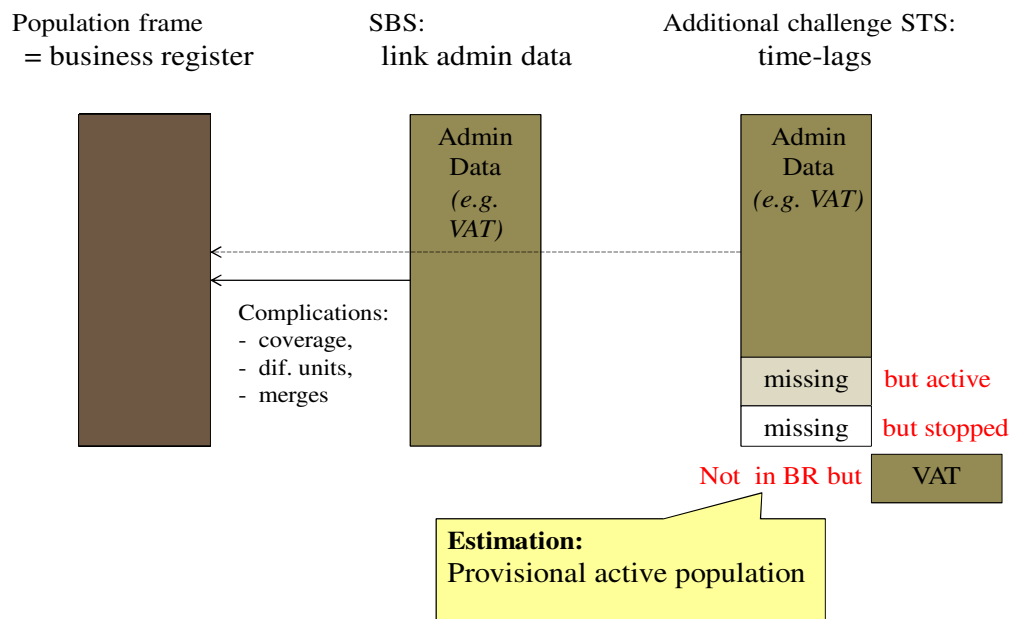


Figure 2. Schematic sketch of a) general challenges when linking admin data to the SBR (middle column) and b) specific challenges for STS when linking incomplete admin data to the SBR.

Step 4: Combining the large-enterprise (LE) survey data with the administrative data for small and medium-sized enterprises (SME)

To ensure stable timeseries, it is recommended to use a ‘frozen’ LEsurvey within a reference year, i.e., the LEsurvey remains unchanged within a year unless an enterprise stops. The use of a ‘frozen’ LEsurvey also prevents enterprises switching from survey to admin data (and vice versa) within a year. The LEsurvey can be updated at the beginning of a new calendar year, at which time new enterprises can be added to LEsurvey and other enterprises may be removed. It is recommend to keep the ‘to be removed’ enterprises within a survey for one extra period in order to maintain the stability of the crucial LEsurvey timeseries.

Step 5 Combining the large-enterprise (LE) survey data with the administrative data for small and medium-sized enterprises (SME)

Definitional differences between VATturnover and STS turnover and administrative variables and statistical variables in general do exist. More information about this topic can be found on the Information Centre of the ESSnet AdminData (<http://essnet.admindata.eu/>).

The most common approach is the use of linear regression analyses to correct for definition differences between administrative data variables and survey variables. These analyses are carried out on observed survey and administrative data of enterprises with similar activities in a reference period, i.e., x years before current year. More specifically, the factor β describing the linear relationship between administrative variables in this reference period like VATturnover and statistical variables like turnover are used to correct the “VATturnovers”. Note, however, that this technique is applicable only if:

- the relationship between administrative data and survey variables is linear;
- the relationship between administrative and survey data is constant in time;
- the relationship between administrative and survey data is not dominated by errors or other sources of noise.

A review by the ESSnet AdminData project of current practices in the use of VAT for annual and short-term statistics showed that differences in definition have little impact on level and growth rate estimates for most industrial activity sections. Therefore, several NSIs do not carry out corrections for definition differences for all variables.

Step 6: Administrative data editing and detection of outlier detection

The topic “Statistical Data Editing” in the Memobust handbook describes several methods for statistical data editing and detecting outliers. Specific for administrative data is that many ‘suspicious values’ may be caused by uncertainties in the link between admin data and the SBR. Some of these suspicious values may be more easily (and reliably) resolved at a later stage when more ‘confirmed’ admin and SBR data are available. Especially when using administrative data for STS, it is advisable not to correct too many suspicious values at the first estimates, but to exclude these suspicious values in the first estimates (and consider them as missing) and correct them when more information becomes available.

Furthermore, it can be recommended that a relationship is established between the stratification level used for administrative based estimates and the stratification level used for detecting:

- influential erroneous values which need to be corrected (= data editing);
- influential correct values which need to be considered as ‘unique’ cases when estimating aggregates (= outliers).

If these two stratification levels do differ, it is hard to determine whether a suspicious and outlying values is influential on the estimates or not. Current practices in the use of VAT data for turnover estimates differ in respect of the stratification level at which missing values with group-specific Y_t/Y_{t-1} or Y_t/Y_{t-12} growth rates are calculated.

Some NSIs use detailed groups. Detailed groups have the advantage that they are theoretically more homogeneous, because growth rates may differ for:

- enterprises with (slightly) different activities; and
- enterprises of different sizes.

The disadvantage of using small groups is that the number of available (donor) units may become too small, which increases the effect of outliers etc. Hence, a good outlier filter to detect anomalous growth rates in the available VATdata should be developed if detailed groups are used for imputation.

Disadvantage of this approach is that it is – in practice – for most NSIs impossible to check VATdata structurally at microlevel, due to the enormous dataset and the generally short production time. As a result, the cause for outlying values (errors, change in reporting unit between the reference period and the current period, or a valid economic explanation for a deviating growth rate) remains often unclear. This implies that if too many outliers are detected in small groups and a valid (economic) explanation for these values does not exist, some selectivity in the remaining values used for imputation should be introduced by filtering out all these outliers. To prevent this, and to keep the process more transparent, other NSIs use higher stratification levels (= bigger groups). This has the advantage that the impact of outliers on the imputed ratios is generally smaller because more (donor) admin data are available. For this reason, some NSIs use higher stratification levels (= bigger groups). This has the advantage that the impact of outliers on the imputed ratios is generally smaller because more (donor) admin data are available.

2.6 *Determining the active population*

Statistical registers and frames are described in the topic “Statistical Registers and Frames” of the handbook. Specific for the use of administrative data in enterprise statistics, and especially STS, is the determination of the active population. For example in STS, the most important issue with respect to determining the active population is to detect whether VATdata are missing because:

- the enterprise has stopped (or changed) its economic activity; or
- the enterprise is a late reporter.

This is especially a problem for small enterprises. Larger enterprises are generally well-recorded and are usually quickly updated in the SBR. This problem is enhanced by the possibility that enterprises do not always report their closure immediately to the Chambers of Commerce and/or the tax office. Therefore the SBR might include them improperly for a long time after their closure. Alternatively, the tax register may apply different rules to the NSI for declaring the administrative unit (enterprise) dead. For example, the tax authority may need to keep the enterprise alive until all outstanding transactions between it and the tax office are completed.

A common method for determining whether enterprises are still active is simply to check whether the enterprise has reported any turnover to the tax office for the last few months. When the enterprise has not reported any turnover to the tax office for the last x months (in the case of monthly reporting) or the last x quarters (in the case of quarterly reporting), the enterprise is considered inactive, otherwise it is considered to be still active. The ESSnet Admin Data has tested different rules and suggests that x should be chosen to be larger than 1, in order to minimise errors in the active population estimate due to late reporting. The most suitable values for x seem to be 2 or 3.

Detecting starting enterprises in time is also an issue. Starting enterprises are reported by the Chambers of Commerce and/or the tax office. Subsequently they are included in the SBR, with some delay after they started. If this delay is small, the starting enterprises should be present in the SBR. If this delay is longer, some of the starting enterprises might not be included in the SBR but in the

VATdata. However these enterprises, when present in the VATdata source, can be included in the population (and in the estimates) provided that a reliable NACE code can be obtained from the administrative data source or elsewhere.

This ESSnet has analysed the impact of starters and stoppers on the estimates. More specifically, it has analysed the impact of:

- incorrectly assumed active enterprises; and
- missing starters

on revisions in growth rate between the preliminary and final estimate. These analyses have been performed on VATdata in Estonia, Finland and Germany and on social security data in Italy. The conclusions are similar:

1. The contributions of starting and stopping enterprises do not average out. This may lead to a systematic over- or under-estimation (bias) in the preliminary STSestimate.
2. The relative contribution of starting and stopping enterprises to the total revisions is large, compared with the small share of starting and stopping enterprises to the total estimate.

These conclusions seemed to be independent of the exact estimation methodology, because the bias effect differs per period. Therefore, it is recommended that NSIs invest in the development of a suitable and efficient approach to determine the active population, before using administrative data for STSestimates. For annual statistics, this challenge has less impact because the administrative data are complete. For these statistics, it is however still crucial that all admin data can be linked to the SBR to prevent estimation challenges due to selectivity problems of the ‘matched’ administrative data which may vary per year.

2.7 Estimation: available administrative data when the estimates have to be made

As previously mentioned, the most frequently used administrative data sources to replace survey are VATdata for monthly, quarterly and annual turnover estimates. VAT may be reported on a monthly, quarterly or annual basis to the tax office. It has to be reported between 30 and 40 days after these reporting periods. The thresholds between these obligatory reporting periods differ per country, but as a rule of thumb it can be stated that:

- Large enterprises report VAT on a monthly base. Monthly reporting is also common for enterprises expecting a VATrefund;
- Most enterprises report VAT on a quarterly basis;
- Only very small enterprises (those having a turnover less than a few thousand Euros) are annual VAT reporters. In some countries, these enterprises do not need to declare VAT at all.

Therefore, it can be concluded that:

- A. VATdata are complete when the annual estimates are produced. This is because all monthly, quarterly and annual VAT has been reported. VATdata are only missed due to matching problems with the SBR.
- A. VATdata provide generally very good coverage for quarterly estimates which are produced (more than) 45 days after the quarter.

This is because all monthly and quarterly VAT has been reported and the share of the missing annual reporters is very limited (in general < 5%). Hence, the available VAT should provide information about almost all enterprises apart from the smallest ones and some matching problems with the SBR.

Altogether is the implications that estimation is only needed for relatively few missing enterprises when producing annual and quarterly statistics with VAT.

- B. No or few (and selective) VAT data are available at the time the monthly estimates have to be published in most countries (30-45 days after the end of the month).

This is because many enterprises report quarterly or annually and publication deadlines for STS estimates are often earlier than deadlines for VAT reporting to the tax office.

Obviously under these circumstances, other estimation techniques are required than in case almost all admin data are available. This is illustrated in Figure 3.

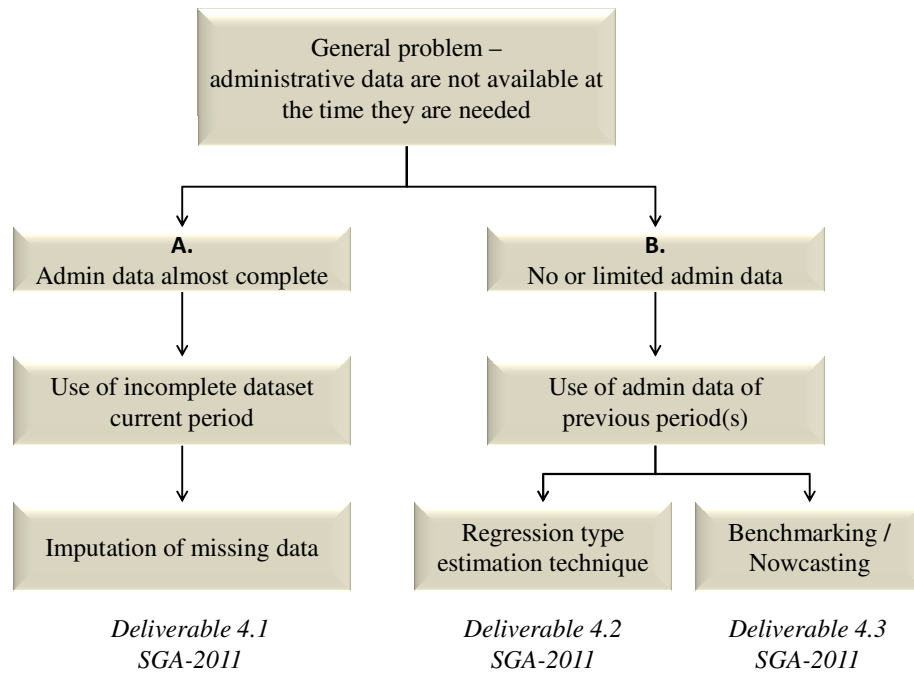


Figure 3. Relationships between estimation techniques when using administrative data as replacement for survey data and available administrative data. This figure also shows the relationship between the techniques and the documentation of the ESSnet Admin Data project (deliverable 4.1 = Maasing et al., 2013; deliverable 4.2 = Kavaliauskiene et al., 2013; deliverable 4.3 – Vlag et al., 2013).

Note VAT reporting periods differ from standard month, quarter and year in a few countries (United Kingdom, Ireland and Iceland). In this case, the VAT data have to be calendarised into values that cover the standard intervals such as month, quarter and year (Maasing et al., 2013). As a consequence, the timeliness of calendarised VAT is worse and a complete set of calendarised VAT is only available for the annual turnover estimates (Vlag et al., 2013).

Pragmatically the ESSnet Admin Data has used 80% coverage as the lower limit for “good coverage of admin data”. This threshold has been chosen because many NSIs require that the large enterprises survey and the available admin data together cover about 80% of the total turnover before reliable turnover estimates can be published. This 80% threshold is arbitrary. Revision analyses, as performed by Langford and Teneva (2013) or by Baldi et al. (2013), can provide more objective criteria to determine whether or not a dataset is almost complete and representative.

In the remaining part of this module, estimation techniques will be discussed in case of:

- almost complete coverage of the available administrative data;
- few administrative data available.

2.8 *Estimation in case of almost complete coverage of administrative data*

As mentioned in section 2.3, the most common practice is to produce level estimates and micro-imputations when using VAT for STS. This practice is recommended when the VAT (or other administrative data sources) are almost complete. The consequence is that the few missing VAT data need to be imputed. The most common practice for imputing missing values is using the available VAT turnover data from enterprises with plausible data for current period. To impute missing values, enterprises are divided into several groups by size class, activity (e.g., NACE code) and in some cases the period of VAT declaration (monthly/quarterly payers). These groups are the so-called stratification groups. The main assumption behind this stratification is that within these groups the available ‘average’ VAT data are representative for the enterprises without data.

Current practices differ with respect to the exact imputation technique, but the two most common imputation techniques are:

1. Average growth rates between current period and previous period of available VAT data. The imputation of the variable y for unit (enterprise) i at month t belonging to a generic stratification group is in formula:

$$\hat{y}_{it} = y_{it-1} \frac{\sum y_{jt}}{\sum y_{jt-1}} = y_{it-1} \frac{Y_t}{Y_{t-1}}$$

where the summation is on the units reporting in both t and $t-1$ belonging to the same stratification group. These are the so-called Y_t/Y_{t-1} imputations.

2. Average growth rates between current period and the corresponding period of the previous year of available VAT data. In formula:

$$\hat{y}_{it} = y_{it-12} \frac{\sum y_{jt}}{\sum y_{jt-12}} = y_{it-12} \frac{Y_t}{Y_{t-12}}$$

These are called Y_t/Y_{t-12} imputations.

The main advantage for choosing Y_t/Y_{t-12} imputations is that these ratios should provide more robust estimates in case of strong seasonality patterns. The disadvantage of using Y_t/Y_{t-12} growth rates for imputation is that these growth rates are affected by changes in activity between t and $t-12$. This is especially true when using detailed stratification levels and therefore relatively few $t; t-12$ growth rates of enterprises with change of activity are available. Another disadvantage is that enterprises which

started during the last 12 months cannot be taken into account when imputing missing data, which may lower the quality. If these disadvantages are dominant, Y_t/Y_{t-1} imputations are preferred over Y_t/Y_{t-12} imputations

The fundamental question is whether the theoretical pros and cons of Y_t/Y_{t-12} versus Y_t/Y_{t-1} lead in practice to differences in publications (Vlag et al., 2012). This question was raised because these techniques are used when at least 80% and generally more than 90% of the estimated VATturnover is available. The same is true for the choice of aggregation levels at which these ratios are calculated.

Testing by the ESSnet Admin Data on VATdata by Statistics Estonia and Statistics Finland and testing on social security data by ISTAT has demonstrated that the impact of the different imputation methods on the published results is negligible, due to the high coverage of the available administrative data combined with the use of a LEsurvey. Hence, when choosing an imputation method one should aim for an optimal trade-off between benefits and costs, rather than aiming for the “best” theoretical quality. The testing of the ESSnet Admin Data also revealed that the impact of different imputation rules on the STSestimates is less than the impact of the uncertain active population on the estimates (i.e., **which** units are to be imputed). Hence, when developing a statistical production system for admin data based STSestimates, it is recommended that research and development should be concentrated on choosing the best method for determining the active population (i.e., **which** units are to be imputed).

2.9 *Estimation in case of few administrative data available*

The work of the ESSnet Admin Data has demonstrated that if few VAT (or other admin data) are available due to timeliness issues, this VAT cannot be used to replace a survey. Main reason is that, analyses in Finland, the Netherlands and the United Kingdom show that the few available VAT is selective and that this selectivity varies in time. As the extend of the selectivity can only be determined afterwards, it is not straightforward to correct for this selectivity with weighting techniques at the time the estimates are needed. Hence, provided that turnover levels and growth rates can be estimated with a later stage, the challenge is to find estimation methods for (early) month when no or only a few VATdata are available. This as alternative for a costly standard monthly survey among all enterprises within a branch.

The application of possible alternatives for a standard survey depends on the observation whether the long-term trends and short-term movements of the timeseries are similar for the larger enterprises, covered by a LEsurvey, and smaller enterprises, covered by administrative data like VAT for quarterly and annual estimates. Depending on the outcome, this information can be used to decide whether for first monthly estimates:

- a small survey under small medium-sized and small enterprises should be added to the LEsurvey which is also used for quarter and annual. This option is called alternative I in the remainder of this module.
- a survey under the largest enterprises (a LEsurvey) only is sufficient for the (first) monthly estimates, knowing that VAT covering the entire population becomes available at a later stage or for the quarters. This option is called alternative II in the remainder of this module.

- the LEsurvey should be combined with a separate estimate for the smallest enterprises based on extrapolation of the VAT series. This option is called alternative III in the remainder of this module.

When one of these three alternatives have been chosen, all alternatives do have in common that VAT of previous periods is used as auxiliary information for the estimates. Note that alternative I uses a small survey. Alternatives II and III are im- or explicitly model-based. Basically, the latter two alternatives provide implicitly a temporal estimation for growth of small and medium sized enterprises. This temporal estimation is ‘overwritten’ as soon as sufficient VAT data become available.

Alternative I is described by Kavaliauskiene et al. (2013). The basic idea is that VAT is too late to produce turnover estimates for the current month. As alternative a mini-survey for current month t is weighted by using VAT of previous period $t-1$ as auxiliary information. This can be done by using GREGtype estimators. The use of GREGtype estimators in combination increases the precision of the estimates. Hence, a smaller survey can be used compared to the Horvitz-Thompson estimator. The use of GREGtype estimators is an established technique, which provides acceptable results. Disadvantage of the method is that it is elaborative in terms of detection and handling of outlying values. Furthermore, the reduction of the survey might be limited as this method requires a minimum amount of data.

As a result, the decision whether the smaller enterprises should still be sampled for estimates until VAT becomes available (alternative I) or whether temporal estimations for small enterprises can be considered (alternatives II and III) basically depends on five factors:

- the target of a National Statistical Institute (NSI) to reduce production costs;
- the target of a NSI to reduce administrative burden;
- the desired quality;
- the output level;
- the risk factor of using temporal estimations in case of unforeseen circumstances.

In general it can be stated that higher targets of reducing production costs and administrative lead to a lower survey coverage. Quality and output level may generally lead to a larger coverage of the surveyed part. However, this is not necessarily correct because if the sample size becomes too small, a survey estimate may have a large imprecision and a temporal estimation may have a better quality. The Standard Mean Error (SME), a combination of bias and imprecision, may help to compare the quality of a (small) survey estimate with temporal estimation (Vlag et al., 2013).

Alternatives II and alternatives III provide a temporal estimation for the smaller enterprises for current period t without using survey and VAT data (as the latter are not available yet).

Alternative II is based on the assumptions that:

- the short-term movement of the growth of the non-surveyed small enterprises is similar to the short-term movement of the surveyed large enterprises;
- changes in the business cycle and sudden events are simultaneously registered in the surveyed and non-surveyed part.

Alternative III is based on the assumptions that:

- the short-term movement of the growth of the non-surveyed small enterprises differ from the short-term movement of the surveyed large enterprises due to time-lags. Time-lags may occur if the small enterprises within a branch supply goods and services to larger enterprises or a subcontractors of larger enterprises;
- changes in the business cycle are differently recorded in the surveyed and non-surveyed part.

Note that available VAT of previous periods are needed to test whether the above mentioned assumptions are valid or not. Hence, although is not directly used in alternatives II and III, VAT is implicitly used for estimation.

If the growth of the larger enterprises is related to the growth of the smaller enterprises (= alternative II), the total estimation can be determined by

$$G_{t,t-1} = \frac{Y_{MLE} \cdot G_{t,t-1;MLE} + Y_{SE} \cdot G_{t,t-1;MLE} \cdot C}{Y_{MLE} + Y_{SE}}$$

with:

$G_{t,t-1}$, $G_{t,t-1;MLE}$ the growth rates of the entire target population and the surveyed medium and large enterprises (MLE) respectively;

Y_{MLE} , Y_{SE} the(extrapolated) turnover level for MLE and small enterprises (SE), respectively;

C factor to correct for systematic differences in growth between MLE and SE.

The most simple model is assuming that $C = 1$. In more sophisticated models C is basically based on bias corrections or by benchmark nowcasting (Fortier et al., 2007; Brown et al., 2012; Vlag et al., 2013).

If the growth of the larger enterprises is not related to the growth of the smaller enterprises (=alternative III), the total estimation can be determined by

$$G_{t,t-1} = \frac{Y_{MLE} \cdot G_{t,t-1;MLE} + Y_{SE} \cdot G_{t-x,t-x-p;SE} \cdot C}{Y_{MLE} + Y_{SE}}$$

with:

$G_{t-x,t-x-p}$ the growth rates of the SE of a previous period;

C correction factor, in this case basically a nowcasting factor.

The most simple model is assuming that $p=1$ and $C=1$. In this case the growth rate of previous period is applied to the current period. Several timeseries models exist to determine C , including Holt-Winters, ARIMA and SSA nowcasting techniques.

The approaches and underlying models are sketched in Figure 4.

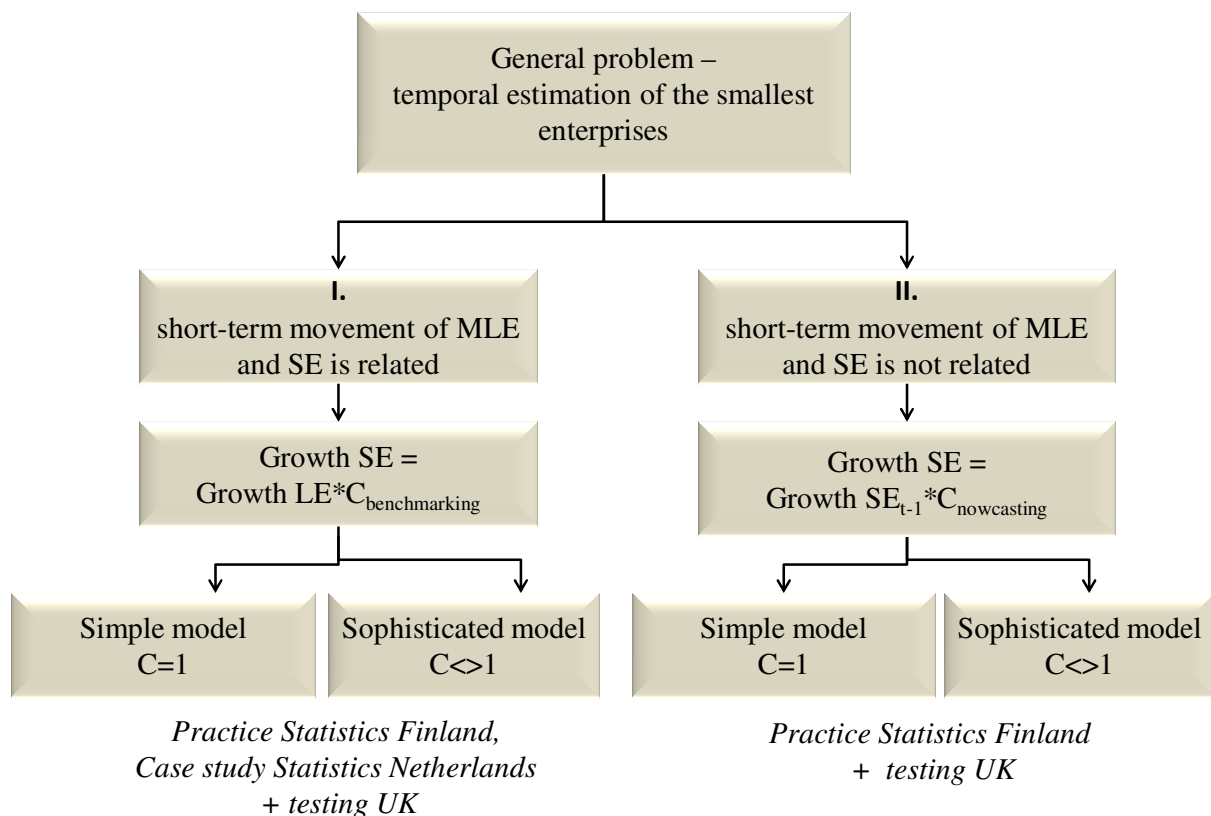


Figure 4. Simplified sketch of temporal estimation methods for small enterprises depending on the relationship between growth rate of large enterprises versus small and medium sized enterprises. Note that the ESSnet Admin Data project has tested these alternatives on data in Finland, the Netherlands and the United Kingdom.

The ESSnet Admin Data has analysed and described three cases (on real data) in which alternative II is used and two cases (on real data) in which alternative III is used. It is beyond the scope of this theme to describe the findings in details. Therefore, only the most important considerations are summarised.

In general these methods provide acceptable results. This especially the case for alternative II, which is also more data-based. However, even for alternative II, it can never be excluded that the underlying assumptions for the temporal estimations are invalid in case of unexpected changes in the business cycle or sudden events. If the surveyed part of the enterprise population (LEsurvey) covers 70-80% of the turnover and analysis on longseries of historical survey or VATdata demonstrate that maximum difference in growth rates between the surveyed part and the non-surveyed part is limited (e.g., a few per cent points), the potential impact of an incidental less performing temporal estimation for small enterprises is limited on the published total estimate for a branch. In this case a National Statistical Institute may accept the risk of an incidental less performing temporal estimation. However, the risks may be considered as unacceptable if the potential impact of an incidental less performing temporal estimation on the total estimate is larger. Risks may be high if the coverage of the LEsurvey is small and historical data suggest that the maximum difference in growth rates between the surveyed part and

the non-surveyed part might be large. Therefore, it is recommended to perform risk analysis before determining the size of small enterprise part which is temporal estimated.

Another finding that whatever alternative is used, artefacts in the (implicitly) extrapolated VATdataset can easily be magnified, leading to erratic temporal estimates of small enterprises. Therefore, is recommended to:

- consider the use of index series, which are panel based series, rather than level based series, i.e., that is using all available data;
- to spend time for correcting ‘previous’ VATdata for outliers, level shifts and other irregularities when implicitly using these data of previous periods for 1st estimates.

3. Design issues

4. Available software tools

Several productions system do exist for producing statistical estimates with VAT and/or social security administrative data. For more information, we refer to Maasing et al. (2013) and references herein.

Statistics Canada has developed SASprocedures for benchmarking and benchmark-nowcasting. The module “tempdisagg” in R can also be used for benchmarking, benchmark-nowcasting and other nowcasting techniques.

5. Decision tree of methods

The first decision to be made is whether administrative data can be used to replace surveys. This decision may depends whether:

1. the NSI has legal access to these admin data;
2. the data transfer from the tax authorities to the NSI is guaranteed;
3. the NSI is able to process large amounts of (administrative) data in a short time;
4. the administrative data can be linked to statistical business register (SBR).

Then decisions have to be taken whether the aim is:

1. to produce estimates for population totals (and implicitly also for growth rates) or to produce growth rates only;
2. the estimates produced at the microlevel, i.e., for individual enterprises, or at the macrolevel, i.e., using combinations of activities and size classes.

In the next step decisions have to be made about:

1. matching the administrative data to the SBR;
2. dataediting and outliers detection;
3. the stratification level at which estimation and detection of influential erroneous or outlying values takes place;

4. the determination of the target population, i.e., active enterprises.

In the next step one has to analyse the completeness and selectivity of the available administrative data when the estimates have to be made, because it determines:

1. whether imputation techniques using growth rates (or levels) of available VAT of current period can be used to impute few missing VAT data need to be imputed;
2. whether available VAT for current period cannot be used for estimations and estimation have to be based on indirect and implicit use of VAT of previous periods.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Baldi, C., Tuzi, D., Ceccato, F., Pacini, S., Karus, E., and Vlag, P.A. (2013), STS-estimates based on admin data: dealing with revisions. *Deliverable 4.3 of the ESSnet on AdminData – SGA2011*, <http://essnet.admindata.eu>.
- Brown, I. (2012), An Empirical Comparison of Benchmarking Methods for Economic Stock Time Series. *Proceedings ICES-IV conference- June 18-21, 2012, Montreal, Quebec, Canada*, 399–412.
- Constanzo, L. (2013), Report to Eurostat on the “Overview of Existing Practices”. *Deliverable 1.2 of the ESSnet on AdminData -SGA2011*, <http://essnet.admindata.eu>.
- Fortier, S. and Quenneville, B. (2007), Theory and application of benchmarking in Business Surveys. *Proceedings ICES-III conference- June 18-21, 2007, Montreal, Quebec, Canada*, 422–434.
- Kavaliauskiene, D., Slickute-Sestokiene, M., and Vlag, P.A. (2013), The use of regression estimators for admin data based STS estimates. *Deliverable 4.2 of ESSnet AdminData – SGA2011*, <http://essnet.admindata.eu>.
- Langford, A. and Teneva, M. (2012), Analysis of revisions of admin data based short term statistics. Application to UK retail sales data and implications for the definition of the boundary between survey and administrative data coverage. Internal report ESSnet AdminData (upon request).
- Maasing, E., Remes, T., Baldi, C., and Vlag, P.A. (2013), STS estimates based solely on administrative data: final results and recommendations. *Deliverable 4.1 of the ESSnet on AdminData*, <http://essnet.admindata.eu>.
- Vlag, P.A. (2012), Imputing missing values when using administrative data for short-term enterprise statistics. Paper for UNECE work session on Statistical Data Editing, Oslo.
- Vlag, P.A., Bikker, R., de Waal, T., Toivanen, E., and Teneva, M. (2013), Extrapolating admin data for early estimation: some findings and recommendations for the ESS. *Deliverable 4.3 of the ESSnet on AdminData*, <http://essnet.admindata.eu>.

Waal, A.G. de, Vlag, P.A., Baldi, C., and Tuzi, D. (2012), The use of administrative data for STS. Situation I: Good coverage provided by administrative data. *Milestone of work package 4*, <http://essnet.admindata.eu>.

Interconnections with other modules

8. Related themes described in other modules

1. Overall Design – Overall Design
2. Statistical Registers and Frames – Main Module
3. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
4. Dynamics of the Business Population – Business Demography
5. Data Collection – Collection and Use of Secondary Data
6. Statistical Data Editing – Main Module
7. Statistical Data Editing – Editing Administrative Data
8. Weighting and Estimation – Main Module
9. Quality Aspects – Revisions of Economic Official Statistics

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.7: Calculate aggregates

Administrative section

14. Module code

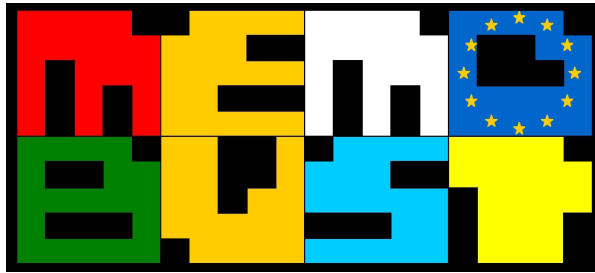
Weighting and Estimation-T-Estimation with Administrative Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.2	23-10-2013	draft version	Pieter Vlag	CBS
0.3	07-02-2014	revised after comments by Editorial Board	Pieter Vlag	CBS
0.3.1	12-02-2014	revised after comments Leon Willenborg and Sander Scholtus	Pieter Vlag	CBS
0.3.2	12-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:36



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Quality of Statistics

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Quality of statistics and its dimensions	3
2.2 Quality and risk management of statistics	5
2.3 Relevance	5
2.4 Accuracy.....	6
2.5 Reliability	9
2.6 Timeliness	9
2.7 Punctuality.....	9
2.8 Coherence.....	9
2.9 Comparability	10
2.10 Accessibility	10
2.11 Clarity.....	11
3. Design issues	11
4. Available software tools.....	11
5. Decision tree of methods	12
6. Glossary.....	12
7. References	12
Interconnections with other modules.....	14
Administrative section.....	15

General section

1. Summary

Quality may be defined as “the degree to which a set of characteristics fulfils requirements” using the much cited ISO standard 9000 (2005). This is valid also for quality of statistical output. The European Statistical System (Eurostat, 2011, principles 11-15; EU, 2009a) uses nine major quality characteristics of statistical output: *relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity*.

Accuracy is generally considered to be a key measure of quality. Total survey error is a conceptual framework describing errors that can occur in a sample survey and the error properties. It may be used as a tool in the design of the survey, working with accuracy, other quality characteristics, and costs. Accuracy is often measured by the mean squared error (MSE) of the estimator. Error sources are considered one by one to estimate the uncertainty and also to obtain some indication of the importance of that source. The errors arise from: sampling, frame coverage, measurement, non-response, data processing, and model assumptions.

Even if statistics are accurate, they cannot be considered as of good quality if, for instance, they are outdated or cannot be easily accessed or there is conflict with other statistics. The quality may be viewed as a multi-faceted concept. Although a major objective of the survey design may be to somehow ‘optimise’ the accuracy, additional quality criteria such as relevance, timeliness, comparability and coherence, and accessibility and clarity are critical to a survey's quality. There needs to be a balance in line with, for instance, regulations and user needs.

2. General description

There is a lot of literature on quality. Here the emphasis is on quality of the statistical output in the European Statistical System. Some useful references are *ESS Handbook for Quality Reports* (Eurostat, 2009b and 2013c), *European Statistics Code of Practice* (Eurostat, 2011), *Handbook on Data Quality Assessment Methods and Tools* (Eurostat, 2007), and *Quality Assurance Framework of the European Statistical System* (Eurostat, 2012).

The description of managing data quality by Brackstone (1999) gives a somewhat broader perspective. Eurostat (1997) has focus on quality reports and provides examples from business statistics. This module is general rather than focused on business statistics, but there are references to other handbook modules. Some characteristics of business surveys and business statistics are described in the handbook modules “Overall Design – Overall Design”, “Repeated Surveys – Repeated Surveys”, and “Weighting and Estimation – Design of Estimation – Some Practical Issues” with connections, for instance, to survey design and successive improvements.

2.1 Quality of statistics and its dimensions

Quality of statistics refers to the degree to which the characteristics of statistics fulfil the requirements of users of statistical information.

In the European Statistical System (ESS), the characteristics of statistics are referred to as quality criteria, quality dimensions or quality components. The product quality dimensions defined by

Eurostat in the European Statistics Code of Practice (Eurostat, 2011) principles covering statistical output are mentioned and defined in Table 1.

Table 1. Quality dimensions of statistics, the associated objects and their definitions.

Nr	Quality dimension	Associated object	Definition
1	Relevance	Concept	The degree to which statistical outputs meet current and potential user needs.
2	Accuracy	Data	The closeness of estimates to the true values.
3	Reliability	Data	Closeness of the initial estimated value to the subsequent estimated value.
4	Timeliness	Release of statistical output	The length of time between the event or phenomenon the statistical output describe and their availability.
5	Punctuality	Release of statistical output	The time lag between the date of the release of the data and the target date on which they were scheduled for release as announced in an official release calendar.
6	Coherence	Concepts and methods	The degree to which the statistical processes by which statistics were generated used the same concepts – classifications, definitions and target populations – and harmonised methods.
7	Comparability	Concepts and methods	The degree to which the same data items can be compared but for different reference periods or different sub populations (regions or domains).
9	Accessibility	Statistical output	The ease and conditions under which statistical information can be obtained.
9	Clarity	Metadata	The extent to which easily comprehensible metadata are available, where these metadata are necessary to give a full understanding of the statistical data.

More criteria of statistics could be added such as reproducibility, level of detail, plausibility, completeness, periodicity and availability (Van Nederpelt, 2009). However, we will not elaborate these characteristics in this document. These criteria are less current.

2.1.1 Statistics

The term statistics can be subdivided into the following objects or components:

1. The concept of the statistical output (concept) and the methods used to compile the statistical output (method)
2. The values of the statistical characteristics (data)
3. The release of the statistical output (release)
4. Statistical output: a combination of data and metadata (statistical output)
5. The description of the statistical output (metadata)

2.1.2 *Focus areas*

Each quality dimension is associated with one or more of these five abovementioned objects (Table 1). A combination of a quality dimension and an object is called a focus area, e.g., accuracy of the data (cf. handbook module “General Observations – Quality and Risk Management Models”).

The concept of focus areas makes it possible to indicate relationships with or dependencies on other focus areas that are related to other objects such as statistical process, administrative data and methodology. These latter objects have their own set of characteristics or quality dimensions. Focus areas are also, e.g., efficiency of the statistical process, timeliness of the administrative data and soundness of methodology.

2.2 *Quality and risk management of statistics*

The quality of statistics is managed by taking the right measures, decisions or actions. Most of these measures are taken in the development stage. However, changes could be necessary in the production stage as well. These measures are necessary for each of the nine quality dimensions of statistical output.

According to the OQRM model (see the module “General Observations – Quality and Risk Management Models”), the following steps can or should be taken to manage quality and risk of each quality dimension:

1. Define requirements for each quality dimension
2. Define and implement quality indicators (measurements, evaluation)
3. Define relationships or dependencies with other focus areas or quality dimensions.
4. Analyse possible causes and effects of problems with a quality dimension (risk analysis)
5. Define and implement measures (decisions, actions) to manage the quality dimension

In the module “General Observations – Quality and Risk Management Models”, these steps of the OQRM model are further elaborated. Quality indicators of output data (step 2) can be found in the specific section of each module of this handbook. Causes of problems (step 4) with the accuracy of the data are described in section 2.4.1–4 about errors.

2.3 *Relevance*

The relevance of statistics is the degree to which statistics meet current and potential users’ need (Eurostat, 2013a).

2.3.1 *Assessment of relevance*

Although relevance is not an inherent characteristic of statistical data, it can be evaluated and measured through analysing the data from users’ satisfaction surveys, and recording the data requirements of Commission Regulations, and International Organisations (e.g., IMF, OECD). The point of departure of every statistical survey has to do with recording the users’ needs and the users’ demands on product quality. Maintaining relevance requires keeping in touch with the current and potential users, not only to record their current needs but also to anticipate their future needs. Usually,

data needs are not clearly formulated by users in statistical terms. Thus, a major challenge is to translate data needs in particular topics into likely statistical terms (Brackstone, 1999).

When reporting on relevance, the aim is to describe the extent to which the statistics are useful to, and used by, the broadest array of users. For this purpose, statisticians need to compile information, firstly about their users (who they are, how many they are, how important is each one of them), secondly on their needs, and finally to assess how far these needs are met. There may be information on user satisfaction and possibly on completeness of the statistical information in comparison with regulations. See also the handbook modules “User Needs – Specification of User Needs for Business Statistics” and “Evaluation – Evaluation of Business Statistics”.

2.4 Accuracy

The accuracy is defined as the closeness of estimates to the unknown true values (Eurostat, 2009b).

Commonly, the objective of a statistical survey is to estimate a set of target parameters referring to a target finite population. Within the framework of quality, *accuracy* of estimates is generally considered a key measure of quality.

2.4.1 Total survey error and mean squared error

A conceptual framework for accuracy is the *total survey error*, which describes, ideally, the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data (Biemer, 2010). For quantifying the total survey error, the most common metric approach is the *mean squared error (MSE)*. Each estimate computed from the survey data has a corresponding estimated MSE.

The total survey error accumulates all errors, which may arise in the sample design, data collection, processing and analysis of survey data, and it comprises both *sampling* and *not sampling errors*.

The mean squared error of an estimator of a population parameter is defined as the hypothetical average of the squared differences between the repeated estimates – when the survey is repeated with sampling, data collection, coding, editing etc. – and the true value of the parameter. In statistical terms, the MSE is the expected squared difference between an estimator and the parameter which is intended to estimate. The mean squared error is equal to the square of the bias plus the variance of the estimator.

2.4.2 Systematic error (bias) and random error (variance)

Accuracy in the general statistical sense denotes the closeness of estimates to the (unknown) exact or true values. Statistics are (nearly) never identical to the true values because of variability (the statistics change from implementation to implementation of the survey due to random errors and effects) and bias (the average of the estimates from each implementation is not equal to the true value due to systematic errors and effects):

- The bias of an estimator equals the difference between its expected value and the true value. Systematic differences may, for instance, be due to systematic measurement errors or systematic effects of non-response that are not overcome in the estimation procedure. The systematic error is the systematic deviation of the estimated value from the true value: the target.

- The variance of the estimator is a measure of the accumulated random errors. The term precision is sometimes used, in general or especially for the square root of the variance.

The total survey error accumulates all errors, which may arise in the sample design, data collection, processing and analysis of survey data, and it comprises both *sampling* and *non-sampling errors*. Both error categories are subject to variability as well as bias.

2.4.3 Sampling errors

Sampling error is that part of the difference between an estimate of a population value and the true value, which is due to the fact that only a subset of the population is selected for the survey.

2.4.4 Non-sampling errors

Non-sampling errors are errors in estimates which cannot be attributed to sample fluctuations. They arise mainly from misleading definitions and concepts, frames that have delays or are inadequate, unsatisfactory questionnaires, defective methods of data collection, non-response, coding, and tabulation. The non-sampling errors appearing to all statistical processes can be categorised as:

- Coverage errors
- Measurement errors
- Non-response errors
- Processing errors

2.4.4.1 Coverage errors

Coverage errors are caused by a failure to cover adequately all units of the target population, which results in differences between the frame population and the target population. We can distinguish the following types of coverage error:

- *Over-coverage* means that units accessible via the frame do not belong to the target population. In business surveys, the *over-coverage* mainly has to do with units (e.g., enterprises) that were included in the business register, they were selected in the sample, but they were not actually existing at the time of the survey (closed enterprises). The decrease of the number of useful sampling units from the initial to the actual size inflates the variance of the parameter's estimate. See the handbook module "Weighting and Estimation – Main Module".
- *Misclassification* is (erroneous) classification of a unit into a category in which the unit does not belong. For instance, a business is classified in Trade instead of Industry. Due to problems of *misclassification*, a number of sampling units turn out to belong to domains of estimation that differ from their design strata. Such units and changes can be handled in the estimation, for instance using post-stratification. See the handbook module "Weighting and Estimation – Main Module".
- The *under-coverage* refers to units which belong to the target population but are not in the frame population. This may, for instance, be due to reporting delays to the business register. Corrections and weighting for *under-coverage* is difficult, because the information cannot be obtained from the sample itself, but only from external sources. See the handbook module "Weighting and Estimation – Main Module".

2.4.4.2 Measurement errors

Measurement errors occur during the data collection, and they mean that the recorded values of variables are different from the true ones.

Their causes are commonly categorised as:

- *Survey instrument*: Questionnaire or measuring device used for data collection may lead to recording of wrong values. Also, the survey mode (CAPI, CATI, CAWI, etc.) can be a potential error source. A wrong mode for a survey could generate, for example, unit or item nonresponse.
- *Respondent*: Respondents may, consciously or unconsciously, provide erroneous data.
- *Interviewer*: Interviewers may influence the answers given by respondents in a way that leads to measurement errors.

Hence, survey results are affected by measurement errors, which occur in the course of the observation of the data. Generally, they can be regarded as random errors, which increase the variance, or as systematic error, which influence the bias. The extra variance (for instance, interviewer variance) due to measurement errors is important to measure in order to assess the effect on the total survey error. See, for instance, the handbook module “Response – Response Process” for the importance and for methods to work with measurement errors.

2.4.4.3 Non-response errors

Non-response errors occur when the survey fails to collect the data as intended, with regard to statistical units and items. The difference between the statistics computed from the collected data and those that would be computed if there were no missing values is the *non-response error*. There are two types of non-response:

- *Unit non-response*, which occurs if there is no information from the statistical unit (respondent) or if the information provided is so limited or possibly erroneous that it is deemed not usable.
- *Item non-response*, which occurs when a statistical unit (respondent) does not provide some of the requested information, or if some of the reported information is not usable.

The effect of non-response on the produced statistics is that it increases variance and bias. Bias is introduced by the fact that non-respondents may be different than respondents in their values of some survey variables in a systematic way that the estimation procedure does not account for. Variability increases due to decreased effective sample size possibly due to the adjustments made. See the handbook module “Weighting and Estimation – Main Module”.

2.4.4.4 Processing errors

Once data have been collected, a range of processes is performed before the production of final estimates, e.g., coding, editing, checks and corrections, imputation of microdata, and later weighting and tabulating etc. Errors that arise at these stages are called processing errors. For example, in coding open-ended answers, wrong codes may be assigned to occupations or economic activities of enterprises. This applies to manual, semi-automated as well as automatic coding. There may also be mistakes in computer programs and when “moving” data and results. Manual handling under time pressure is risky.

There are both systematic and random processing errors.

2.5 *Reliability*

Reliability is the closeness of the initial estimated value to the subsequent estimated value. The subsequent estimated values relate to the same reference period. It regards revisions of data. There may be several revisions. Hence there may be several measures of reliability, due to different combinations of estimated values. See the handbook module “Quality Aspects – Revisions of Economic Official Statistics”.

Reliability is related to accuracy. However, it does not refer to the true value but to a later estimate. It is also related to the coherence between provisional and final data. The revision size depends on both random errors and possible systematic differences between the estimators. See, for instance, the handbook module “Weighting and Estimation – Main Module” for possible early estimates and also references there.

2.6 *Timeliness*

The timeliness of statistical outputs is the length of time between the event or phenomenon they describe and their availability. This is a quality dimension, which is obvious, and there may be user requests. For example, monthly data must not be available too many months after the reference month.

2.7 *Punctuality*

Punctuality is the time lag between the actual delivery of the data and the target date when it should have been delivered.

2.8 *Coherence*

The coherence of two or more statistical outputs refers to the degree to which the statistical surveys and processes by which they were generated used the same concepts – classifications, definitions, and target populations – and harmonised methods. Coherent statistical outputs have the potential to be validly combined and used jointly. An example of joint use is where the statistical outputs refer to the same population, reference period, and region, but where they comprise different sets of data items (say, employment data and production data).

Comparability may be regarded as a special case of coherence where the statistical outputs refer to the same data items and the aim of combining them is to make comparisons over time, or across regions, or across other domains.

When bringing together statistical outputs, the errors occurring (i.e., lacks of accuracy) in the surveys and processes have the potential to cause numerical inconsistency of the corresponding estimates. This can easily be confused with a lack of coherence/comparability. In some cases the estimation procedure eliminates such numerical inconsistencies, for instance through calibration or benchmarking. See the handbook module “Weighting and Estimation – Main Module”.

Different categories of coherence are distinguished:

- Coherence of provisional and final statistics (see also reliability above).
- Coherence of short term and long term statistics

- Coherence of statistics in the same domain
- Coherence of statistics of business statistics with national accounts

2.8.1 *Coherence of short term and annual statistics*

In business surveys, an essential point of quality assessment is the coherence between short term and annual statistics. When comparing the annual growth rates of annual and short-term statistics (STS), divergent trends sometimes appear, provoking inconvenience to the users, especially when the target populations and the definitions of the variables coincide between annual and short-term statistics (e.g., turnover and employment between Short Time Statistics and Structural Business Surveys). Reasons for deficiencies in coherence – influential differences in definitions and methodology – need to be studied. Their effects should be assessed.

2.8.2 *Coherence of statistics in the same domain*

Frequently, a group of statistics, possibly of a different type (e.g., in monetary value, in volume or constant price, price indicators) measures the same phenomenon, but from different approaches. It is very important to check that these representations do not diverge too much in order to anticipate users' questions and prepare corrective actions.

2.8.3 *Coherence of business statistics with national accounts*

Finally, in order to advise users on the information source best suited to their needs, it may also be useful to compare survey statistics with national accounts. The methodology used for compiling national accounts would need to be taken into consideration as well the primary data source used and the adjustments made. Divergences in the concepts should also be taken into account.

2.9 *Comparability*

Comparability is the degree to which the same data items can be compared but for different reference periods or different sub populations (regions or domains). Statistics should be coherent in order to be comparable. Three types of comparability are distinguished:

- *Comparability over time:* It refers to the degree of comparability between two or more instances of data on the same phenomenon measured at different points in time.
- *Comparability between geographical domains:* It refers to the degree of comparability between similar surveys measuring the same phenomenon for different geographical domains.
- *Comparability between non-geographical domains:* It refers to the comparability between different surveys results which target similar characteristics in different statistical domains.

2.10 *Accessibility*

Accessibility of statistics is the ease and conditions under which statistical information can be obtained (Eurostat, 2013a). It depends on the physical conditions by means of which users obtain data: where to go, how to order, delivery time, pricing policy, marketing conditions (copyright, etc.), availability of micro- or macrodata, various formats and media.

To achieve the accessibility of information, the following three principal aspects need to be fulfilled (United Nations, 2003):

- A catalogue system, which allows the users to find out what information is available and assist them to locate it.
- A delivery system, which provides access to information through distribution channels, and in formats, that suit users.

The traditional printed catalogue has given way to on-line catalogues of statistical products, linked to metadata bases in which the characteristics of the information can be found. Access to the catalogue system can be through the Internet, and users who find what they want can immediately place an order to request the desired information or retrieve the information themselves. On-line databases, accessible by internet are the dominant component of the delivery system.

2.11 Clarity

Clarity is the extent to which easily comprehensible metadata are available (for the user), where these metadata are necessary to give a full understanding of statistical data (Eurostat, 2013a). It is determined by the information environment within which the data are presented, whether the data are accompanied with appropriate metadata, whether use is made of illustrations such as graphs and maps, whether information on accuracy and other quality aspects are available (including any limitations on use) and the extent to which additional assistance is provided by the producer

According to the United Nations (2003), the clarity of statistical information is primarily achieved by providing users with metadata, which help them to properly interpret the produced statistical information. The information needed to understand statistical data has to do with (United Nations, 2003):

- The concepts and classifications that underlie the data (what has been measured).
- The methodology used to collect and compile the data (how it was measured).
- The accuracy measures of the data (how well it was measured).

Quality information and indicators for other dimensions than accuracy could be added to this list. For instance, some information on comparability and coherence may be important.

These elements could be compiled in a quality report (EU, 2009b; Eurostat, 2009a) or as explanation of a statistical table.

3. Design issues

4. Available software tools

The *ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys* (Eurostat, 2013b) presents variance estimation and many software packages (in its Appendix 7.5) available which can calculate variance estimates for linear and non-linear statistics under simple and complex sampling designs. Its focus is household statistics having sampled individuals, but there are general texts and useful information also for business statistics.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Biemer P. P. (2010), Total Survey Error. Design, Implementation and Evaluation. *Public Opinion Quarterly* **74**, 817–848. <http://poq.oxfordjournals.org/content/74/5/817.full.pdf+html>
- Brackstone, G. (1999), Managing Data Quality in a Statistical Agency. *Survey Methodology* **25**, 139–149. <http://www.statcan.gc.ca/pub/12-001-x/1999002/article/4877-eng.pdf>
- EU (2009), Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities. Also referred to as “StatLaw”.
<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:en:PDF>
- Eurostat (1997), *Model Quality Report in Business Statistics*. Four volumes (methodological overview, variance estimation and software, examples from Sweden and the UK, and implementation guidelines) from a European project. Accessible on line, e.g., (retrieved on 2014-01-29) <http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/MODEL%20QUALITY%20REPORT%20VOL%201.pdf>
- Eurostat (2007), *Handbook on Data Quality Assessment Methods and Tools*. Editors: Manfred Ehling and Thomas Körner.
<http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf>
- Eurostat (2009a), *ESS Handbook for Quality Reports*. Methodologies and Working papers, European Commission. See next edition Eurostat (2013c).
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/EHQR_FINAL.pdf
- Eurostat (2009b), *ESS Standard for Quality Reports*. Methodologies and Working papers, European Commission.
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf
- Eurostat (2011), *European Statistics Code of Practice. For the National and Community Statistical Authorities*. Adopted by the European Statistical System Committee 28th September 2011.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/CoP_October_2011.pdf
- Eurostat (2012), *Quality Assurance Framework of the European Statistical System (QAF)*. Version 1.1. http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF

- Eurostat (2013a), *Eurostat's Concepts and Definitions Database*. Website: http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GLOSSARY&StrNom=CODED2&StrLanguageCode=EN. Retrieved 25 October 2013.
- Eurostat (2013b), *ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys*. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-13-029/EN/KS-RA-13-029-EN.PDF
- Eurostat (2013c), *ESS Handbook for Quality Reports*. Methodologies and Working papers, European Commission. Not available on Internet yet at 1 March 2014. See previous edition Eurostat (2009a).
- ISO 1179 (2004), *ISO/IEC-FDIS 1179-1. Information technology – Metadata registers – Part 1: Frameworks*. International Organization for Standardization, Geneva.
- ISO 9000 (2005), *ISO 9000:2005. Quality management systems – Fundamentals and vocabulary*. International Organization for Standardization, Geneva.
- Nederpelt, P. W. M. van (2009), *Checklist Quality of Statistical Output*. Statistics Netherlands, The Hague/Heerlen. <http://www.cbs.nl/NR/rdonlyres/4119715F-7437-4379-9A70-90A0893F949E/0/2009ChecklistQualityofStatisticalOutput.pdf>
- NQAF (2012), *Glossary*. Compiled by the Expert Group on National Quality Assurance Frameworks. 3 February 2012.
- United Nations (2003), *Managing Data Quality in a Statistical Agency*. LC/ L.1891 (CEA.2003/6). <http://www.eclac.cl/scaeclac/documentos/lcl1891i.pdf>

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Quality and Risk Management Models
2. User Needs – Specification of User Needs for Business Statistics
3. Overall Design – Overall Design
4. Repeated Surveys – Repeated Surveys
5. Response – Response Process
6. Weighting and Estimation – Main Module
7. Weighting and Estimation – Design of Estimation – Some Practical Issues
8. Quality Aspects – Revisions of Economic Official Statistics
9. Evaluation – Evaluation of Business Statistics

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

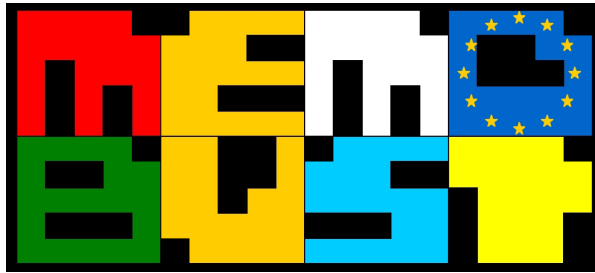
Quality Aspects-T-Quality of Statistics

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	10-03-2012	first version	Ioannis Nikolaidis	ELSTAT
0.1.1	25-10-2013	reviews Norway, Italy, Sweden and the Netherlands processed	Peter van Nederpelt	Statistics Netherlands
0.1.2	20-01-2014	reviews Sweden and Hungary processed.	Peter van Nederpelt	Statistics Netherlands
0.1.3	01-03-2014	reviews EB processed	Peter van Nederpelt	Statistics Netherlands
0.1.4	11-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:27



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Revisions of Economic Official Statistics

Contents

General section.....	3
1. Summary	3
2. General description.....	4
2.1 Definition and classification of revisions	4
2.2 Building a real-time dataset.....	6
2.3 Analysis of revisions	8
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

Macroeconomic indicators are very often revised and the size of revisions, computed comparing subsequent estimates with previous ones, allows to assess their reliability. Along with accuracy, reliability represents a dimension of the statistics quality and is considered in the twelfth of the fifteen principles of the European statistics code of practice (Eurostat, 2011; see also IMF, 2012). Stating that “European Statistics accurately and reliably portray reality”, the code of practice recognises the revision analysis as a tool “... to improve statistical processes”. As stressed in De Vries (2002), accuracy refers to the closeness between the estimated value and the true unknown value and is assessed (when possible) evaluating the error associated with the estimate. On the other hand, reliability refers to the closeness of the initial estimated value to the subsequent up to the final estimated value and is partially assessed comparing estimates over time, i.e., analysing the revisions. In fact reliability is only one indicator of statistical quality and is not completely captured by revision analysis. Other aspects of quality in statistics are established in the Eurostat framework and include timeliness and robustness. Especially for short-term statistics is well known the existing trade-off between timeliness and reliability. For many of these indicators for which seasonally adjusted data are produced it is also important to distinguish between revisions in raw data and revisions in the seasonally adjusted data due to the seasonal adjustment method used.

The importance of the reliability, as one of the dimensions of quality, is confirmed by the growing interest in revisions on official statistics from international organisations (Eurostat, OECD, IMF) and National Statistics Institutes (NSIs). With the aim to produce as much as possible transparent statistics NSIs put efforts describing the revision policy, providing information about past revisions, scheduling future revisions, creating real-time data bases and analysing revisions. Most of these efforts are users-oriented because users see with a certain criticism the fact of revising economic statistics. However, in some cases, revision analysis could be helpful as well for data producer detecting possible “weakness” in the estimation procedures and to suggest suitable measures to counteract them.

Several recommendations have been set by international organisations (OECD, IMF and Eurostat¹) and NSIs (in particular ONS). In particular:

1. revisions should follow a regular and transparent schedule (publicly available);
2. preliminary and/or revised data should be clearly identified;
3. non schedulable revisions due to errors should be communicated as soon as possible by data producers to the general public;
4. real-time databases should be built for performing revisions analysis and made public at least for the main economic indicators;
5. studies and analyses of revisions should be carried out routinely, used internally to inform statistical processes and made public (particularly for short-term statistics).

¹ Recently, Eurostat (2013) has released the guidelines on revision policy for PEEIs (Principal European Economic Indicators) stating eight principles for a common revision policy for European Statistics.

2. General description

The remainder of the module is aimed at presenting the revisions, their sources and causes and the tools to analyse them. It is organised in three subsections. Section 2.1 defines the revisions and presents their classification; section 2.2 describes the real-time datasets and section 2.3 deals with revision analysis and some statistical measures on revisions.

2.1 Definition and classification of revisions

Macroeconomic statistics are typically revised and the size of revisions reflects the trade-off between accuracy/reliability and timeliness. In particular accuracy refers to the closeness between the estimated value and the true value measured by the statistic (usually unknown); reliability refers to closeness of the initial estimate to subsequent (revised) estimates. The latter is measurable and the size of revisions, computed comparing subsequent estimates with previous ones, allows to assess the reliability, though non revised estimates are not to be considered automatically as reliable estimates (see Di Fonzo, 2005). Users often see with a certain criticism the revision of economic statistics. To improve the communication about the revision process many NSIs have made efforts towards transparency, providing information about past revisions, scheduling future revisions both methodological and definitional, classifying revisions, creating real-time datasets where all the vintages are gathered, analysing size, bias and efficiency of revisions.

Hereinafter in this section definitions and classifications of revisions are provided to better understand concepts and practices presented in section 2.2.

Definition of revisions

Although several formal definitions of revisions have been proposed in the literature (for short term economic indicators see Mazzi and Ruggeri Cannata, 2008), here we are interested in providing an analytical definition. In particular, given an indicator and two subsequent estimates referred to a generic period t (month or quarter), a preliminary (or earlier) estimate P_t and a later (more recent) estimate L_t , revision can be defined as

$$R_t = L_t - P_t$$

or, in relative terms,

$$R_t = (L_t - P_t) / L_t.$$

The first definition is exploited to analyse revisions to growth rates (period-on-period or year-on-year growth rates), while the second definition is exploited to analyse revisions to values in level. Since many short term economic indicators are released as index numbers, both users and producers of

² When revision R_t refers to the comparison of growth rates, it may depend on all the estimates exploited to compute the growth rates L_t and P_t , that is the estimates referred to different time points (t and $t-12$, for year-on-year growth rates, and t and $t-1$, for period-on-period growth rates). Using the triangles described in section 3, it is possible to isolate the revisions affecting only one time period, say only t or only $t-1$ ($t-12$). Although it could sound like the correct way to analyse revisions on growth rates, this practice is not followed by NSIs and other international organisations because it does not consider the revisions in the growth rates actually released to users.

official statistics are interested in revisions on growth rates and, consequently, on the first definition of revision.

Classification of revisions

Classifying revisions to official statistics is helpful for both users and producer. Although several bases and criteria have been proposed by NSIs and international organisations, revisions are often classified either by reason or by scheduling (as proposed in Mazzi and Ruggeri Cannata, 2008): the former focuses on the sources or the causes of revisions, while the latter refers to the frequency of revisions.

a) Revisions by reason

- Incorporation of additional data
 - incorporation of late responses (increasing the response rates to surveys)
 - replacement of previous model-based estimates/forecasts with available data (for example in the calculation of national account aggregates early estimates are produced on the basis of models and forecasting techniques and revisions should be expected when more information becomes available)
 - incorporation of data more closely matching concepts and definitions (e.g., more accurate annual data, alignment to annual structural surveys and so on). For example in guidelines for the calculation of index of production it is suggested to use an input method based on hours worked that foresees the calculation of productivity coefficients (calculated as value added per hour worked drawn from national accounts) and their forecast for the current year. Annual changes in these coefficient may lead to revisions in following releases.
- Updating of routine adjustment/treatment or compilation
 - Updating of seasonal factors or time series models exploiting to produce seasonally adjusted data (see the ESS guidelines on seasonal adjustment (Eurostat, 2009) for more details about revision policy of seasonally adjusted data)
 - Change of the base year involving the update of the basket of products, the rotation of business in the sample, the revision of the weighting system
- Introduction of new methods and concepts
 - improvement of estimation methods
 - changes in classifications
 - introduction of new definitions
- Correction of data/estimation errors either caused by the incorrect internal treatment of source data or resulting from wrong information previously provided by respondents and replaced later on (very often after direct contacts with the respondents).

b) Revisions by scheduling (more suitable for short-term indicators)

- Routine revisions generally affecting only the most recent periods
- Annual revisions made when annual (external to surveys) information becomes available and affecting a larger time span (even several years)
- Major revisions recurring at longer intervals (more than three/four years) due to changes of classifications, base period for fixed-based indices, benchmarking and so on. They may require a re-calculation of the whole time series of short term indicators
- Unexpected revisions usually caused by errors or by extraordinary acquisition of new data

Since the main aim of revisions should be the improvement of the estimates previously released, they need to be freely available along with the new statistics and they should be accompanied by supporting and explanatory information aimed at explaining their causes. In order to inform users about this, NSIs should publish general statements describing their practice and policy on revisions accompanying revised statistics, explaining revision sources and describing the effects of revisions.

2.2 Building a real-time dataset

As stated above, the analysis of revision is a tool to assess the quality of the first estimate in relation to later and final estimates. In the recent years, OECD, Eurostat and ONS have stressed that real time datasets (also referred to as *revision triangles*) can represent a useful tool for producers of official statistics to undertake revision analysis and to present revisions and their statistical properties to users. These datasets show how estimates change over time and provide further information about the dissemination policy, timing of revisions, the explanation of revision sources, the status of the published data.

For short-term statistics, a complete history of revisions can be derived collecting the historical vintages of the same indicator in these datasets. According to the definition given by McKenzie and Gamba (2008), a vintage is a “set of data (sequence of values) that represented the latest estimate for each reference point in the time series at a particular moment in time”.

The real time dataset is a table whose rows represent the vintages (identified through their date of release) and whose columns represent the reference periods (months or quarters) of the time series.

As far as the updating of real time datasets is concerned, it is worth stressing two issues: firstly, in order to avoid the loss of information on the revision process, they should be updated whenever a new estimates is available; secondly, when a release is skipped for extraordinary reasons, this missing release should not be replaced by the previous release, because this means to introduce a null revision in the revision process altering its statistical properties.

According to the aim of the analysis to be performed, a revision triangle can be read “horizontally”, “vertically” or “diagonally” (Figure 1). When triangles are read horizontally, they provide time series released at the available dates (such information is useful to analysts interested in assessing their forecasting models). On the contrary when triangles are read vertically, they give the revision history referred to a given period, from the preliminary estimates to the latest (such information measures the

reliability of the earlier estimates). Finally, when triangles are read along the main diagonal (the first sub-diagonal, the second sub-diagonal, ...), they give the time series of the first (second, third, ...) releases.

Release date	Reference Month											
	Jan-09	Feb-09	Mar-09	Apr-09	May-09	Jun-09	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09
Preliminary estimates P	82.5	85.8	90.6	83.6	86.1	87.3	97.6	46.6	93.2	96.2	93.0	77.0
Revised estimate R	81.6	85.4	90.6	83.4	85.9	87.6	97.9	46.4	93.5	96.3	93.1	77.0
Estimate S	81.1	85.4	90.2	83.1	85.8	87.4	H.S.	H.S.	H.S.	H.S.	H.S.	H.S.
Estimate Y1	79.9	84.1	88.8	81.8	84.6	86.2	96.8	46.0	92.5	94.6	91.7	76.1
Estimate Y2	79.4	83.7	88.4	81.3	84.1	85.8	96.3	45.8	92.0	94.1	91.2	75.7
Last estimate L	79.6	83.8	88.4	81.4	84.2	85.9	96.4	45.8	92.1	94.2	91.3	75.8
Mar-09	82.5											
Apr-09	81.6	85.8										
May-09	81.6	85.4	90.6									
Jun-09	81.6	85.4	90.6	83.6								
Jul-09	81.6	85.4	90.6	83.4	86.1							
Aug-09	81.6	85.4	90.6	83.4	85.9	87.3						
Sep-09	81.6	85.4	90.6	83.4	85.9	87.6	97.6					
Oct-09	81.1	85.4	90.2	83.1	85.8	87.4	97.9	46.6				
Nov-09	81.1	85.4	90.2	83.1	85.8	87.4	97.9	46.4	93.2			
Dec-09	81.1	85.4	90.2	83.1	85.8	87.4	97.9	46.4	93.5	96.2		
Jan-10	81.1	85.4	90.2	83.1	85.8	87.4	97.9	46.4	93.5	96.3	93.0	

Figure 1: Example of real time dataset³

Based on the more widespread practice among NSIs and international organisations, the diagonal reading of the triangles is generally used to build time series of revisions on which descriptive analysis is carried out. The idea is that comparing the first estimate with later estimates (for example the second estimate), a time series of homogeneous revisions is derived, that is a time series of revisions having the same features (e.g., routine revisions between the second and the first estimate, in the example the time series of one-step revisions). Since other causes of revisions than routine revisions may occur at lower frequency producing important changes in the estimates, attention should be taken to exclude them from the analysis.

Usually, the real time dataset contains seasonally adjusted data and growth rates on them (Figure 2). However, when users are interested in unadjusted data, triangle should be provided for both unadjusted and adjusted data and revision analysis undertaken on year-on-year growth rates (for unadjusted data) and on period-on-period growth rates (for adjusted data). In fact the seasonal adjustment, causing revisions spanning several years, may mask interesting evidences on the revision process of unadjusted data.

³ For a generic index the preliminary estimate (P), the revised estimate after a month (R), the estimate after a year (Y1), after two years (Y2) and the last estimate (L) are reported. Figure 1 reports a further estimate (S), which is peculiar of the indicator considered in the example, released in October 2009 and revising the period January-June 2009.

Reference month	Growth rate on the same month of the previous year						Revision of Growth rate on the same month of the previous year					
	Preliminary estimate P	Revised estimate R	Estimate S	Estimate Y1	Stima Y2	Last estimate L	R - P (h = 1)	S - P	Y1 - P	Y1 - R	Y2 - Y1	L - P
jan-09	-21.9	-22.8	-23.3	-24.2	-24.7	-24.5	-0.9	-1.4	-2.3	-1.4	-0.5	-2.6
feb-09	-23.7	-24.1	-24.1	-25.0	-25.3	-25.2	-0.4	-0.4	-1.3	-0.9	-0.3	-1.5
mar-09	-18.2	-18.2	-18.6	-19.6	-19.9	-19.9	0.0	-0.4	-1.4	-1.4	-0.3	-1.7
apr-09	-25.4	-25.5	-25.8	-26.7	-27.2	-27.1	-0.1	-0.4	-1.3	-1.2	-0.5	-1.7
may-09	-22.6	-22.8	-22.8	-23.7	-24.2	-24.1	-0.2	-0.2	-1.1	-0.9	-0.5	-1.5
jun-09	-19.7	-19.4	-19.6	-20.4	-20.8	-20.8	0.3	0.1	-0.7	-1.0	-0.4	-1.1
jul-09	-17.5	-17.2	N.S.	-18.2	-18.6	-18.5	0.3		-0.7	-1.0	-0.4	-1.0
aug-09	-14.5	-14.9	N.S.	-15.4	-15.8	-15.8	-0.4		-0.9	-0.5	-0.4	-1.3
sep-09	-15.3	-15.0	N.S.	-15.6	-16.1	-16.0	0.3		-0.3	-0.6	-0.5	-0.7
oct-09	-14.0	-13.9	N.S.	-15.1	-15.5	-15.4	0.1		-1.1	-1.2	-0.4	-1.4
nov-09	-5.2	-5.1	N.S.	-6.0	-6.6	-6.5	0.1		-0.8	-0.9	-0.6	-1.3
dec-09	-2.3	-2.3	N.S.	-2.9	-3.6	-3.4	0.0		-0.6	-0.6	-0.7	-1.1

Figure 2: Revision of growth rate on the same month of the previous year

2.3 Analysis of revisions

Descriptive analysis of revision

It can be also useful for producers of official statistics to better understand the characteristic of the statistical compilation process to identify eventually possible drawbacks and to make improvements studying the information in succeeding revisions. It is an important tool for economic forecasts. Moreover it gives the user the opportunity to analyse different types of revision intervals depending on the purpose of the study: revision between first and final estimates or the incremental effect of revisions between subsequent releases.

Using all vintages for an economic indicator it is possible to identify where any biases might exist, to study the pattern in the revisions that can be used to improve the forecasting processes and finally to provide measures of data quality.

To measure the average size of the revisions without providing an indication of directional bias, mean absolute revision (MAR) is calculated. The range that 90% of revisions lie within gives a normal range expected for the revision without being influenced by outliers.

To reveal whether revisions are systematic or not and to have an idea if the average level of revision is close to zero is useful to calculate the arithmetic average or means of revisions. When the mean is positive it indicates that on average earlier releases have been underestimated. Because revisions of opposite signs cancel out this measure is of limited use but calculating the percentages of positive, negative and zero revisions can be useful supplementary information. Besides, a modified t-statistics test (to take into account the serial correlation because revisions are not independent of each other) is used to see whether there is statistical evidence that the bias (mean revision) is significantly different from zero. In the case of not significance it is implied that the pattern of revisions may have occurred by chance.

The relative mean of absolute revisions (RMAR) can be calculated along the previous statistics when making comparisons in the size of revision across different indicators.

It is useful to have a measure of the variability of the revisions calculating the standard deviation of revision to give an indicator of the volatility of revisions for a given revision interval together with the minimum and maximum revision.

The mean squared revisions (MSR) and its decomposition (UM, UR and UD) displays possible systematic components in the revision process.

News or noise

Other approaches are available in order to analyse revisions. Among them an interesting tool, usually applied to analyse revisions to GDP, is the news or noise approach (Mankiw and Shapiro, 1986; Fixler, 2007), which provides an evidence about the way in which the available information is used.⁴ In fact revisions may add new available information (they contain news) or may arise because of measurement errors and inefficiencies in the preliminary estimates (they contain noise). In order to assess if revisions are news or noise, the correlation between $R_t = L_t - P_t$ and the estimates L_t and P_t are considered. If revisions R_t are significantly correlated with P_t (and uncorrelated with L_t), they contain noise and the preliminary estimates do not fully utilise the information available. On the contrary, if revisions R_t is significantly correlated with L_t (and uncorrelated with P_t), they contain news enabling the subsequent estimates to embody new information correctly.

Another technique, based on regressions, could be used to assess whether revisions embody news or noise (see Mankiw and Shapiro, 1986). As stressed in Fixler (2007), although it can be shown that the two techniques are related, there is a difference between them: in the computation of the correlation coefficient revisions and estimates are considered symmetrically, while in the regression equation the (preliminary or revised) estimate of an economic indicator represents the independent variable and the revision represents the dependent variable.

Analysis of revisions to detect the sources of large/biased revisions

Since revisions can be widely reported and criticised in the media threatening to undermine confidence in official statistics and in NSIs, most of their efforts to improve revision policy, to build real-time database and to analyse revisions are users-oriented. This contributes to explain why revision analysis is mainly restricted to key economic indicators (often in seasonally adjusted form). Revision analysis could be utilised as well as in different contexts and specifically can help detecting problems in the statistical estimation/compilation process. In fact, as the Statistics Commission stated in its report (Statistics Commission, 2004, p. 4) some revisions are not the consequence of additional information, but are potentially “avoidable”, as they are “due to errors or to weakness in the estimation procedures, or to tractable weakness in the underlying data systems”. In particular the report highlights four categories of avoidable circumstances that affect the revision process (p. 24):

- substantial mistakes in early processing;
- the models used to compute early estimates are not “best practice”;
- timetables could be more rapid than they actually are;
- the methods used are “best practice”, but they are implemented without sufficient resources.

⁴ Revisions due to changes in definitions, estimation methods, nomenclature, ... should not be considered in this analysis.

Revision analysis could be a useful tool both to detect such circumstances and to suggest suitable measures to counteract them. Two examples are provided in Hoven (2008) and in Ciammola et al. (2008): the former refers to the Dutch estimates of GDP volume growth, the latter to the Italian index of industrial production. Both propose a top-down approach (i.e., analysing first the highest level and then proceeding to the more detailed levels) to identify the specific area(s)/domain(s) “responsible” for large or biased revisions. Based on this approach, the following steps can be implemented⁵:

- 1) analysis of revisions on the highest level of aggregation (total aggregate);
- 2) if revisions are large or biased, computation of their contribution to the revision of the total aggregate;
- 3) if *one* or *few* areas/domains show a large contribution, analysis of revisions on these selected areas;
- 4) replication of steps 2 and 3 up to the most disaggregated components;
- 5) detection of the causes of large/biased revisions on early estimates (through a decomposition of revisions if the latter are generated by several sources).

It is worth noting that analysing revisions requires the analysis of a time series of homogeneous revisions.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Ciammola, A., Mancini, A. R., and Gambuti, T. (2008), Revision Analysis to Detect Possible Weakness in the Estimation Procedures. An application to the Italian IIP. *Proceedings of Q2008 - European Conference on Quality in Official Statistics*.

De Vries, W. (2002), Dimensions of Statistical Quality. Report presented at the “Inter-agency Meeting on Coordination of Statistical Activities” (UNSD), New York, 17-19 September 2002. <http://unstats.un.org/unsd/accsub/2002docs/sa-02-6add1.pdf> (August 2013)

⁵ The procedure here presented fits well the domain of short-term economic business indicators (industrial production, turnover, retail trade, ... derived through the aggregation of many components), while for GDP, due to its complex estimation process, many other factors have to be considered.

- Di Fonzo, T. (2005), The OECD project on revisions analysis: First elements for discussion. Paper presented at the OECD STESEG Meeting, Paris, 27-28 June 2005.
<http://www.oecd.org/dataoecd/55/17/35010765.pdf>
- Eurostat (2011), *European Statistics Code of Practice*.
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF
- Eurostat (2013), *ESS guidelines on revision policy for PEEIs*.
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-13-016/EN/KS-RA-13-016-EN.PDF
- Fixler, D. (2007), How to interpret whether revisions to economic variables reflect ‘News’ or ‘Noise’. Contribution to OECD / Eurostat taskforce on performing revisions analysis for sub-annual economic statistics.
- Hoven, L. (2008), Using results from revisions analysis to improve compilation methods: a case study on revisions of Dutch estimates of GDP volume growth. Contribution to OECD / Eurostat taskforce on performing revisions analysis for sub-annual economic statistics.
- IMF (2012), *Data Quality Assessment Framework – Generic Framework*.
http://dsbb.imf.org/images/pdfs/dqrs_Genframework.pdf
- Jenkinson, G. (2004), ONS Policy on Standards for Presenting Revisions Analysis in Time Series First Releases. *Economic Trends*, No. 604, March 2004.
<http://www.statistics.gov.uk/cci/article.asp?ID=793>
- Jenkinson, G. and George, E. (2005), Publication of Revisions Triangles on the National Statistics Website. *Economic Trends*, No. 614, January 2005.
<http://www.statistics.gov.uk/cci/article.asp?ID=1026>
- Mankiw, N. G. and Shapiro, M. D. (1986), News or Noise: An Analysis of GNP Revisions. *Survey of Current Business* **66**, 20–25.
- Mazzi, G. and Ruggeri-Cannata, R. (2008), A Framework for Revisions Policy of Key Economic Indicators. Contribution to OECD / Eurostat taskforce on performing revisions analysis for sub-annual economic statistics.
<http://www.oecd.org/dataoecd/44/39/40309491.pdf?contentId=40309492>
- McKenzie, R. and Gamba, M. (2008), Data and metadata requirements for building a real-time database to perform revisions analysis. Contribution to OECD / Eurostat taskforce on performing revisions analysis for sub-annual economic statistics.
- Statistics Commission (2004), *Revisions to economic statistics*. Technical Report 17, Statistics Commission.

Interconnections with other modules

8. Related themes described in other modules

1.

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1.

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Quality Aspects-T-Revisions of Economic Official Statistics

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	18-03-2013	first version	Anna Ciammola Roberto Iannaccone	Istat
0.2	10-06-2013	numeric example, glossary	Anna Ciammola Roberto Iannaccone	Istat
0.3	26-08-2013	several changes to embody comments and remarks by SE	Anna Ciammola Roberto Iannaccone	Istat
0.4	11-09-2013	changes to embody comments by SE	Anna Ciammola Roberto Iannaccone	Istat
0.4.1	30-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:27

Theme: Macro integration**0. General information****0.1 Module name**

Theme: Macro integration

0.2 Module type

Theme

0.3 Module code

Theme-Macro integration

0.4 Version history

Version	Date	Description of changes	Author	NSI	Person-id
1.0p1	31-3-2011	First version	Jacco Daalmans	CBS	JDAS

Template version used	1.0 d.d. 25-3-2011
Print date	12-8-2011 14:41

Contents

General description – Theme: Macro Integration	3
1. Summary	3
2. General description.....	3
3. Glossary.....	6
4. Literature	7
Specific description – Theme: Macro Integration	9
A.1 Relationship with other modules.....	9

General description – Theme: Macro Integration

1. Summary

Macro integration is defined as the process to integrate data from different sources on an aggregate level, to enable a coherent analysis of the data, and to increase the accuracy of estimates. It entails the correction of data for major measurement error and for the remaining noise. The latter is also called data reconciliation. In this module we focus on formal methods for data reconciliation.

2. General description

2.1.1 Aim

Macro integration has two objectives. The first is to facilitate analysis of the interrelationships in data by organizing economic information into an accounting framework. The second is to make more accurate estimates of economic reality through the reconciliation of the various statistical information contained in a framework of this kind. An economic accounting framework is defined by a set of variables and a set of relationships between them (accounting rules). The data generally come from a wide variety of sources. Some variables may be obtained from external sources, or through sampling, but, where no suitable source exists, variables may be based on model estimates, or ‘expert guesses’.

Usually, the data that are collected by statistical offices are not consistent with the accounting rules. This happens, for example, because economic data are frequently collected by different methods, using different sample surveys and different data processing methods and because of estimation error in case of missing data.

Macro integration first ‘translates’ the source data to comply with the correct definitions and then identifies and adjusts for major measurement errors. After the major errors are corrected (the so-called bias), the remaining, usually smaller, discrepancies have to be solved (the so-called noise). These small discrepancies appear more or less by accident, for instance due to sampling errors.

National Statistical Institutes (NSIs) have often applied informal methods for macro integration. These methods heavily depend on mutual agreement of different subject matter experts on the necessary adjustments to the data. Although these informal methods work well in practise there are also some drawbacks. Informal methods are not transparent and therefore irreproducible. Furthermore the process of achieving consistency is often time-consuming for large data sets.

As an alternative to informal methods, formal methods can be used. In the literature a lot of formal methods are described. Some statistical offices have adopted these formal methods. In this section formal macro integration methods are discussed, that are aimed at the second step of macro integration, i.e. for correcting the noise, the small discrepancies that cannot be attributed to measurement errors or other sources of bias. In practise it is hard to differentiate between bias and noise. It is usually necessary to resort to elimination by hand of the largest differences, and to distribute the mass of smaller discrepancies through modelling.

2.2.2 Related process steps

Data reconciliation can be extended with a temporal component, which involves using data from two time periods simultaneously. Data that have unequal frequencies, such as quarterly and annual, are allowed. Without loss of generality, the high-frequency period will be called the sub annual period, while the low-frequency period will be called the annual period.

Two process steps are mentioned frequently in the literature:

1. *Benchmarking* which is a process to achieve consistency between sub annual and annual data.
2. *Temporal disaggregation* Deriving sub annual data (for instance quarterly data) from annual data, possibly by using indicators of the sub annual data (i.e. related time series).

The difference is that benchmarking assumes the same definitions for the annual and the sub annual data, while these definitions may be different for temporal disaggregation. In the problem of benchmarking the sub annual data are already available, but for temporal disaggregation only indicators may (or may not) be available. Another difference is that for temporal disaggregation more than one indicator time series may be used for the disaggregation of one time-series, while in case of benchmarking always one sub annual time series is aligned to one sub annual time-series.

Although benchmarking and temporal disaggregation differ from a conceptual point of view, both process steps are closely related from a methodological point of view. The same methods can be used for both problems.

Temporal disaggregation can be divided into:

- 2a. *Temporal distribution* which deals with flow variables (variables that are measured over an interval of time)
- 2b. *Interpolation*, which is the estimation of missing values of a stock variable (variables that are measured at one point of time).

Closely related to temporal disaggregation is:

- A. *Extrapolation* an estimation method of generating values outside the temporal range.
- B. *Calenderization*: converting data to a different unit of time. The problem occurs for instance if respondents send their data at 4-weeks intervals (i.e. 13 times a year), while the statistical office needs to publish the data at a monthly basis. Calenderization is as a special case of temporal distribution.

2.2.3 Related Methods

Although many macro integration are described in the literature, in this handbook we choose to describe three (classes of) methods: RAS (M-RAS), the Stone Method (M-Stone) and Denton (M-Denton_for_benchmarking) and (M-Denton_for_temporal_disaggregation). The motivation of this choice is that these three methods are well-known, computationally easy, and suitable for large-scale application. References will be given to other macro integration methods.

Macro integration

The literature refers to various formal data reconciliation methods, each with its own origins. There is a correspondingly great variety in applicability, interpretability and generality.

The simplest methods were devised at a time before powerful computers were widely available. An example is the *RAS method* (M-RAS), a numerical method that allows the entries of a rectangular matrix to be aligned with a set of row and column totals. For this specific problem other methods may also be applied. For an overview we refer to Lahr and De Mesnard (2004) and Lenzen et al. (2009).

There are more general methods, with a better statistical foundation and a broader scope of applicability, that estimate reconciled results from source data while complying with certain constraints, in accordance with a specific procedure. Many of the common methods can be classified as *generalized least-squares methods*, which all have quadratic error terms in the objective function. Although approaches with a quadratic objective function are often used, other forms can be used as well. The best statistical estimate corresponds with the optimum value of some objective function. Different additional assumptions lead to specific model variants.

The assumption that there are linear equality constraints that should hold exactly (without an error term) leads to the method of *Stone* (M-Stone), which is one of the older (1942) and most rudimentary of the least-squares methods.

Benchmarking

For an overview of benchmarking methods in the literature we refer to Bloem et al. (2001). Here, we describe a well known method for benchmarking: the *Denton method* (M-Denton_for_benchmarking), which may also be applied to temporal disaggregation (M-Denton_for_temporal_disaggregation). It attempts to adjust the sub annual data, so that consistency is achieved with the annual data, while preserving as much as possible the trend of the sub annual data.

Originally, the Denton method is described for univariate data, Denton (1971). However, there are extensions in the literature, for instance for multivariate data.

Temporal disaggregation

An overview of temporal disaggregation methods is given by Chen (2007). A distinction can be made for methods that require the availability of indicator time-series on the sub annual level and methods for disaggregation that can be applied in absence of sub annual indicator series.

When sub annual data are not available smoothing methods can be used, for instance Cubic Splines, or the Boot, Feibes and Lisman smoothing method, see for instance Boot et al. (1967) and Wei and Stram (1990). For the case when indicator series on the sub annual are available, the Chow-Lin regression method (1971) and its variants are often used. As pointed out by Fernández (1981), the Denton method may also be used. The Denton method of Fernández is an extension of the more common Chow and Lin method. Where the Chow-Lin regression approach attempts to preserve as much as possible the initial values of the indicator series, the Denton method aims at the preservation of the changes of the initial indicators.

Calenderization

The topic of calenderization is described by Cholette and Chhab (1991) and Dagum and Cholette (2006). These references also describe the relationship between calenderization and temporal disaggregation.

Extrapolation

For a reference to this topic and the relationship with temporal disaggregation we refer to Dagum and Cholette (2006).

3. Glossary

Note 1: The definitions of the terms marked by an asterix (*) are taken from the *Statistical Data and Metadata Exchange* (SDMX).

Note 2: The term “Benchmarking” also appears in the SDMX, but we give a different definition, because of the specific context of the problem.

Term	Definition
Accuracy*	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.
Bias (of an estimator)*	An effect which deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average.
Benchmarking (hyper link to process step)	Achieving <u>consistency</u> between <u>data</u> that are published at different <u>frequencies</u> (for instance quarterly <u>data</u> that has to comply with annual <u>data</u>).
Calenderization	The problem of converting <u>data</u> to a different unit of time. (for instance to construct monthly data from data that are observed on 4-weeks intervals).
Coherence*	Adequacy of statistics to be combined in different ways and for various uses.
Consistency*	Logical and numerical <u>coherence</u> .
Constraint*	Specification of what may be contained in a <u>data</u> or metadata set in terms of the content or, for data only, in terms of the set of key combinations to which specific attributes (defined by the data structure) may be attached.
Data*	Characteristics or information, usually numerical, that are collected through observation.
Data Integration*	The process of combining <u>data</u> from two or more sources to produce statistical outputs.
Data Reconciliation*	The process of adjusting <u>data</u> derived from two different sources to remove, or at least reduce, the impact of differences identified.
Data Set*	Any organised collection of <u>data</u>
Disaggregation*	The breakdown of observations, usually within a common branch of a hierarchy, to a more detailed level to that at which detailed observations are taken.
Extrapolation	An estimation method of generating values outside the temporal scope
Flow variable	A flow variable is measured over an interval of time. (see also stock variable)
Frequency*	The time interval at which observations occur over a given time period.
Indicator*	A data element that represents statistical data for a specified time,

	place, and other characteristics, and is corrected for at least one dimension (usually size) to allow for meaningful comparisons.
Macrodata*	The result of a statistical transformation process in the form of aggregated information.
Macro integration (hyper link to theme)	Integrating <u>data</u> from different sources on an aggregate level, to enable a <u>coherent</u> analysis of the <u>data</u> , and to increase the <u>accuracy</u> of estimates.
Measurement error*	Error in reading, calculating or recording numerical value.
Microdata*	Non-aggregated observations, or measurements of characteristics of individual units.
Micro integration	A method that matches <u>data</u> on individual statistical units from different sources, to obtain a combined data file with better information. The quality of the <u>data</u> is measured in terms of validity, reliability and <u>consistency</u> .
Missing data*	Observations which were planned and are missing
Stock Variable	A stock variable is measured at one specific time, and represents a quantity existing at that point in time. See also flow variable
Temporal Disaggregation (hyper link to process step)	Deriving <u>sub annual data</u> (for instance quarterly data) from <u>annual data</u> , by using <u>indicators</u> of the <u>annual data</u> (i.e. related time series), see disaggregation
Time Series*	A set of ordered observations on a quantitative characteristic of an individual or collective phenomenon taken at different points of time.

4. Literature

- Bloem, A.M., Dippelsman, R.J., Maehle, O.N. (2001) Quarterly National Accounts Manual: Concepts, Data Sources, and Compilation, International Monetary Fund, Washington, DC.
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/CA-22-99-781/EN/CA-22-99-781-EN.PDF
- Boot J.C.G., W. Feibes and J.H.C. Lisman (1967), Further methods of derivation of quarterly figures from annual data, *Cahiers Economiques de Bruxelles*, 36: 539-546.
- Chen B. (2007), An Empirical Comparison of Methods for Temporal Distribution and Interpolation at the National Accounts, Bureau of Economic Analysis
- Chow G. and A.L. Lin (1971), Best linear unbiased interpolation, distribution and extrapolation of time series by related series, *The Review of Economics and Statistics*, 53, 372-375.
- Cholette P.A., N.B. Chhab (1991), Converting Aggregates of Weekly Data into Monthly Values, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40, 3, 411-422.
- Dagum E.B. and P.A. Cholette (2006), *Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series*, Springer, New York.
- Denton F.T. (1971), Adjustment of monthly or quarterly series to annual totals: An Approach based on quadratic minimization, *Journal of the American Statistical Association*, 66, 333, pp. 99-102.
- Fernández, R.B. (1981), Methodological note on the estimation of time series, *Review of Economic and Statistics*, 63, no.3, p. 471-478.

- Lahr M.L. and L. de Mesnard (2004), Biproportional Techniques in Input-Output Analysis: Table Updating and Structural Analysis, *Economic Systems Research*, Vol. 16, No.2, 115-134.
- Lenzen, M., B. Gallego, and R. Wood (2009), Matrix balancing under conflicting information. *Economic Systems Research* 21, 23–44.
- Wei W.W.S. and D.O. Stram (1990), Disaggregation of time series models, *Journal of the Royal Statistical Society*, 52, 453-467.

Specific description – Theme: Macro Integration**A.1 Relationship with other modules***A.1.1 Related themes described in other modules*

1. n/a

A.1.2 Methods explicitly referred to in this module

1. RAS (hyper link)
2. Stone (hyper link)
3. Denton method for benchmarking (hyper link)
4. Denton method for temporal disaggregation (hyper link)

A.1.3 Mathematical techniques explicitly referred to in this module

1. Quadratic optimization under linear constraints
2. Interpolation
3. Extrapolation

A.1.4 GSBPM phases explicitly referred to in this module

1. GSBPM Phase 6.2

A.1.5 Tools explicitly referred to in this module

1. n/a

A.1.6 Process steps explicitly referred to in this module

1. Data Reconciliation
2. Benchmarking
3. Temporal Disaggregation
4. Temporal distribution
5. Calenderization

Method: RAS

0. General information

0.1 Module name

Method: RAS

0.2 Module type

Method

0.3 Module code

Method-RAS

0.4 Version history

Version	Date	Description of changes	Author	NSI	Person-id
1.0p1	31-3-2011	First version	Jacco Daalmans	CBS	JDAS

Template version used	1.0 d.d. 25-3-2011
Print date	12-8-2011 15:04

Contents

General description – Method: RAS	3
1. Summary	3
2. General description.....	3
3. Examples – not tool specific	4
4. Examples – tool specific.....	5
5. Glossary.....	5
6. Literature	5
Specific description – Method: RAS.....	7
A.1 Purpose of the method.....	7
A.2 Recommended use of the method	7
A.3 Possible disadvantages of the method.....	7
A.4 Variants of the method.....	7
A.5 Input data sets	7
A.6 Logical preconditions.....	7
A.7 Tuning parameters	8
A.8 Recommended use of the individual variants of the method	8
A.9 Output data sets.....	8
A.10 Properties of the output data sets	8
A.11 Unit of processing	8
A.12 User interaction - not tool specific.....	8
A.13 Logging indicators	8
A.14 Quality indicators of the output data.....	9
A.15 Actual use of the method	9
A.16 Relationship with other modules.....	9

General description – Method: RAS

1. Summary

The RAS method is a well known method for data reconciliation. Its aim is to achieve consistency between the entries of some nonnegative matrix and pre-specified row- and column totals.

The method was devised in a time when powerful computers were not available. It is very easy to apply and to understand. However, it has a narrow scope of applicability. It can only be applied to nonnegative matrices. Mathematically, the method is an iterative scaling method.

2. General description

Below we give a non-technical description of the RAS method. For a more detailed explanation we refer to Chapter IX of United Nations (1993). The original paper is by Bacharach (1970), but due to its technical character, we do not advise this for readers who are unfamiliar with the RAS method.

Mathematically, RAS is basically an iterative scaling method whereby a non-negative matrix is adjusted until its column sums and row sums equal to some pre-specified totals. It multiplies each entry in one row or column by some factor, that is chosen in such a way that the sum of all entries in the row or column becomes equal to its target total. This operation is first applied to all rows of the matrix. As a consequence the matrix becomes consistent with all target row totals. Then, the columns are made consistent with their required totals. As a result consistency is achieved with the column totals, but the constraints on the row totals may be violated again. The rows and columns are adjusted in turn, until the algorithm converges to a matrix that is consistent with all required row and column totals.

The adjustment of the entries of the matrix always happens to be biproportional to the row and column totals, in order to preserve the structure of the matrix as much as possible. This means that all ratios between an entry of the inner part of a matrix and the corresponding row and column totals are kept as close as possible to their initial values.

In the literature several extensions of the RAS method are given, see for instance Lahr and De Mesnard (2004). Amongst others are the following methods:

- GRAS (Generalised RAS) allows for matrices in which some of the elements are predefined, in addition to the row and the column totals;
- Another GRAS method (Generalised RAS) allows for matrices with negative entries, Lenzen et al. (2007);
- TRAS (Three-stage RAS) extends RAS by including constraints on arbitrary subsets of matrix elements, see Cole (1992) and Gilchrist and St. Louis (1999);
- KRAS (Konfliktfreies RAS), by Lenzen et al. (2009), generalizes the GRAS methods for the case of multiple (conflicting) source data for the same matrix entry and for differences in reliability between different data sources.

3. Examples – not tool specific

3.1 Example: the RAS method

The entries of the matrix in Table 1 below are inconsistent with the associated row and column totals. The RAS method is applied in order to produce a consistent table.

Table 1. Initial matrix

2	4	12
2	4	6
9	9	18

The rows are modified first. The sum of the entries in the first row is six, while the row total is twice as large. The entries of the matrix must therefore be multiplied by two. The second row of the matrix entries is consistent with the row total and is therefore left unmodified. The result of the above is Table 2. The matrix entries of this table are consistent with the row totals, but not with the column totals.

Table 2. First intermediate result

4	8	12
2	4	6
9	9	18

The next step of the algorithm modifies the columns. The sum of the first column is six and the column total is nine. The entries in the first column must therefore be increased by a factor of 9/6, or 1.5. The sum of the entries in the second column is 12, while the column total is 9. The two entries of the matrix must therefore be multiplied by 9/12, or 0.75, which produces Table 3. This table is entirely consistent. The algorithm stops. The algorithm would continue with rows if the table were still inconsistent.

Table 3 Final result

6	6	12
3	3	6
9	9	18

This example is quite simple. Real-life examples of this method may involve much more iterations. Further, the final matrix may involve broken numbers, even if all initial values are integers. In that case a stopping criterion has to be applied, to stop the algorithm when the discrepancies are sufficiently small.

4. Examples – tool specific

5. Glossary

Note 1: The definitions of the terms marked by an asterisk (*) are taken from the *Statistical Data and Metadata Exchange* (SDMX).

Term	Definition
Accuracy*	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.
Coherence*	Adequacy of statistics to be combined in different ways and for various uses.
Consistency*	Logical and numerical <u>coherence</u> .
Constraint*	Specification of what may be contained in a data or metadata set in terms of the content or, for data only, in terms of the set of key combinations to which specific attributes (defined by the data structure) may be attached.
Data*	Characteristics or information, usually numerical, that are collected through observation.
Data Integration*	The process of combining <u>data</u> from two or more sources to produce statistical outputs.
Data Reconciliation*	The process of adjusting <u>data</u> derived from two different sources to remove, or at least reduce, the impact of differences identified.
Data Set*	Any organised collection of <u>data</u>
Iterative proportional fitting	See <u>multiplicative weighting</u> .
Macrodata*	The result of a statistical transformation process in the form of aggregated information.
Macro integration (hyper link to theme)	Integrating <u>data</u> from different sources on an aggregate level, to enable a <u>coherent</u> analysis of the <u>data</u> , and to increase the accuracy of estimates.
Microdata*	Non-aggregated observations, or measurements of characteristics of individual units.
Micro integration	A method that matches <u>data</u> on individual statistical units from different sources, to obtain a combined data file with better information. The quality of the data is measured in terms of validity, reliability and consistency.
Missing data*	Observations which were planned and are missing
Multiplicative weighting	A form of weighting for which the weights are obtained by multiplying relevant weight factors, determined in an iterative process. Multiplicative weighting is also referred to as raking or iterative proportional fitting.
Raking	See <u>multiplicative weighting</u> .
Reconciliation	See <u>Data reconciliation</u>
Source Data*	Characteristics and components of the raw statistical data used for compiling statistical aggregates.

6. Literature

Bacharach, M. (1970), *Biproportional matrices & input-output change*. Cambridge University Press, Cambridge.

Cole, S. (1992) A Note on a Lagrangian Derivation of a General Multi-Proportional Scaling Algorithm. *Regional Science and Urban Economics*, 22, 291–297.

- Gilchrist, D. & St. Louis, L. (1999), Completing input–output tables using partial information with an application to Canadian data, *Economic Systems Research*, 11, pp. 185–193.
- Günlük-Şenesen G. and J. M. Bates (1988), Some experiments with methods of adjusting unbalanced data matrices. *Journal of the Royal Statistical Society, Series A* 151, 473–490.
- Lahr M.L. and L. de Mesnard (2004), Biproportional Techniques in Input-Output Analysis: Table Updating and Structural Analysis, *Economic Systems Research*, Vol. 16, No.2, 115-134.
- Lenzen, M., R. Wood and B. Gallego (2007), Some Comments on the GRAS Method. *Economic Systems Research*, 19, 461–465.
- Lenzen, M., B. Gallego, and R. Wood (2009), Matrix balancing under conflicting information. *Economic Systems Research* 21, 23–44.
- United Nations (1993). *Handbook of National Accounting; Handbook of Input-Output Table Compilation and Analysis. Series F, No. 74.* New York: United Nations.
http://unstats.un.org/unsd/publication/SeriesF/SeriesF_74E.pdf

Specific description – Method: RAS

A.1 Purpose of the method

The method is used for Data reconciliation ([hyperlink](#)), which is a specific process step used in the context of Macro integration ([hyperlink](#)).

A.2 Recommended use of the method

1. This method is only interesting if it is possible to represent the data in one rectangular matrix, in which the entries of the matrix are subject to change, while the row and column totals have to be equal to pre-specified values.
2. The matrix does not have to be square: the number of rows may differ from the number of columns.
3. The method is useful if the row- and column totals are of a higher precision than the entries of the matrix.

A.3 Possible disadvantages of the method

1. There is no way of differentiating between the reliability of different variables in the entries of the matrix. It is therefore impossible to make sure that a given value in the entries of the matrix is to undergo minimal or no change. However, some of the extensions of the RAS method that are described in the literature do allow for differences in reliability.
2. There is likewise no facility to impose constraints other than the row and column totals. It is therefore impossible to prescribe that the sum of two entries in the matrix equals a third entry, even if the condition was met in the initial situation. However, some extended versions of the RAS method in the literature do allow for these constraints.

A.4 Variants of the method

1. n/a

A.5 Input data sets

1. Ds-input1 = the entries of a matrix (required);
2. Ds-input2 = the required row and column totals of Ds-input1 (required).

Remark: the input data may involve micro- or macrodata.

A.6 Logical preconditions

A.6.1 Missing values

1. In Ds-input1 individual missing data values are not allowed.
2. In Ds-input2 individual missing data values are not allowed.

A.6.2 Erroneous values

A.6.3 Other preconditions

1. The sums of the row and column totals in Ds-input2 are equal.
2. Rows and columns consisting entirely of zeros do not occur in combination with a nonzero row or column total;
3. All values in Ds-input1 and Ds-input2 are non-negative.

Remark: A more complete description of the preconditions is given in Bacharach (1970). In order to avoid technical details we do not state the preconditions mentioned in this article.

A.7 Tuning parameters

1. Threshold values (optional), which specifies the maximum tolerated violation between the sum of the entries in one row or column of a matrix and the required row- or column total.

A.8 Recommended use of the individual variants of the method

1. n/a

A.9 Output data sets

1. Ds-output1 = the entries of a matrix

A.10 Properties of the output data sets

1. Ds-output1 is consistent with the pre-specified row- and column sums (Ds-input2)
2. The amount of adjustments is biproportional to the row- and columntotals (Ds-input2)

A.11 Unit of processing

Processing full data sets.

A.12 User interaction - not tool specific

1. Before execution of the method, the tuning parameters and input datasets are specified.
2. During operation no user interaction is needed, but inspection of quality indicators and subsequent adjustment of tuning parameters and recurrent use is optional.
3. After use of the method the quality indicators and logging should be inspected.

A.13 Logging indicators

1. The running time of the application
2. The number of iterations
3. Characteristics of the input data, for instance problem size, and the largest discrepancies between the row and column totals of the initial matrix (Ds-input1) and the desired totals (Ds-input2).

A.14 Quality indicators of the output data

1. The most important quality indicator is *how* much the figures are adjusted. Relative or absolute differences may be explored. Because of the relationships between the various entries of the input matrix, the differences must be examined in their mutual context.

Remark 1: It is possible to explore how the ratios between matrix cells and the row and column totals change in the reconciliation process. The RAS method attempts to preserve these ratios as much as possible. If a ratio has to change in the reconciliation process nonetheless, it is advisable to review the suitability of RAS.

Remark 2: Special attention is needed for zeros, both before and after reconciliation. The RAS method may create zeros if a row or column total is zero, but cannot adjust existing zero entries of the matrix. In both cases it must be verified that the data set has the correct structure.

A.15 Actual use of the method

1. At Statistics Netherlands the RAS method is used for updating the Input and Output Tables of National Accounts. At the current time, t , the row and column totals are fixed, since they have to be consistent with another statistic: the supply and use tables. Figures about the entries of the matrix are available only up to and including $t - 1$. Updating involves modifying the entries of the matrix at $t - 1$ in such a way as to make them consistent with the row and column totals at time t , also preserving the structure at $t - 1$ as much as possible.

A.16 Relationship with other modules

A.16.1 Themes that refer explicitly to this module

1. Macro integration

A.16.2 Related methods described in other modules

1. Method of Stone ([hyper link](#))
2. Denton for temporal disaggregation ([hyper link](#))
3. Denton for benchmarking ([hyper link](#))

Remark 1. The RAS method is the easiest method to apply, small problems can even be solved without a computer. However its field of application is more narrow than for the other methods. It is restricted to updating a matrix in such a way that it conforms to predefined row and column totals. The Stone method is more general: it does not require input data that can be represented in one matrix and it allows for more general type of constraints than the alignment to pre-specified row- and columntotals.

Remark 2. For the specific problem of updating a matrix to known row- and columntotals a lot of alternative methods are described in the literature as well, for an overview see for instance Lahr and De Mesnard (2004) and Lenzen et al. (2009). We choose to describe the RAS-method, since this method is the most well-known and most rudimentary.

A.16.3 Mathematical techniques used by the method described in this module

1. Iterative Scaling (also called raking)

A.16.4 GSBPM phases where the method described in this module is used

1. GSBPM phase 6.2 “Validate Outputs” ([hyper link](#))

A.16.5 Tools that implement the method described in this module

A.16.6 The Process step performed by the method

Data reconciliation ([hyper link](#))

Method: Stone

0. General information

0.1 Module name

Method: Stone

0.2 Module type

Method

0.3 Module code

Method-Stone

0.4 Version history

Version	Date	Description of changes	Author	NSI	Person-id
1.0p1	31-3-2011	First version	Jacco Daalmans	CBS	JDAS

Template version used	1.0 d.d. 25-3-2011
Print date	12-8-2011 15:05

Contents

General description – Method: Stone	3
1. Summary	3
2. General description.....	3
2.1 Literature.....	3
2.2 Determining a covariance matrix.....	3
3. Examples – not tool specific	4
4. Examples – tool specific.....	6
5. Glossary.....	6
6. Literature	7
Specific description – Method: Stone.....	8
A.1 Purpose of the method.....	8
A.2 Recommended use of the method	8
A.3 Possible disadvantages of the method.....	8
A.4 Variants of the method.....	8
A.5 Input data sets	8
A.6 Logical preconditions.....	8
A.7 Tuning parameters	9
A.8 Recommended use of the individual variants of the method	9
A.9 Output data sets.....	9
A.10 Properties of the output data sets	9
A.11 Unit of processing	9
A.12 User interaction - not tool specific.....	9
A.13 Logging indicators	9
A.14 Quality indicators of the output data.....	10
A.15 Actual use of the method	10
A.16 Relationship with other modules.....	10

General description – Method: Stone

1. Summary

This Stone method is a method for data reconciliation. It adjusts data in order to satisfy a set of linear constraints. The adjustments made to the data are as small as necessary to remove all discrepancies. In adjusting the initial data the method of Stone uses information on the relative reliabilities of the initial data, in particular a covariance matrix. Data that are considered most reliable are modified least and vice versa. The Stone method yields a set of fully reconciled data, with minimum variance.

The method of Stone translates the reconciliation problem into a mathematical optimization problem. From a mathematical perspective, the method of Stone is a weighted quadratic optimization problem under linear conditions. The formulation of this problem is relatively easy to understand.

The solution of the model includes the reconciled data as well as its covariance matrix. Analytical expressions can be derived for both results.

2. General description

2.1 Literature

A detailed description of the Stone method is given in the original paper, Stone (1942). In view of the extremely technical nature of this article, readers who are unfamiliar with the method are referred to the appendix in Wroe et al. (1999). A mathematical derivation of the results is given in Sefton and Weale (1995).

The Stone method is widely researched in the literature. Several extensions are described. For instance for reconciliation problem with soft constraints, (hard and soft) nonlinear constraints (for instance ratios) and inequalities, see for instance Magnus et al. (2000).

2.2 Determining a covariance matrix

In practical applications of the method, a covariance matrix of the initial data is often unavailable. Therefore applications generally use estimates of relative variances. Relative variances have no intrinsic meaning, but the ratio of relative variances is an indicator of the reliability of figures relative to each other. There are several ad hoc methods for estimating relative variances. One method is to have a specialist estimate 95% confidence intervals and to use the interval sizes as an approximation for variances. Another method is to distinguish several categories, such as relatively unreliable, normally reliable and relatively reliable. All variables within the same group are assigned the same variance.

It is often desirable in practice for reconciliation to affect large values more than small values in an absolute sense. If this is what is intended, the following variances may be chosen:

$$Var(x_i) = \theta_i^2 x_i^2,$$

where θ_i is a parameter that depends on the reliability, or reliability category, of x_i .

Determining the correct ratios between the various variances is a process of trial and error in practice, which means that one particular ratio is chosen based on a degree of prior knowledge and simple assumptions (e.g. that variances are equal in the absence of prior knowledge), and then judging whether the results are acceptable. If not, the variances are modified.

In practice, in the absence of quantitative measures, all covariances are usually assumed to be zero, or, in other words, that the variables are assumed to be mutually independent.

3. Examples – not tool specific

This example is based on the greatly simplified supply and use tables, which belong to the national accounts, as shown in Table 1 and 2. The rows of Table 1 are related to the supply of products and services, and columns to the producing sectors. The first two rows of Table 2 show the demand for products and services, and the first two columns show the customer sectors. The grand total of the whole table is empty, since it was opted not to include it in the mathematical model. This grand total can be derived directly from the other totals.

There are only two sectors, industry and services, and two goods groups, industrial products and services. The economy depicted is moreover a closed one, since there is no trading with foreign countries. Other items ignored are taxes on products, subsidies, trade and transport margins, and all categories of final use other than consumption.

The constraints are that:

- total supply equals total use for industry and services (the column totals of Table 1 equal the first two column totals of Table 2);
- total supply equals total use for industrial products and services (row totals in Table 1 equal the first two row totals of Table 2).

In addition, the sums of the entries of Tables 1 and 2 must also equal its row and column totals.

Table 1. Supply

	Industry	Services	Total
Industrial products	700	300	1000
Services	100	400	500
Total	800	700	

Table 2. Use

	Industry	Services	Consumption	Total
Industrial products	50	190	860	1100
Services	170	100	180	450
Wages	450	350		800
Operating surplus	130	60		190
Total	800	700	1040	

Two constraints are not satisfied in the starting situations: total supply is unequal to total use for industrial products and services (the row totals of Table 1 are inconsistent with the first two row totals of Table 2). The variances are shown in Tables 3 and 4; they were chosen arbitrarily.

Table 3. Variances: supply table

	Industry	Services	Total
Industrial products	100	1000	1100
Services	1000	100	1100
Total	1100	1100	X

Table 4. Variances: use table

	Industry	Services	Consumption	Total
Industrial products	500	1000	1000	2500
Services	1000	1000	1000	3000
Wages	700	700		1400
Operating surplus	1200	1200		2400
Total	3400	3000	2000	X

Note that the row and column totals are not fixed, since their variance is greater than zero.

The figures are reconciled with the method of Stone. The reconciled values in Tables 5 and 6 satisfy all constraints. Small differences in the row sums in Table 6 are attributable only to rounding errors. The figures before reconciliation are shown in brackets.

Table 5. Table of reconciled supply values, rounded

	Industry Services				Total	
Industrial products	705	(700)	318	(300)	1023	(1000)
Services	92	(100)	396	(400)	488	(500)
Total	797	(800)	714	(700)	1511	(1500)

Table 6. Table of reconciled use values, rounded

	Industry Services				Consumption		Total	
Industrial products	33	(50)	164	(190)	827	(860)	1023	(1100)
Services	179	(170)	118	(100)	191	(180)	488	(450)
Wages	452	(450)	358	(350)			810	(800)
Operating surplus	133	(130)	74	(60)			207	(190)
Total	797	(800)	714	(700)	1017	(1040)	2527	(2540)

A covariance matrix is also derived for the reconciled figures. This covariance matrix is not diagonal, and there are also nonzero covariances. The variances are shown in Tables 7 and 8. The values are less than in the initial situation. The variances before reconciliation are shown in brackets.

Table 7. Variances for the table of reconciled supply values

	Industry	Services	Total
Industrial products	84 (100)	270 (1000)	280 (1100)
Services	277 (1000)	85 (100)	292 (1100)
Total	293 (1100)	289 (1100)	

Table 8. Variances for the table of reconciled use values

	Industry	Services	Consumption	Total
Industrial products	346 (500)	524 (1000)	463 (1000)	280 (2500)
Services	541 (1000)	523 (1000)	489 (1000)	292 (3000)
Wages	415 (700)	420 (700)		519 (1400)
Operating surplus	575 (1200)	591 (1200)		667 (2400)
Total	293 (3400)	289 (3000)	563 (2000)	

4. Examples – tool specific

5. Glossary

Note 1: The definitions of the terms marked by an asterisk (*) are taken from the *Statistical Data and Metadata Exchange* (SDMX).

Term	Definition
Accuracy*	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.
Coherence*	Adequacy of statistics to be combined in different ways and for various uses.
Consistency*	Logical and numerical <u>coherence</u> .
Constraint*	Specification of what may be contained in a <u>data</u> or metadata set in terms of the content or, for data only, in terms of the set of key combinations to which specific attributes (defined by the data structure) may be attached.
Covariance matrix	A mathematic measure of reliability.
Data*	Characteristics or information, usually numerical, that are collected through observation.
Data Integration*	The process of combining <u>data</u> from two or more sources to produce statistical outputs.
Data Reconciliation*	The process of adjusting <u>data</u> derived from two different sources to remove, or at least reduce, the impact of differences identified.
Data Set*	Any organised collection of <u>data</u>
Iterative proportional fitting	See <u>multiplicative weighting</u> .
Macrodata*	The result of a statistical transformation process in the form of aggregated information.

Macro integration (hyper link to theme)	Integrating <u>data</u> from different sources on an aggregate level, to enable a coherent analysis of the <u>data</u> , and to increase the <u>accuracy</u> of estimates.
Microdata*	Non-aggregated observations, or measurements of characteristics of individual units.
Micro integration	A method that matches <u>data</u> on individual statistical units from different sources, to obtain a combined data file with better information. The quality of the <u>data</u> is measured in terms of validity, reliability and consistency.
Missing data*	Observations which were planned and are missing
Multiplicative weighting	A form of weighting for which the weights are obtained by multiplying relevant weight factors, determined in an iterative process. Multiplicative weighting is also referred to as <u>raking</u> or <u>iterative proportional fitting</u> .
Raking	See <u>multiplicative weighting</u> .
Reconciliation	See <u>Data reconciliation</u>
Source Data*	Characteristics and components of the raw statistical <u>data</u> used for compiling statistical aggregates.

6. Literature

Magnus, J.R., J.W. van Tongeren, and A.F. de Vos (2000), *National Accounts Estimation using Indicator Ratios*, The Review of Income and Wealth **3**, 329-350,

Mantegazza S., F. Di Leo (2007), Integration of SUT/IOT into the National Accounts: The Italian experience. 16th International Conference on Input - Output Techniques, 2-6 July 2007 (Istanbul Turkey).

United Nations, Statistics Division (2000), *Handbook of National Accounting: Use of Macro Accounts in Policy Analysis*. Studies Methods, United Nations, New York.

Sefton, J. and M.R. Weale (1995), *Reconciliation of national income and expenditure: balanced estimates for the United Kingdom, 1920-95*. Cambridge University Press, Cambridge.

Stone, J.R.N., D.A. Champerowne and J.E. Maede (1942), *The Precision of National Income Accounting Estimates*. Reviews of Economic Studies **9**, 111-125.

Wroe D., P. Kenny, U. Rizki and I. Weerakoddy (1999), *Reliability and Quality Indicators for National Accounts Aggregates*. Office for National Statistics (ONS). Document CPNB 265-1 for the 33rd meeting of the GNP Committee, http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47143266/RELIABILITY%20AND%20QUALITY%20INDICATORS%20FOR%20NATIONAL%20ACCOUNTS%20AGGREGATES.PDF

Specific description – Method: Stone

A.1 Purpose of the method

The method is used for data reconciliation ([hyperlink](#)), which is a specific process step used in the context of Macro integration ([hyperlink](#)).

A.2 Recommended use of the method

1. The method may be applied to any problem in which consistency has to be achieved towards some set of equality constraints, which satisfies the preconditions in A.6.
2. Both positive and negative values are allowed. However, there is no way to constrain positive figures to remain positive and negative values to remain negative.
3. The method allows for exogenous variables, which are values that must remain unmodified.
4. The Stone method is relatively simple to program in R, Matlab and other packages. Quadratic programming (QP) solvers, such as CPLEX and XPRESS, can also be used.
5. The method may be applied to micro or macro data.
6. The method should be used to unbiased source figures. All source figures are consistent with their definitions. They are therefore already adjusted for systematic errors (nonresponse errors, measurement errors, processing errors and conceptual differences). Any errors in the input data (as mentioned in subsection A.5) will propagate to the results.

A.3 Possible disadvantages of the method

1. There is no way to constrain positive figures to remain positive and negative values to remain negative.

A.4 Variants of the method

1. n/a

A.5 Input data sets

1. Ds-input1 = a data set (micro data or macro data) (required)

A.6 Logical preconditions

A.6.1 Missing values

1. In Ds-input1 individual missing data values are not allowed.

A.6.2 Erroneous values

1. n/a

A.6.3 Other preconditions

1. The constraints (A.7-2) must not be mutually conflicting.

A.7 Tuning parameters

1. A covariance matrix (Required). The relative variance determine which of the variables are adjusted the most.

Remark 1: When all variables are equally reliable an identity matrix may be used.

Remark 2: This covariance is usually called the ex-ante covariante matrix. It differs from the ex-post covariance matrix, as mentioned in A.14-1.

2. Constraints (Required). These specify the constraints that should be satisfied.
 - a. Only equality constraints. Inequality constraints, such as 'revenue > 100*number of active employees' are not supported. Therefore non-negativity cannot be modelled.
 - b. Only linear constraints.

A.8 Recommended use of the individual variants of the method

1. n/a

A.9 Output data sets

1. Ds-output1 = a dataset with reconciled (micro or macro data) sub annual time-series.

A.10 Properties of the output data sets

1. The output data (Ds-output1) satisfy all constraints (A.7-2).
2. The ex-post variances (in A.14-1) can be at most as large as the corresponding ex-ante variances (in A.7-1).
3. The amount of adjustment is at least as possible.

A.11 Unit of processing

Processing full data sets.

A.12 User interaction - not tool specific

1. Before execution of the method, the tuning parameters and input datasets are specified.
2. During operation no user interaction is needed, but inspection of quality indicators and subsequent adjustment of tuning parameters and recurrent use is optional.
3. After use of the method the quality indicators and logging should be inspected

A.13 Logging indicators

1. The run time of the application
2. Characteristics of the input data, for instance problem size, the largest discrepancies of the input data towards the constraints.

A.14 Quality indicators of the output data

1. A covariance matrix corresponding to Ds-output1 (usually called the ex-post covariance matrix).
2. *How* the figures (in DS-input1) were adjusted. Attention may be focused on relative or absolute differences. Because of the relationships between the various variables in the system, the differences must be examined in their mutual context. A quantitative measure for the degree of inconsistency in the data before reconciliation is the value of a weighed sum of the squared reconciliation adjustments. A high value implies many large adjustments were needed to achieve consistency.
3. The accuracy. The method of Stone gives reconciled figures with minimum variance, assuming the variance of the figures to be reconciled are given. The diagonal entries of the ex-post covariance matrix (A.14-1) give information about the relative reliability of the reconciled results. Comparison with the ex-ante covariance matrix (A.7-1) yields information about how the reconciliation reduces the data variance. The nondiagonal entries of the ex post covariance matrix (A.14-1) yield information about inter-variable correlations introduced by reconciliation.

Remark 1: This process can become complicated with very large numbers of variables or internal relationships, in which case it may be simpler to analyse the differences before reconciliation (i.e. the constraint violations), as opposed to the reconciliation adjustments.

Remark 2: A need for many large reconciliation adjustments may indicate biased source data, meaning that the model conditions were not satisfied, and therefore that the method should not have been applied.

Remark 3: An important quality indicator of the method *implementation* in a tool is *whether* the figures are successfully reconciled. To this end, the remaining differences can be calculated on all linear constraints (A.7-2). Numerical error will generally cause these differences to deviate slightly from zero, which is not usually a problem as long as the differences are less than a certain threshold value.

A.15 Actual use of the method

1. The method of Stone adapted by National Statistical Offices in the compilation of National Accounts, for instance by ISTAT (Mantegazza and Di Leo, 2007).

A.16 Relationship with other modules

A.16.1 Themes that refer explicitly to this module

1. Macro integration (hyper link)

A.16.2 Related methods described in other modules

1. RAS method (hyper link)
2. Denton for benchmarking (hyper link)
3. Denton for temporal disaggregation (hyper link)

Remark 1: The Stone method is more general than the RAS method. The RAS method adjusts the entries of an matrix to achieve consistency with given row- and columntotals. The method of Stone however does not need the precondition that the data can be represented in a two-dimensional matrix. Furthermore, the method of Stone allows for different types of constraints than the alignment to row and column totals. And a third difference is that the method of Stone allows for differences of reliability of the source data, while the RAS method does not. However, from a technical point of view, the RAS method is easier to apply than the Stone method. The RAS method is an easy iterative scaling procedure, while the Stone method requires the computation of the solution of a least square problem.

Remark 2: In comparison with the Denton method, the Stone method is less specific. The Denton method is meant for achieving consistency between data of different frequencies (for instance quarterly data that has to comply with annual data), while the Stone method does not include a time component.

A.16.3 Mathematical techniques used by the method described in this module

1. Generalised Regression
2. Quadratic programming under linear constraints

A.16.4 GSBPM phases where the method described in this module is used

1. GSBPM phase 6.2 “Validate Outputs” ([hyper link](#))

A.16.5 Tools that implement the method described in this module

A.16.6 The Process step performed by the method

Data reconciliation

Method: Denton for benchmarking

0. General information

0.1 Module name

Method: Denton for benchmarking

0.2 Module type

Method

0.3 Module code

Method-Denton-for-benchmarking

0.4 Version history

Version	Date	Description of changes	Author	NSI	Person-id
1.0p1	31-3-2011	First version	Jacco Daalmans	CBS	JDAS

Template version used	1.0 d.d. 25-3-2011
Print date	12-8-2011 15:07

Contents

General description – Method: Denton for benchmarking.....	3
1. Summary	3
2. General description.....	3
3. Examples – not tool specific	5
4. Examples – tool specific.....	8
5. Glossary.....	8
6. Literature	9
Specific description – Method: Denton for benchmarking	11
A.1 Purpose of the method.....	11
A.2 Recommended use of the method	11
A.3 Possible disadvantages of the method.....	11
A.4 Variants of the method.....	11
A.5 Input data sets	12
A.6 Logical preconditions.....	12
A.7 Tuning parameters	13
A.8 Recommended use of the individual variants of the method	13
A.9 Output data sets.....	13
A.10 Properties of the output data sets	13
A.11 Unit of processing	14
A.12 User interaction - not tool specific.....	14
A.13 Logging indicators	14
A.14 Quality indicators of the output data.....	14
A.15 Actual use of the method	14
A.16 Relationship with other modules.....	14

General description – Method: Denton for benchmarking

1. Summary

The Denton method is a well known method for benchmarking and temporal disaggregation. Its aim is to achieve consistency between data that are published at different frequencies (for instance annual data with quarterly data). Following the literature, these periods will be called annual and sub annual periods, respectively. This terminology can be used without loss of generality, sub annual and annual periods can be any combination of two different periods with unequal lengths, such that one annual period covers a whole number of sub annual periods.

The method may be applied to time-series, consisting of at least one annual period. In achieving consistency, the sub annual data are adjusted, while the annual data are not changed (i.e. at least not in the method that is originally described by Denton in 1971). Furthermore, the Denton method attempts to preserve the trend of the high-frequency data as much as possible.

Originally, the Denton method is defined for univariate data. However, in the literature a lot of extensions are described, for instance for the multivariate case.

Mathematically, the Denton method translates a data reconciliation problem into a weighted quadratic optimization problem under linear conditions. As mentioned by Bloem et al. (2001) the Denton method is very well suited for large-scale applications.

2. General description

Below we give a non-technical description of the Denton method. For a more comprehensive and a more technical description we refer to Denton (1971). A reference for an extended version for multivariate data is Bikker et. al. (2010).

The Denton method is a macro integration method ([hyperlink: M-Macro_integration](#)), which is specially aimed at benchmarking. Therefore, the Denton method is more specific than other macro integration methods, like RAS ([hyperlink: M-Ras](#)) and Stone ([hyperlink: M-Stone](#)).

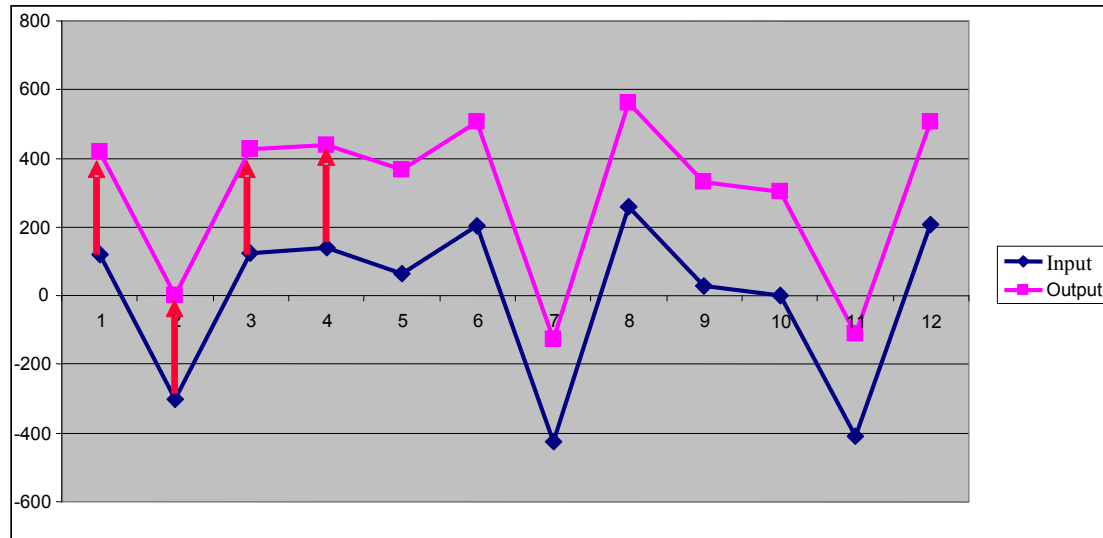
Although the Denton method is originally described for the problem of benchmarking, it can also be used for temporal disaggregation (Fernández, 1981). However in the field of temporal disaggregation other methods are mentioned more frequently in the literature, for instance the Chow–Lin regression method and its variants (Chen, 2007). In fact, the Denton method is an extension of the Chow-Lin method.

The aim of a Denton method is to achieve consistency between annual and sub annual data. For instance four quarterly figures that have to add up to an annual total (the so-called *annual alignment*)

In achieving consistency, the Denton method assumes that the annual data sources are of high precision and provide reliable information on the overall levels. On the other hand, the sub-annual data are less precise, but they provide the only information on the short-term movements. The Denton method combines the (assumed) strengths of both types of data.

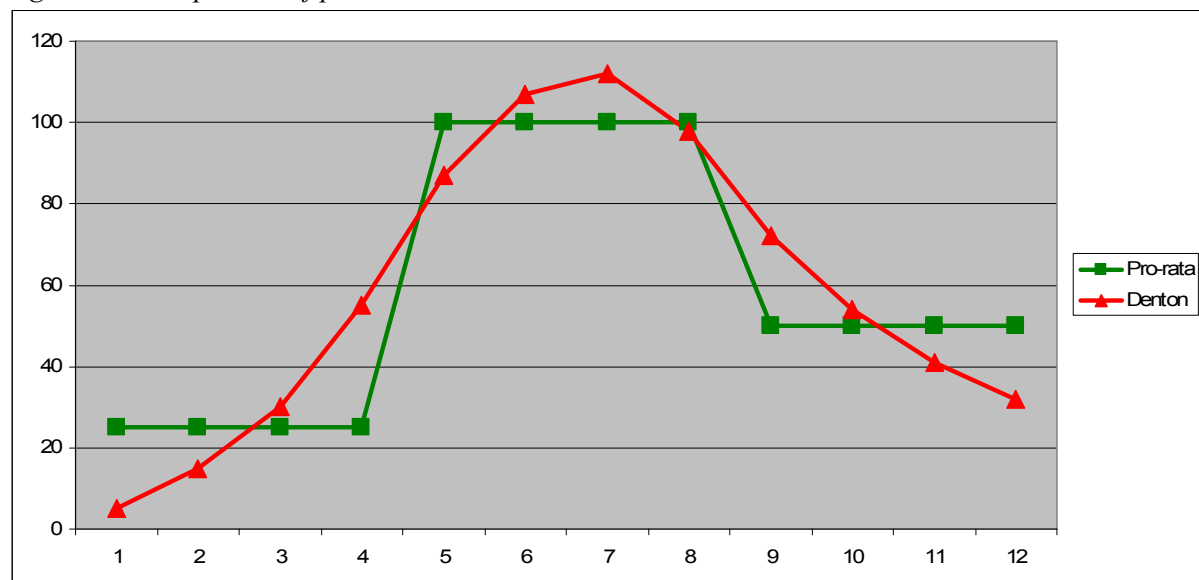
The Denton methods adjust the sub annual time-series (or indicators, in case of temporal disaggregation) to align with the annual time-series, while preserving as much as possible the trend of the sub annual data. The latter property is often called the *movement preservation principle*. Figure 1 illustrates this property.

Figure 1. The movement preservation principle



Due to this property the so-called *step problem* is prevented, which means that there are no large gaps between the last sub annual period of one year and the first sub annual period of the next year. Obviously, the most straightforward method of benchmarking is simply dividing an annual value by the number of sub annual periods in one annual period. However, this so-called *pro-rata method* heavily suffers from step-problems. For instance, Figure 2 compares the results of the Denton method with a pro-rata method, for a case of fictitious annual and quarterly data.

Figure 2. A comparison of pro-rata and Denton



There are several variants of the Denton method that differ in the way how the changes of the sub annual data are defined. For instance, a *proportional model* tries to preserve the procentual changes, while an *additive model* can be used to preserve the difference in absolute terms.

Initially, the Denton method was proposed for univariate data, Denton (1971). Several extensions are described in the literature, for instance:

- Di Fonzo and Marini (2003) have extended the Denton method for the multivariate case;
- Bikker and Buijtenhek (2006) have added reliability weights to a multivariate Denton method;
- Bikker et al. (2010) have added inequality constraints, soft constraints and ratio constraints to a multivariate Denton model.

In addition to the annual alignment, multivariate Denton methods also allow for *contemporaneous constraints*, i.e. constraints between different variables that should hold at one time period. For instance: for each time period total supply should be equal to total use.

In the extended multivariate Denton method, by Bikker et. al. (2010), weights can be used to differentiate the variable by the reliability of its source. As a consequence it can be established that variables that are considered highly reliable can be adjusted less than variables that are considered less reliable.

Another extension is for soft constraints, i.e. constraints that should approximately. These kind of constraints can be used to include subject matter knowledge in the model. For instance: the value of the stocks of some perishable good should be approximately equal to zero, since these kind of goods are usually not kept in stock. A nonzero value, is allowed, but the soft restriction prevents the occurrence of an undesirably high outcome.

Another possibility is to model the annual alignment as a soft constraint, for instance because some annual figure is considered not very reliable. As a consequence the annual data may be adjusted. A benchmarking problem with soft annual alignments is called *nonbinding benchmarking* by Dagum and Cholette (2006).

Weights can be attached to soft constraints that determine the relative importance of each constraint.

Furthermore ratio constraints can be included in the model. This is a usual feature since ratio types of interrelationships may play an important role in practical applications. For instance, in the national accounts, it is often assumed that the ratio between production and intermediate use is quite constant over time.

Finally, it is possible to include inequality constraints. A very commonly used type of inequality is the requirement that values cannot be negative.

3. Examples – not tool specific

3.1 Example: the univariate Denton method

To illustrate the univariate model, we use an artificial data set of twelve quarters and three annual totals. The quarterly figures were chosen to include pronounced changes that follow the seasons. They

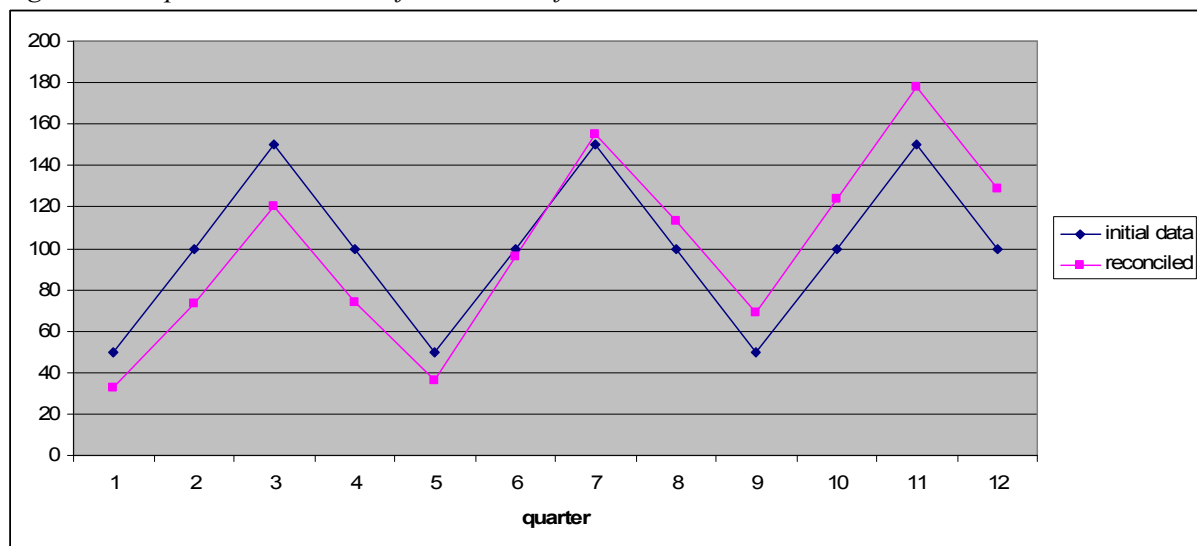
must then be modified in a way that the sums of the four quarters for each year exactly equal the corresponding annual totals. We imposed no other constraints.

Table 1. Input and Output of a fictitious example

	Input		Output
	Quarterly data	Annual totals	Reconciled data
Year 1	50	300	33
	100		73
	150		120
	100		74
Year 2	50	400	36
	100		96
	150		155
	100		113
Year 3	50	500	69
	100		124
	150		178
	100		129

We applied an univariate, proportional Denton method and rounded the results. The quarter-to-quarter changes were preserved as much as possible, see also figure 3. The reconciled data of the first year were lower than their initial values because of a lower annual total, and those of the third year were higher because of a higher annual total.

Figure 3. Graphical illustration of the results of table 1



3.2 Example: the multivariate Denton method

Let us consider a benchmarking problem, consisting of 12 quarters and two time series x_1 and x_2 . Suppose initially, each quarterly value is 10 and that annual benchmarks are available for both time series. These are 50, 75 and 95 for the three consecutive years and both time-series. Now assume that the first annual alignment is binding, whereas the second and the third are not.

This example is not very realistic, we intentionally choose for large discrepancies between the quarterly and annual data in order to illustrate more vividly how the Denton method works.

Furthermore, there is one soft ratio constraint, defined by

$$\hat{x}_{1t} / \hat{x}_{2t} \approx 1.1 \quad \text{for } t = 1, \dots, 12.$$

and the proportional model will be used for both time series. Note that the soft, ratio constraint is inconsistent with the annual figures of both time series (both time-series have the same annual values, which implies a target value of 1 for the ratio). The relative values of the weights of both model components determine their influence on the model outcome.

The results of the benchmarking method, depicted in figure 4, are two time series, whose values increase gradually over time. This increase is due to the connection to the annual benchmarks. Further note as a result of the ratio from the fifth quarter onwards \hat{x}_{1t} increases more rapidly than \hat{x}_{2t} . During the first four quarters, the influence of the ratio constraint is negligible, since the quarters of both time series have to strictly add up to the same annual values. In the second and third year the annual alignment is soft, and therefore the ratio constraint is relatively more important than for the first year.

Figure 4 The benchmarked time-series

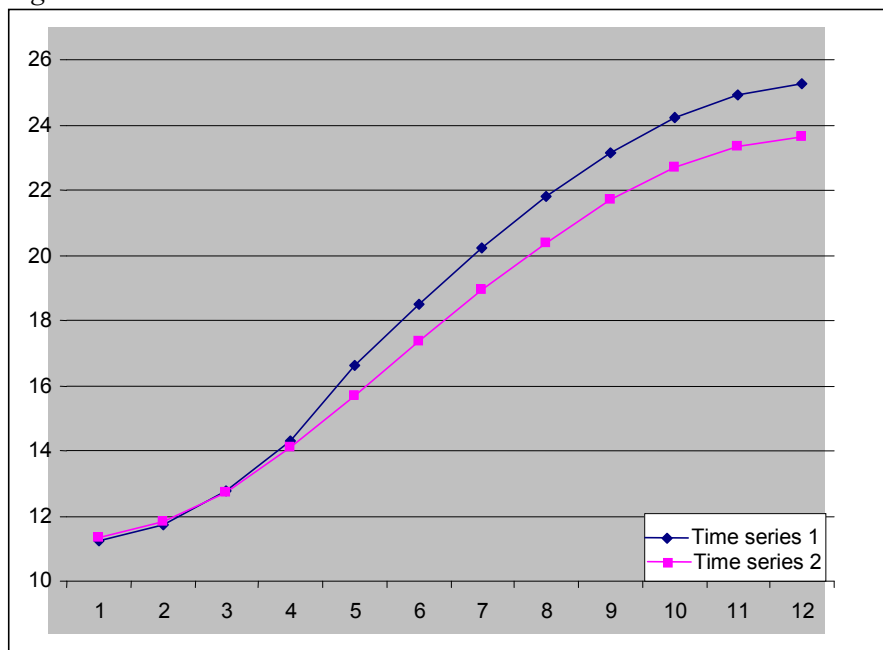


Table 2 shows that the reconciled annual figures (i.e. the sum of the underlying four quarterly values) of the second and third year closely approximate their target values.

Table 2. Reconciled annual figures, computed as the sum of four underlying quarterly values;

	Year 1	Year 2	Year 3
Time Series 1	50.00	77.16	97.61
Time Series 2	50.00	72.32	91.42
Time Series 1 / Time Series 2	1.000	1.067	1.068

4. Examples – tool specific

5. Glossary

Note 1: The definitions of the terms marked by an asterisk (*) are taken from the *Statistical Data and Metadata Exchange* (SDMX).

Note 2: The term “Benchmarking” also appears in the SDMX, but we give a different definition, because of the specific context of the problem.

Term	Definition
Accuracy*	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.
Annual Alignment	The constraint that annual data has to be consistent with sub annual <u>data</u> . Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Bias (of an estimator)*	An effect which deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average.
Benchmarking (hyper link to process step)	Achieving <u>consistency</u> between <u>data</u> that are published at different <u>frequencies</u> (for instance quarterly data that has to comply with annual data).
Binding constraint	See <u>hard constraint</u>
Coherence*	Adequacy of statistics to be combined in different ways and for various uses.
Consistency*	Logical and numerical <u>coherence</u> .
Constraint*	Specification of what may be contained in a <u>data</u> or metadata set in terms of the content or, for data only, in terms of the set of key combinations to which specific attributes (defined by the data structure) may be attached.
Contempeous constraints	Constraints within one period, between different <u>time-series</u>
Coverage error*	Error caused by a failure to cover adequately all components of the population being studied, which results in differences between the target population and the sampling frame.
Data*	Characteristics or information, usually numerical, that are collected through observation.
Data Integration*	The process of combining <u>data</u> from two or more sources to produce statistical outputs.
Data Reconciliation*	The process of adjusting <u>data</u> derived from two different sources to remove, or at least reduce, the impact of differences identified.
Data Set*	Any organised collection of <u>data</u>
Disaggregation*	The breakdown of observations, usually within a common branch of a hierarchy, to a more detailed level to that at which detailed observations are taken.
Frequency*	The time interval at which observations occur over a given time period.
Hard Constraint	A <u>constraint</u> that should hold exactly
Indicator*	A data element that represents statistical data for a specified time, place, and other characteristics, and is corrected for at least one dimension (usually size) to allow for meaningful comparisons.
Lagrange multiplier technique	In mathematical optimization, the technique of Lagrange multipliers (named after Joseph Louis Lagrange) provides a strategy for finding the maxima and minima of a function subject to <u>constraints</u> .
Macrodata*	The result of a statistical transformation process in the form of

	aggregated information.
Macro integration (hyper link to theme)	Integrating <u>data</u> from different sources on an aggregate level, to enable a <u>coherent</u> analysis of the data, and to increase the <u>accuracy</u> of estimates.
Measurement error*	Error in reading, calculating or recording numerical value.
Microdata*	Non-aggregated observations, or measurements of characteristics of individual units.
Micro integration	A method that matches data on individual statistical units from different sources, to obtain a combined data file with better information. The quality of the data is measured in terms of validity, reliability and consistency.
Missing data*	Observations which were planned and are missing
Movement preservation principle	The property that <i>all</i> changes of the <u>sub annual data</u> are kept as much as possible at their initial values.
Nonbinding benchmarking	A <u>benchmarking</u> problem with at least one <u>nonbinding annual alignment</u> constraint.
Nonbinding Constraint	See <u>soft constraint</u>
Pro-rata method	A simple <u>benchmarking</u> method in which the reconciled values are computed by dividing the <u>annual data</u> by the number of <u>sub annual</u> periods in one <u>annual</u> period. Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Soft constraint	A <u>constraint</u> that does not have to hold exactly, but approximately.
Step problem	The phenomenon of a large gap between the last sub annual period of one annual period and the first sub annual period of the next annual period. (for instance: a large gap between December and Januar). Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Temporal constraint	Constraints in the same <u>time-series</u> for different periods
Temporal Disaggregation (hyper link to process step)	Deriving <u>sub annual data</u> (for instance quarterly data) from <u>annual data</u> , by using <u>indicators</u> of the <u>sub annual data</u> (i.e. related time series), see <u>disaggregation</u> . Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Time Series*	A set of ordered observations on a quantitative characteristic of an individual or collective phenomenon taken at different points of time.
Weight*	The importance of an object in relation to a set of objects to which it belongs.

6. Literature

Barcellan R and D. Buono (2002), 'ECOTRIM interface user manual', Eurostat

Bikker, R.P. and S. Buijtenhek (2006), Alignment of Quarterly Sector Accounts to Annual Data, Statistics Netherlands, Voorburg, http://www.cbs.nl/NR/rdonlyres/D918B487-45C7-4C3C-ACD0-0E1C86E6CAFA/0/Benchmarking_QSA.pdf

Bikker R.P., J.A. Daalmans and N. Mushkudiani (2010), A multivariate Denton method for benchmarking large data sets, Discussion Paper 10002, Statistics Netherlands. <http://www.cbs.nl/NR/rdonlyres/7B2387F2-5773-42CF-8C50-5F02B451A2E4/0/201002x10pub.pdf>

- Bloem, A.M., Dippelsman, R.J., Maehle, O.N. (2001) Quarterly National Accounts Manual: Concepts, Data Sources, and Compilation, International Monetary Fund, Washington, DC.
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/CA-22-99-781/EN/CA-22-99-781-EN.PDF
- Boonstra H.J., C.J. De Blois, and G.J. Linders (2010), Macro integration with inequality constraints an application to the integration of transport and trade statistics, Statistics Netherlands,
- Boot J.C.G., W. Feibes and J.H.C. Lisman (1967), Further methods of derivation of quarterly figures from annual data, *Cahiers Economiques de Bruxelles*, 36: 539-546.
- Chen B. (2007), An Empirical Comparison of Methods for Temporal Distribution and Interpolation at the National Accounts, Bureau of Economic Analysis
- Dagum E.B. and P.A. Cholette (2006), *Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series*, Springer, New York.
- Denton F.T. (1971), Adjustment of monthly or quarterly series to annual totals: An Approach based on quadratic minimization, *Journal of the American Statistical Association*, 66, 333, pp. 99-102.
- Di Fonzo, T. (1990), The Estimation of M Disaggregate Time Series when Contemporaneous and Temporal Aggregates are Known, *The review of Economics and Statistics*, 72, pp. 178-182.
- Di Fonzo, T. and M. Marini (2003), *Benchmarking systems of seasonally adjusted time series according to Denton's movement preservation principle*. University of Padova, Working Paper 2003.09 <http://www.oecd.org/dataoecd/59/19/21778574.pdf>.
- Fernández, R.B. (1981), Methodological note on the estimation of time series, *Review of Economic and Statistics*, 63, no.3, p. 471-478.
- Magnus, J.R., J.W. van Tongeren and A.F. de Vos, (2000), National Accounts Estimation Using Indicators Analysis, *Review of Income and Wealth* (46), pp. 329-350.
- Magnus, J.R. and D. Danilov, (2008), On the estimation of a large sparse Bayesian system: The Snaer program, *Computational Statistics & Data Analysis*, 52, pp. 4203-4224.
- Wei W.W.S. and D.O. Stram (1990), Disaggregation of time series models, *Journal of the Royal Statistical Society*, 52, 453-467.

Specific description – Method: Denton for benchmarking

A.1 Purpose of the method

The method is used for Benchmarking ([hyperlink](#)), which is a specific process step used in the context of Macro integration ([hyperlink](#)).

A.2 Recommended use of the method

1. The method can be applied to any problem in which consistency has to be achieved between time-series that are published at different frequencies, assuming that the preconditions of Section A.6 are satisfied.
2. The method may be applied on micro or macro data.
3. The method is suitable when the sub-annual series are less reliable than the annual time-series. The Denton method will revise the sub-annual data and thus, a willingness to revise is necessary.
4. It is especially useful for practical applications in which it is important to preserve the trend of the sub annual series. For instance: the reconciliation of national accounts data.
5. As mentioned by Bloem et al. (2001) the method is very well suited for large-scale applications.
6. The method should be applied to unbiased source figures. All source figures are consistent with their definitions. They are therefore already adjusted for systematic errors (nonresponse errors, measurement errors, processing errors and conceptual differences). Any errors in the input data (as mentioned in subsection A.5) will propagate to the results.
7. The method may be used in a context of seasonal adjustment, when there are discrepancies between the yearly sums of the raw and the corresponding yearly sums of the seasonally adjusted series. The seasonally adjusted series may be benchmarked to the annual totals derived from the raw series.

A.3 Possible disadvantages of the method

1. Do not use Denton-for-benchmarking when the annual values are less reliable than the annual sums from the sub-annual series. In this case, using Denton-for-benchmarking will essentially diminish the reliability of sub annual time-series.

A.4 Variants of the method

1. A Denton method preserves as much as possible the period-to-period changes of the initial sub- annual data. Variants of the Denton method that differ in the way how these changes are defined:
 - 1.1. The additive model attempts to keep the additive corrections as constant as possible over all periods.

- 1.2. The proportional model is designed to preserve the percentage differences as constant as possible over all periods.
- 1.3. The additive and proportional model may be combined when the Denton method is applied to a multivariate data sets, i.e. the additive method may be applied to some time-series, and the proportional model to the other time-series.
2. The Denton method can minimize
 - 2.1 First order differences or
 - 2.2 Second order differences (differences of differences).

Remark: in the applications that are mentioned in the literature always first-order differences are used.

A.5 Input data sets

1. Ds-input1 = a data set (micro data or macro data) with sub annual time-series of a quantitative variable (Required).
2. Ds-input2 = a data set (micro data or macro data) with annual totals of the same quantitative variable (Required).

Remark: each time-series may comprise one or several annual periods.

A.6 Logical preconditions

A.6.1 Missing values

1. In Ds-input1 individual missing data values are not allowed.

Remark: In case of missing source data, one may use a dummy time-series, consisting of the value '1' (or any other value) for all sub-annual periods.

2. In Ds-input2 individual missing data values are allowed (meaning that there is no annual alignment).

A.6.2 Erroneous values

1. All non-empty values can be processed if of the right data type, that is the input must be quantitative values.

A.6.3 Other preconditions

1. One annual period covers a whole number of sub annual periods
2. The annual and sub annual data describe the same variables.
3. The constraints (Subsection A.7-2) must not be mutually conflicting. In particular this means that: for a multivariate benchmark problem, in which :
 - all annual alignments (Subsection A.7-2-a) are binding,
 - contemporaneous constraints (Subsection A.7-2-b) are defined, that have to be exactly satisfied;

the annual values, that are input to the model (Subsection A.5), also have to satisfy all the contemporaneous constraints.

4. The proportional variants of the Denton method (Subsection A.4) can only be applied if the initial sub annual data (Subsection A.5-1) do not contain any zeros.

A.7 Tuning parameters

1. Weights (Optional). These weights determine which of the time-series are adjusted the most (in a multivariate model) and which of the soft constraints are the most stringent. Weights may be omitted; in that case all data and constraints are considered equally reliable.
2. Constraints. These specify the constraints that should be satisfied. The following type of constraints can be used:
 - a. Annual alignment: a sum of sub annual values that should add up to an annual value (Fixed, i.e. this type of constraint is typical for the Denton method).
 - b. Contemporaneous constraints (for the multivariate case only): constraints between different time-series, within the same period (Optional).

Technically, a distinction can be made between

- i. soft constraint en hard constraints;
- ii. inequality constraints and equality constraints;
- iii. linear and nonlinear constraints.

A.8 Recommended use of the individual variants of the method

1. The proportional model (A.4 variant 1.2) is often preferred in application to economic data. The reason for this is that this model better preserves the seasonal changes, as these changes are often measured as a percentage change. However an additive model (A.4 variant 1.1) should be used:
 - a) For time-series that include zeros, for the proportional model is not defined properly for such time-series. The model attempts to preserve the initial percentage changes. A percentage change, however is not defined when the value of the first period is zero.
 - b) For time-series that involve both positive and negative values. A proportional model may yield undesirable results, for instance that each sub annual value is multiplied by some negative constant, meaning that all signs change.

A.9 Output data sets

1. Ds-output1 = a dataset with reconciled (micro or macro data) sub annual time-series.

A.10 Properties of the output data sets

1. The reconciled time-series (Subsection A.9-1) satisfy all restrictions (Subsection A.7-2).
2. The seasonal pattern is as close as possible to the seasonal pattern of the initial sub annual data (Subsection A.5-1)

A.11 Unit of processing

Processing full data sets

A.12 User interaction - not tool specific

1. Before execution of the method, the tuning parameters and input datasets are specified.
2. During operation no user interaction is needed, but inspection of quality indicators and subsequent adjustment of tuning parameters and recurrent use is optional.
3. After use of the method the quality indicators and logging should be inspected.

A.13 Logging indicators

1. The run time of the application
2. Characteristics of the input data, for instance problem size, the largest discrepancies of the input data towards the constraints.

A.14 Quality indicators of the output data

1. A quality indicator is *how* the sub annual time-series are modified. Of interest are the changes made in the first differences, given the starting point of the Denton method. The size of these changes is important, but the trend of the changes in time is particularly important. This trend can usually be assessed fastest by graphical means.

Remark 1: The Denton method adjusts the sub annual time-series in such a way that the adjustments are as smooth as possible over time. If the model still needs to largely adjust the initial changes, this may be because the preconditions of the model are not satisfied and therefore that the method should not have been applied. In other words: the ratio between the reconciled and the raw data should be stable over time.

Remark 2: A quality indicator of an implementation in a tool is how accurately the sub annual series is aligned with the annual series. Numerical error will generally cause these differences to deviate slightly from zero. The differences are not usually a problem as long as they are less than a certain threshold value.

A.15 Actual use of the method

1. Statistics Netherlands applies the method for reconciliation of quarterly and annual Supply and Use tables, see Bikker et. al. (2010).

A.16 Relationship with other modules

A.16.1 Themes that refer explicitly to this module

1. Macro integration (hyper link)

A.16.2 Related methods described in other modules

1. RAS method (hyper link)
2. Method of Stone (hyper link)

3. Denton for temporal disaggregation (hyper link)

Remark: the method of Stone and RAS are also data reconciliation methods, but these are not specially aimed at benchmarking and temporal disaggregation. Many other methods for benchmarking and temporal disaggregation are given in the literature. For an overview we refer to Bloem et al. (2001).

A.16.3 Mathematical techniques used by the method described in this module

1. Quadratic optimization under linear constraints (hyper link).

A.16.4 GSBPM phases where the method described in this module is used

1. GSBPM phase 6.2 “Validate Outputs” (hyper link)

A.16.5 Tools that implement the method described in this module

1. ECOTRIM

Remark: Freely available from: <http://circa.europa.eu/Public/irc/dsis/ecotrim/library> However, ECOTRIM is not designed for dealing with thousands of time series, it does not include features like weights, ratio's, soft constraints, and the possibility to combine the proportional and additive methods of benchmarking into one model.

2. De Kwartaalmachine (dutch)

Remark: This software has been developed by Statistics Netherlands and is currently used in the reconciliation of their national accounts. This software is designed for large-scale applications of the multivariate Denton method of Bikker et al. (2010), over 200,000 free variables. The software is not freely available. It makes use of XPRESS, a commercial solver for quadratic programming problems. A license of XPRESS is required to use the Kwartaalmachine.

A.16.6 The Process step performed by the method

Benchmarking (hyper link)

Method: Denton for temporal disaggregation

0. General information

0.1 Module name

Method: Denton for temporal disaggregation

0.2 Module type

Method

0.3 Module code

Method-Denton-for-temporal-disaggregation

0.4 Version history

Version	Date	Description of changes	Author	NSI	Person-id
1.0p1	31-3-2011	First version	Jacco Daalmans	CBS	JDAS

Template version used	1.0 d.d. 25-3-2011
Print date	12-8-2011 14:43

Contents

General description – Method: Denton for temporal disaggregation	3
1. Summary	3
2. General description.....	4
3. Examples – not tool specific	6
4. Examples – tool specific.....	8
5. Glossary.....	8
6. Literature	10
Specific description – Method: Denton for temporal disaggregation.....	12
A.1 Purpose of the method.....	12
A.2 Recommended use of the method	12
A.3 Possible disadvantages of the method.....	12
A.4 Variants of the method.....	12
A.5 Input data sets	13
A.6 Logical preconditions.....	13
A.7 Tuning parameters	14
A.8 Recommended use of the individual variants of the method	14
A.9 Output data sets.....	14
A.10 Properties of the output data sets	14
A.11 Unit of processing	15
A.12 User interaction - not tool specific.....	15
A.13 Logging indicators	15
A.14 Quality indicators of the output data.....	15
A.15 Actual use of the method	15
A.16 Relationship with other modules.....	15

General description – Method: Denton for temporal disaggregation

1. Summary

The Denton method is a well known method for benchmarking, but it can also be used for temporal disaggregation. In that case its aim is to achieve consistency between data that are published at different frequencies (for instance annual data with quarterly data or quarterly indicators). Following the literature, these periods will be called annual and sub annual periods, respectively. This terminology can be used without loss of generality, sub annual and annual periods can be any combination of two different periods with unequal lengths, such that one annual period covers a whole number of sub annual periods.

As mentioned by (Fernández, 1981) the Denton method can also be used for temporal disaggregation, i.e. the problem to reconcile annual figures with sub annual indicator time-series. However in the field of temporal disaggregation other methods are mentioned more frequently in the literature, for instance the Chow–Lin regression method and its variants (Chen, 2007). In fact, the Denton method of Fernández is an extension of the Chow-Lin method.

The problem of temporal disaggregation is to make annual values consistent with sub annual indicators. For each time-series of annual data, one or more sub-annual indicators may be used. Temporal disaggregation may also be applied to some annual time-series for which no sub annual indicators are available. However, a Denton method cannot be applied to that case: it requires at least one sub annual indicator time-series for each time-series to be estimated¹.

In achieving consistency, the sub annual indicators are adjusted, while the annual data are not changed (i.e. at least not in the method that is originally described by Denton in 1971). Furthermore, the Denton method attempts to preserve the trend of the indicator time-series as much as possible. At this point the Denton differs from the Chow-and-Lin method. The Chow-and-Lin method is aimed at the preservation of the levels of the sub annual indicators, not at the differences of these data.

Originally, the Denton method is defined for univariate time-series, that include one or more annual periods. However, in the literature a lot of extensions are described, for instance for the multivariate case. Mathematically, the Denton method translates a data reconciliation problem into a weighted quadratic optimization problem under linear conditions. The relationship between the indicator time-series and the time-series to be estimated is modelled by a linear regression relationship.

As mentioned by Bloem et al. (2001) the Denton method is very well suitable for large-scale applications.

¹ When sub annual indicators are not available, a Denton method may still be applied by using a constant indicator time-series.

2. General description

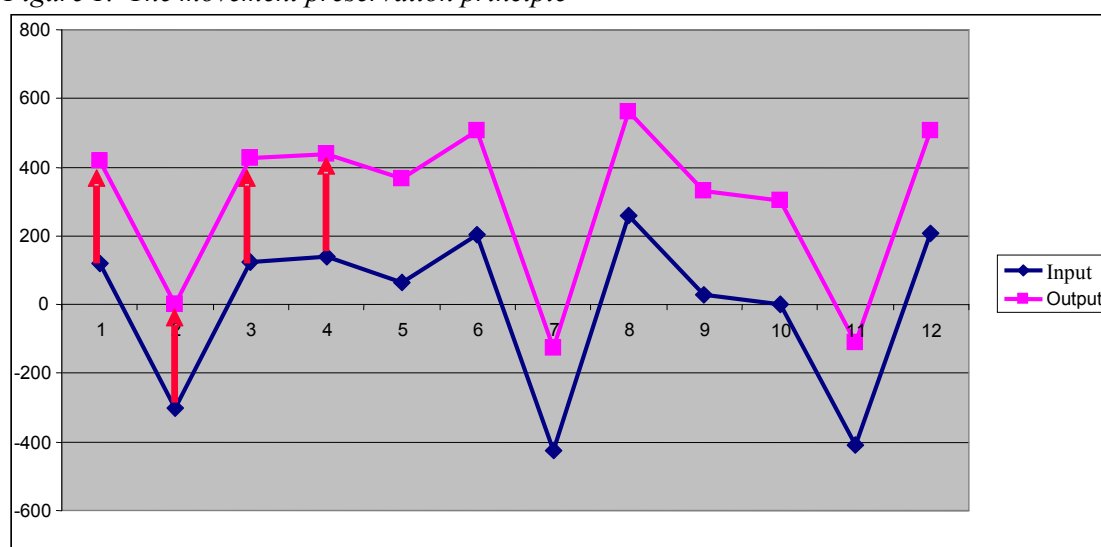
Below we give a non-technical description of the Denton method. For a more comprehensive and a more technical description we refer to Denton (1971). The reference for an extended version for multivariate data is Bikker et. al. (2010).

The Denton method is a macro integration method (hyperlink: M-Macro_integration), which is specially aimed at benchmarking and temporal disaggregation. Therefore, the Denton method is more specific than other macro integration methods, like RAS (hyperlink: M-Ras) and Stone (hyperlink: M-Stone).

In achieving consistency, the Denton method assumes that the annual data sources are of high precision and provide reliable information on the overall levels. On the other hand, the sub-annual indicators are less precise, but they provide the only information on the short-term movements. The Denton method combines the (assumed) strengths of both types of data.

The Denton methods adjust the sub annual indicator time-series to align with the annual time-series, while preserving as much as possible the trend of the sub annual indicator series. The latter property is often called the *movement preservation principle*. Figure 1 illustrates this property.

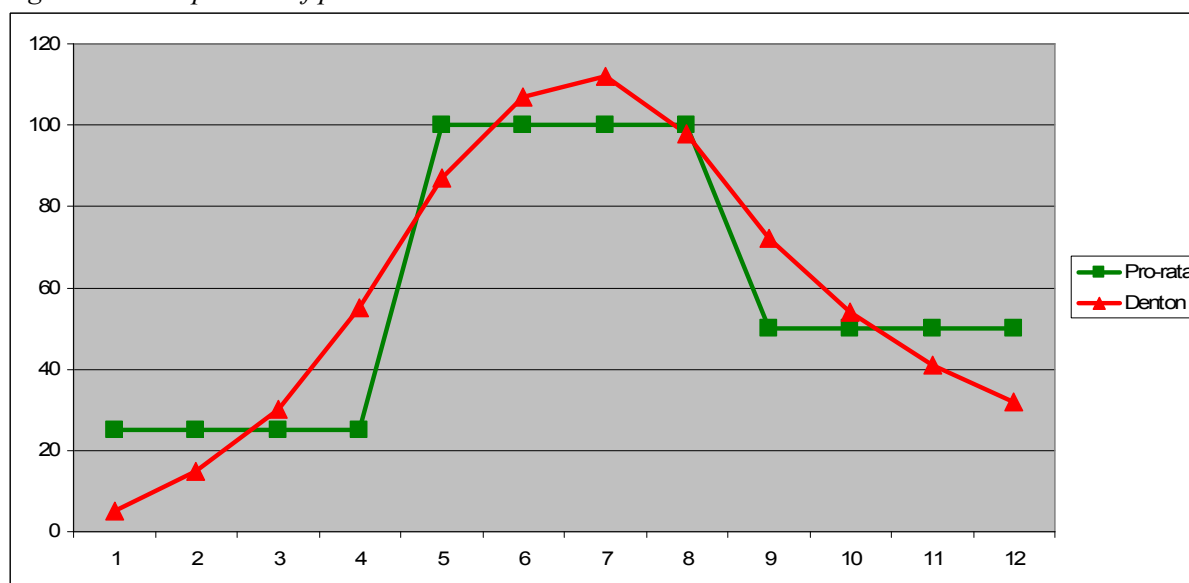
Figure 1. The movement preservation principle



Due to this property so-called *step problem* is prevented, which means that there are no large gaps between the last sub annual period of one year and the first sub annual period of the next year. Obviously, the most straightforward method of temporal disaggregation is simply dividing an annual value by the number of sub annual periods in one annual period, neglecting the sub annual indicators, ignoring the sub annual indicator series.

However, this so-called *pro-rata method* heavily suffers from step-problems. For instance, Figure 2 compares the results of the Denton method with a pro-rata method, for a case of fictitious annual data and quarterly indicators.

Figure 2. A comparison of pro-rata and Denton



There are several variants of the Denton method that differ in the way how the changes of the sub annual data are defined. For instance, a *proportional model* tries to preserve the procentual changes, while an *additive model* can be used to preserve the difference in absolute terms.

Initially, the Denton method was proposed for univariate data, Denton (1971). Several extensions are described in the literature, for instance:

- Di Fonzo and Marini (2003) have extended the Denton method for the multivariate case;
- Bikker and Buijtenhek (2006) have added reliability weights to a multivariate Denton method;
- Bikker et al. (2010) have added inequality constraints, soft constraints and ratio constraints to a multivariate Denton model.

In addition to the annual alignment, multivariate Denton methods also allow for *contemporaneous constraints*, i.e. constraints between different variables that should hold at one time period. For instance: for each time period total supply should be equal to total use.

In the extended multivariate Denton method, by Bikker et. al. (2010), weights can be used to differentiate the variable by the reliability of its source. As a consequence it can be established that variables that are considered highly reliable can be adjusted less than variables that are considered less reliable.

Another extension is for soft constraints, i.e. constraints that should approximately. These kind of constraints can be used to include subject matter knowledge in the model. For instance: the value of the stocks of some perishable good should be approximately equal to zero, since these kind of goods are usually not kept in stock. A nonzero value, is allowed, but the soft restriction prevents the occurrence of an undesirably high outcome.

Another possibility is to model the annual alignment as a soft constraint, for instance because some annual figure is considered not very reliable. As a consequence the annual data may be adjusted². A

² The reconciled annual figure is the sum of the underlying sub annual reconciled figures.

benchmarking problem with soft annual alignments is called *nonbinding benchmarking* by Dagum and Cholette (2006). Weights can be attached to soft constraints that determine the relative importance of each constraint.

Furthermore ratio constraints can be included in the model. This is a usual feature since ratio types of interrelationships may play an important role in practical applications. For instance, in the national accounts, it is often assumed that the ratio between production and intermediate use is quite constant over time.

Finally, it is possible to include inequality constraints. A very commonly used type of inequality is the requirement that values cannot be negative.

3. Examples – not tool specific

3.1 Example: the univariate Denton method

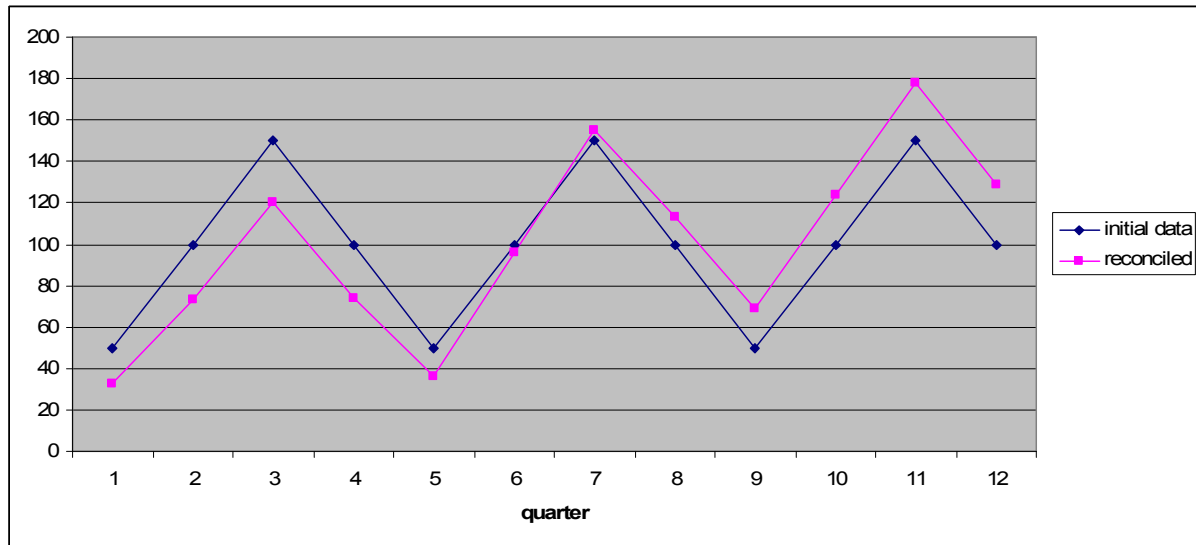
To illustrate the univariate model, we use an artificial data set of twelve quarters and three annual totals. The quarterly indicators were chosen to include pronounced changes that follow the seasons. They must then be modified in a way that the sums of the four quarters for each year exactly equal the corresponding annual totals. We imposed no other constraints. For convenience only one indicator series is used in the example.

Table 1. Input and output of a fictitious example

	Input		Output
	Quart. indicators	Annual	Reconciled data
Year 1	50	300	33
	100		73
	150		120
	100		74
Year 2	50	400	36
	100		96
	150		155
	100		113
Year 3	50	500	69
	100		124
	150		178
	100		129

We applied an univariate, proportional Denton method and rounded the results. The quarter-to-quarter changes were preserved as much as possible, see also figure 3. The reconciled data of the first year were lower than their initial values because of a lower annual total, and those of the third year were higher because of a higher annual total.

Figure 3. Graphical illustration of the results of table 1



3.2 Second example (Multivariate Denton method)

Let us consider a temporal disaggregation problem, consisting of two time-series, 12 quarters and two indicator time series x_1 and x_2 . Both indicator series are related to one time-series.

Suppose initially, each quarterly indicator value is 10 and that annual benchmarks are available for both time series. These are 50, 75 and 95 for the three consecutive years and both indicator time-series. Now assume that the first annual alignment is binding, whereas the second and the third are not.

This example is not very realistic, we intentionally choose for large discrepancies between the quarterly indicators and annual data in order to illustrate more vividly how the Denton method works.

Furthermore, there is one soft ratio constraint, defined by

$$\hat{x}_{1t} / \hat{x}_{2t} \approx 1.1 \quad \text{for } t = 1, \dots, 12.$$

and the proportional model will be used for both time series. Note that the soft, ratio constraint is inconsistent with the annual figures of both time series (both time-series have the same annual values, which implies a target value of 1 for the ratio). The relative values of the weights of both model components determine their influence on the model outcome.

The results of the benchmarking method, depicted in figure 4, are two time series, whose values increase gradually over time. This increase is due to the connection to the annual benchmarks. Further note as a result of the ratio from the fifth quarter onwards \hat{x}_{1t} increases more rapidly than \hat{x}_{2t} . During the first four quarters, the influence of the ratio constraint is negligible, since the quarters of both time series have to strictly add up to the same annual values. In the second and third year the annual alignment is soft, and therefore the ratio constraint is relatively more important than for the first year.

Figure 4 The benchmarked time-series

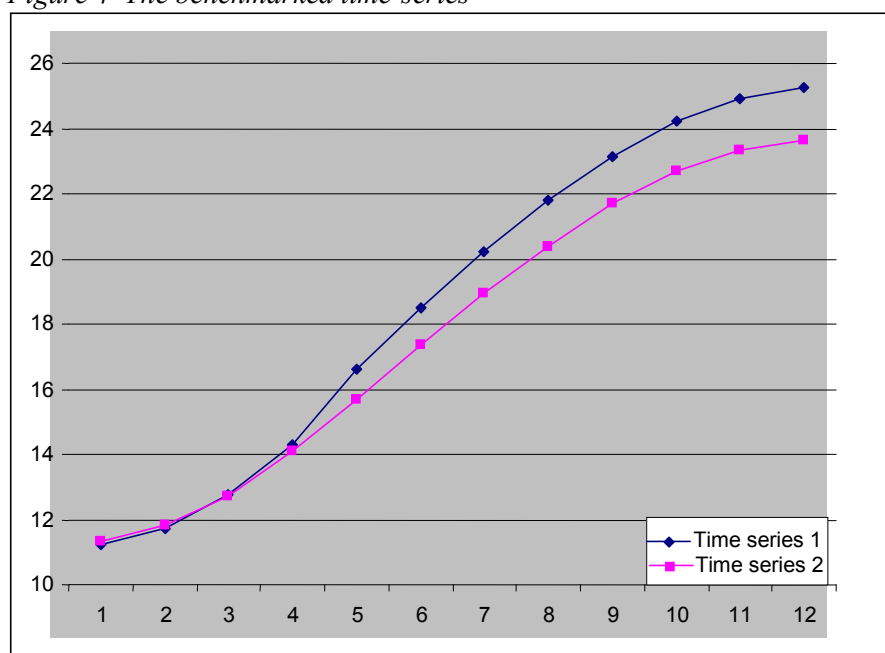


Table 2 shows that the reconciled annual figures (i.e. the sum of the underlying four quarterly values) of the second and third year closely approximate their target values.

Table 2. Reconciled annual figures, computed as the sum of four underlying quarterly values;

	Year 1	Year 2	Year 3
Time Series 1	50.00	77.16	97.61
Time Series 2	50.00	72.32	91.42
Time Series 1 / Time Series 2	1.000	1.067	1.068

4. Examples – tool specific

5. Glossary

Note 1: The definitions of the terms marked by an asterisk (*) are taken from the *Statistical Data and Metadata Exchange* (SDMX).

Note 2: The term “Benchmarking” also appears in the SDMX, but we give a different definition, because of the specific context of the problem.

Term	Definition
Accuracy*	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.
Annual Alignment	The <u>constraint</u> that annual data has to be consistent with sub annual <u>data</u> . Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Bias (of an	An effect which deprives a statistical result of representativeness by

estimator)*	systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average.
Benchmarking (hyper link to process step)	Achieving <u>consistency</u> between <u>data</u> that are published at different <u>frequencies</u> (for instance quarterly data that has to comply with annual data).
Binding constraint	See <u>hard constraint</u>
Coherence*	Adequacy of statistics to be combined in different ways and for various uses.
Consistency*	Logical and numerical <u>coherence</u> .
Constraint*	Specification of what may be contained in a <u>data</u> or metadata set in terms of the content or, for data only, in terms of the set of key combinations to which specific attributes (defined by the data structure) may be attached.
Contempeous constraints	Constraints within one period, between different <u>time-series</u>
Coverage error*	Error caused by a failure to cover adequately all components of the population being studied, which results in differences between the target population and the sampling frame.
Data*	Characteristics or information, usually numerical, that are collected through observation.
Data Integration*	The process of combining <u>data</u> from two or more sources to produce statistical outputs.
Data Reconciliation*	The process of adjusting <u>data</u> derived from two different sources to remove, or at least reduce, the impact of differences identified.
Data Set*	Any organised collection of <u>data</u>
Disaggregation*	The breakdown of observations, usually within a common branch of a hierarchy, to a more detailed level to that at which detailed observations are taken.
Frequency*	The time interval at which observations occur over a given time period.
Hard Constraint	A <u>constraint</u> that should hold exactly
Indicator*	A data element that represents statistical data for a specified time, place, and other characteristics, and is corrected for at least one dimension (usually size) to allow for meaningful comparisons.
Lagrange multiplier technique	In mathematical optimization, the technique of Lagrange multipliers (named after Joseph Louis Lagrange) provides a strategy for finding the maxima and minima of a function subject to <u>constraints</u> .
Macrodata*	The result of a statistical transformation process in the form of aggregated information.
Macro integration (hyper link to theme)	Integrating <u>data</u> from different sources on an aggregate level, to enable a <u>coherent</u> analysis of the data, and to increase the <u>accuracy</u> of estimates.
Measurement error*	Error in reading, calculating or recording numerical value.
Microdata*	Non-aggregated observations, or measurements of characteristics of individual units.
Micro integration	A method that matches data on individual statistical units from different sources, to obtain a combined data file with better information. The quality of the data is measured in terms of validity, reliability and consistency.
Missing data*	Observations which were planned and are missing
Movement preservation principle	The property that <i>all</i> changes of the <u>sub annual data</u> are kept as much as possible at their initial values.
Nonbinding benchmarking	A <u>benchmarking</u> problem with at least one <u>nonbinding annual alignment</u> constraint.
Nonbinding Constraint	See <u>soft constraint</u>
Pro-rata method	A simple <u>benchmarking</u> method in which the reconciled values are

	computed by dividing the <u>annual data</u> by the number of <u>sub annual</u> periods in one <u>annual period</u> . Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Soft constraint	A <u>constraint</u> that does not have to hold exactly, but approximately.
Step problem	The phenomenon of a large gap between the last sub annual period of one annual period and the first sub annual period of the next annual period. (for instance: a large gap between December and Januar). Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Temporal constraint	Constraints in the same <u>time-series</u> for different periods
Temporal Disaggregation (hyper link to process step)	Deriving <u>sub annual data</u> (for instance quarterly data) from <u>annual data</u> , by using <u>indicators</u> of the <u>sub annual data</u> (i.e. related time series), see <u>disaggregation</u> . Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.
Time Series*	A set of ordered observations on a quantitative characteristic of an individual or collective phenomenon taken at different points of time.
Weight*	The importance of an object in relation to a set of objects to which it belongs.

6. Literature

Barcellan R and D. Buono (2002), 'ECOTRIM interface user manual', Eurostat

Bikker, R.P. and S. Buijtenhek (2006), Alignment of Quarterly Sector Accounts to Annual Data, Statistics Netherlands, Voorburg, http://www.cbs.nl/NR/rdonlyres/D918B487-45C7-4C3C-ACD0-0E1C86E6CAFA/0/Benchmarking_QSA.pdf

Bikker R.P., J.A. Daalmans and N. Mushkudiani (2010), A multivariate Denton method for benchmarking large data sets, Discussion Paper 10002, Statistics Netherlands. <http://www.cbs.nl/NR/rdonlyres/7B2387F2-5773-42CF-8C50-5F02B451A2E4/0/201002x10pub.pdf>

Bloem, A.M., Dippelsman, R.J., Maehle, O.N. (2001) Quarterly National Accounts Manual: Concepts, Data Sources, and Compilation, International Monetary Fund, Washington, DC. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/CA-22-99-781/EN/CA-22-99-781-EN.PDF

Boonstra H.J., C.J. De Blois, and G.J. Linders (2010), Macro integration with inequality constraints an application to the integration of transport and trade statistics, Statistics Netherlands,

Boot J.C.G., W. Feibes and J.H.C. Lisman (1967), Further methods of derivation of quarterly figures from annual data, Cahiers Economiques de Bruxelles, 36: 539-546.

Chen B. (2007), An Empirical Comparison of Methods for Temporal Distribution and Interpolation at the National Accounts, Bureau of Economic Analysis

Dagum E.B. and P.A. Cholette (2006), *Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series*, Springer, New York.

- Denton F.T. (1971), Adjustment of monthly or quarterly series to annual totals: An Approach based on quadratic minimization, *Journal of the American Statistical Association*, 66, 333, pp. 99-102.
- Di Fonzo, T. (1990), The Estimation of M Disaggregate Time Series when Contemporaneous and Temporal Aggregates are Known, *The review of Economics and Statistics*, 72, pp. 178-182.
- Di Fonzo, T. and M. Marini (2003), *Benchmarking systems of seasonally adjusted time series according to Denton's movement preservation principle*. University of Padova, Working Paper 2003.09 <http://www.oecd.org/dataoecd/59/19/21778574.pdf>.
- Fernández, R.B. (1981), Methodological note on the estimation of time series, *Review of Economic and Statistics*, 63, no.3, p. 471-478.
- Magnus, J.R., J.W. van Tongeren and A.F. de Vos, (2000), National Accounts Estimation Using Indicators Analysis, *Review of Income and Wealth* (46), pp. 329-350.
- Magnus, J.R. and D. Danilov, (2008), On the estimation of a large sparse Bayesian system: The Snaer program, *Computational Statistics & Data Analysis*, 52, pp. 4203-4224.
- Wei W.W.S. and D.O. Stram (1990), Disaggregation of time series models, *Journal of the Royal Statistical Society*, 52, 453-467.

Specific description – Method: Denton for temporal disaggregation

A.1 Purpose of the method

The method is used for temporal disaggregation ([hyperlink](#)), which is a specific process step used in the context of Macro integration ([hyperlink](#)).

A.2 Recommended use of the method

1. The method can be applied to any problem in which sub annual series have to be estimated consistent with a given annual time-indicator series, assuming that the preconditions in Section A.6 are satisfied
2. The method may be applied on micro or macro data
3. It is especially useful for time-series with seasonal patterns, for instance national accounts data.
4. As mentioned by Bloem et al. (2001) the method is very well suited for large scale applications.
5. It is especially useful when users of economic statistics require data more frequently than the availability of the sources of these sources.
6. The indicator series should be correlated with the time-series to be estimated as much as possible, in particular the trend of the indicator series should not differ too much from the trend of the time-series to be estimated.
7. The method should be applied to unbiased source figures. All source figures are consistent with their definitions. They are therefore already adjusted for systematic errors (nonresponse errors, measurement errors, processing errors and conceptual differences). Any errors in the input data (as mentioned in subsection A.5) will propagate to the results.

A.3 Possible disadvantages of the method

1. Do not use Denton-for-temporal-disaggregation when the annual values are less reliable than the annual sums from the sub-annual indicator series. In this case, using Denton-for-temporal-disaggregation will essentially diminish the reliability of sub annual time-series.
2. The method must not be applied in case there is no or little correlation between the time-series to be estimated and the available indicator time-series, for then the method would introduce an artificial, erroneous correlation in its estimates.

A.4 Variants of the method

1. A Denton method preserves as much as possible the period-to-period changes of the initial sub- annual indicator time-series. Variants of the Denton method that differ in the way how these changes are defined:
 - 1.1 The additive model attempts to keep the additive corrections as constant as possible over all periods.

- 1.2 The proportional model is designed to preserve the percentage differences as constant as possible over all periods.
- 1.3 The additive and proportional model may be combined when the Denton method is applied to a multivariate data sets, i.e. the additive method may be applied to some time-series, and the proportional model to the other time-series.
2. The Denton method can minimize
 - 2.1 First order differences or
 - 2.2 Second order differences (differences of differences).

Remark: in the applications that are mentioned in the literature always first-order differences are used.

A.5 Input data sets

1. Ds-input1 = a data set (micro data or macro data) with indicators of the sub annual time-series (required). In a multivariate model one indicator time-series may be used as a predictor of more than one time-series
2. Ds-input2 = a data set (micro data or macro data) with annual figures (required)

Remark: each time-series may comprise one or several annual periods.

A.6 Logical preconditions

A.6.1 Missing values

1. In Ds-input1 individual missing data values are not allowed.

Remark: In case of missing source data, one may use a dummy time-series, consisting of the value '1' (or any other value) for all sub-annual periods.

2. In Ds-input2 individual missing data values are allowed (meaning that there is no annual alignment).

A.6.2 Erroneous values

- 1.

A.6.3 Other preconditions

1. One annual period covers a whole number of sub annual periods
2. The constraints (Subsection A.7-2) must not be mutually conflicting. In particular this means that: for a multivariate benchmark problem, in which :
 - all annual alignments (Subsection A.7-2-a) are binding,
 - contemporaneous constraints (Subsection A.7-2-b) are defined, that have to be exactly satisfied;

the annual values, that are input to the model (Subsection A.5-2), also have to satisfy all the contemporaneous constraints.

3. The proportional variants of the Denton method (Subsection A.4-b) can only be applied if the initial sub annual indicators (Subsection A.5-1) do not contain any zeros.

A.7 Tuning parameters

1. Weights (Optional). These weights determine which of the indicator time-series are adjusted the most (in a multivariate model) and which of the soft constraints are the most stringent. Weights may be omitted; in that case all indicators and constraints are considered equally reliable.
2. Constraints. These specify the constraints that should be satisfied. The following type of constraints can be used:
 - a. Annual alignment: a sum of sub annual values that should add up to an annual value (This is a fixed constraint, meaning that it is typical for the Denton method).
 - b. Contemporaneous constraints (for the multivariate case only): constraints between different time-series, within the same period (Optional).

Technically, a distinction can be made between

- i. soft constraint en hard constraints;
- ii. inequality constraints and equality constraints;
- iii. linear and nonlinear constraints.

A.8 Recommended use of the individual variants of the method

1. The proportional model (A.4 variant 1.2) is often preferred in application to economic data. The reason for this is that this model better preserves the seasonal changes, as these changes are often measured as a percentage change. However an additive model should be used:
2. The additive model (A.4 variant 1.1) should be used for time-series that include zeros, for the proportional model is not defined properly for such time-series. The model attempts to preserve the initial percentage changes. A percentage change, however is not defined when the value of the first period is zero.
3. The additive model (A.4 variant 1.1) should be used for time-series that involve both positive and negative values. A proportional model may yield undesirable results, for instance that each sub annual value is multiplied by some negative constant, meaning that all signs change.

A.9 Output data sets

1. Ds-output1 = a dataset with estimated (micro or macro data) sub annual time-series.

A.10 Properties of the output data sets

1. The estimated time-series (Subsection A.9-1) satisfy all restrictions (Subsection A.7-2).
2. The seasonal pattern is as close as possible to the seasonal pattern of the initial sub annual indicator time-series (Subsection A.4-1)

A.11 Unit of processing

Processing full data sets

A.12 User interaction - not tool specific

1. Before execution of the method, the tuning parameters and input datasets must be specified.
2. During operation no user interaction is needed, but inspection of quality indicators and subsequent adjustment of tuning parameters and recurrent use is optional.
3. After use of the method the quality indicators and logging should be inspected.

A.13 Logging indicators

1. The run time of the application
2. Characteristics of the input data, for instance problem size, the largest discrepancies of the input data towards the constraints.

A.14 Quality indicators of the output data

1. The most important quality indicator is *how* the sub annual time-series are modified. Of particular interest are the changes made in the first differences, given the starting point of the Denton method. The size of these changes is important, but the trend of the changes in time is particularly important. This trend can usually be assessed fastest by graphical means.

Remark 1: The Denton method tries to adjust the sub annual time-series so that the adjustments are as smooth as possible over time. If the model needs to largely adjust the initial changes, this may be an indication that the preconditions of the model are not satisfied and therefore that the method should not have been applied. The ratio between the reconciled and the raw data should be stable over time.

Remark 2: A quality indicator of an implementation of the method in a tool is how accurately the sub annual series is aligned with the annual series. Numerical error will generally cause these differences to deviate slightly from zero. The differences are not usually a problem as long as they are less than a certain threshold value.

A.15 Actual use of the method

1. n/a

A.16 Relationship with other modules

A.16.1 Themes that refer explicitly to this module

A.16.2 Related methods described in other modules

1. RAS method ([hyper link](#))
2. Method of Stone ([hyper link](#))
3. Denton for benchmarking ([hyper link](#))

Remark: the method of Stone and RAS are also data reconciliation methods, but these are not specially aimed at benchmarking and temporal disaggregation.

A.16.3 Mathematical techniques used by the method described in this module

1. Quadratic optimization under linear constraints (hyper link).

A.16.4 GSBPM phases where the method described in this module is used

1. GSBPM phase 6.2 “Validate Outputs” (hyper link)

A.16.5 Tools that implement the method described in this module

1. ECOTRIM

Remark: Freely available from: <http://circa.europa.eu/Public/irc/dsis/ecotrim/library> However, ECOTRIM is not designed for dealing with thousands of time series, it does not include features like weights, ratio's, soft constraints, and the possibility to combine the proportional and additive methods of benchmarking into one model.

A.16.6 The Process step performed by the method

Temporal disaggregation (hyper link)



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Chow-Lin Method for Temporal Disaggregation

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Method.....	3
2.2 Properties.....	4
3. Preparatory phase	6
4. Examples – not tool specific.....	6
4.1 Example: The Chow-Lin Regression based approach for the multivariate case	6
4.2 Example: The Chow-Lin Regression based approach for the univariate case	8
5. Examples – tool specific.....	9
6. Glossary.....	10
7. References	10
Specific section.....	11
Interconnections with other modules.....	13
Administrative section.....	14

General section

1. Summary

The Chow-Lin method is a technique used for temporal disaggregation or also known as temporal distribution. Temporal disaggregation is the process of deriving high frequency data (e.g., monthly data) from low frequency data (e.g., annual data).

In addition to the low-frequency data, the Chow-Lin method also uses indicators on the high-frequency data, which contain the short-term dynamics of the time series under consideration. These indicators are time series that are related to the target time series and thus measure a different topic than the time series to be estimated. Since the results of the Chow-Lin method depend on information on a different variable, the method can be considered as an indirect approach. The main goal of temporal disaggregation techniques such as the Chow-Lin method is to create a new time series that is consistent with the low frequency data while keeping the short-term behaviour of the higher frequency indicator series. The Chow-Lin method may be applied to time series, generally aided by one (univariate case) or more indicator series (multivariate case). Presumably, these indicator series should be socio-economic variables deemed to behave like the target variable. In absence of such variables, functions of time can be used, as proposed by Chow and Lin (1976).

Temporal disaggregation is closely related to benchmarking, another data integration technique, see for instance the module “Macro-Integration – Denton’s Method”. The main distinction is that in benchmarking, the sub-annual series to be benchmarked consists of the same variable as the annual benchmarks, while in temporal disaggregation the sub-annual series differ from the annual series.

2. General description of the method

In this section we give a non-technical description of the Chow-Lin method. For a more elaborate and technical explanation we refer to Dagum and Cholette (2006) and Chow and Lin (1971).

2.1 Method

Temporal disaggregation is the process of deriving consistent high frequency data from low frequency data. In the past few decades, different methods for disaggregating low frequency data to high frequency data have been developed. These methods can be classified into two types of approaches.

- Models developed without indicator series but which rely upon purely mathematical criteria or time series models to derive a smooth path for the unobserved series;
- Models based on indicator series observed at the desired higher frequency.

In the second case it is assumed that one or more correlated high frequency series are available. The approach includes: the procedure proposed by Denton (1971), which is an adjustment method that does not rely on any statistical model, and ‘optimal’ methods proposed by Chow and Lin (1971) and further developed by Fernandez (1981) and Litterman (1983), which use the best linear unbiased estimator, given some assumed model (Islam, 2009).

The Chow-Lin method is a temporal disaggregation technique that uses a (statistical) relationship between low frequency data and higher frequency indicator variables. This is done by a univariate

regression in the case with one indicator and a multivariate regression if two or more indicators are used. The regression coefficients are estimated at the low frequency level, at this level the target time series and the indicator time series are both available. The indicator series are assumed to be at the high-frequency level, but these data can easily be transformed to the low frequency level, just by aggregating over the high frequency periods.

The results of the regression are used to estimate the high-frequency target series from the high-frequency indicator series. The univariate as well as the multivariate regression consists of two parts. The first part is a normal regression which estimates the monthly values. These values may not be consistent over time with the original low frequency data because a simple regression does not account for consistency. To make up for possible inconsistency, the second part of the regression corrects the estimated monthly values.

2.2 Properties

We can classify time series into three different types: stocks series, flows series and index series.

Flows series measure how much of something has happened over a period of time, for instance exports, production and GDP. Stock series measure a quantity existing at some specific time point, for instance inventories of some good. Temporal disaggregation applied on stock variables is also known as *interpolation*, while the application of the method on flow variables is called *distribution*, see also Chow and Lin (1976).

In this article we only discuss flows series. For these series to be disaggregated it is necessary to do this consistently in the sense that temporal additivity is observed. This basically means that the sum of the three months of the (estimated) disaggregated series must add up to the value of the associated quarter of the original series. An illustration of the additivity constraint is displayed in Figure 1 and Table 1 using index of manufacturing as high frequency indicator.

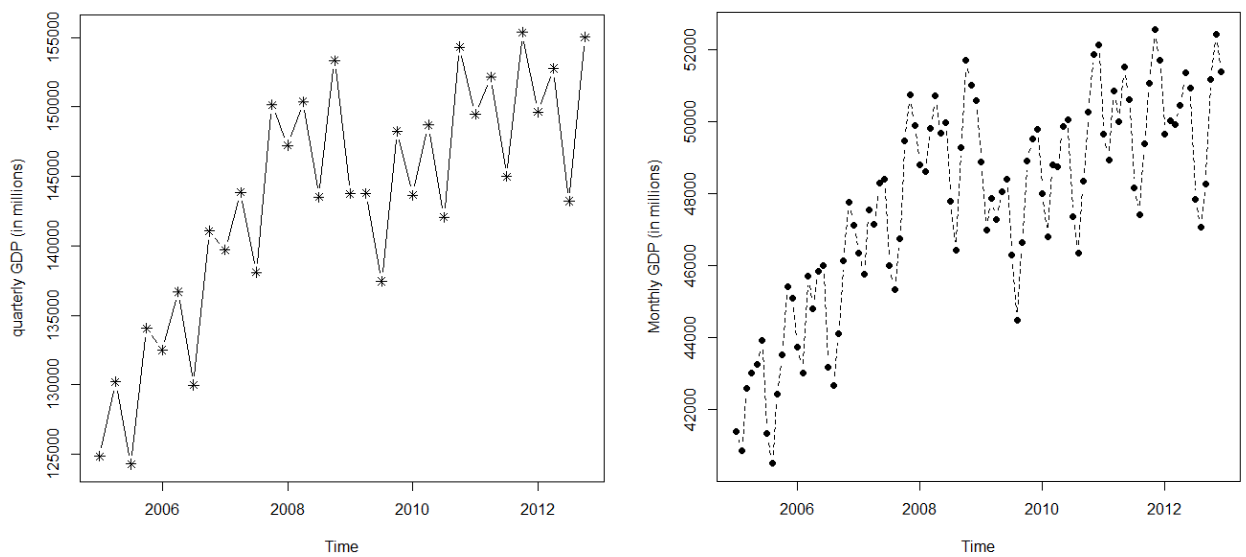


Figure 1. Plots of the quarterly GDP values (left) and estimated monthly values (right)

Table 1. Values of the estimated monthly GDP and quarterly GDP. **Bold** = added quarterly estimates

	Estimated monthly GDP	Quarterly GDP
Jan-05	41398,03	
Feb-05	40865,45	
Mar-05	42587,53	
2005 Q1	124851	124851
Apr-05	43023,87	
May-05	43256,36	
Jun-05	43932,77	
2005 Q2	130213	130213

Table 1 shows the values of the first two quarters in 2005 for the GDP. Also, the estimated monthly GDP values are shown. It is clear that the monthly values add up to the quarterly values.

There is a number of reasons why the Chow-Lin method (or any other disaggregation method) is needed. Quantitative analyses performed by for example National Statistic Institutes (NSIs) rely on statistical data. These data are generally obtained from sample surveys. Due to the need of large resources and high costs to conduct these surveys, NSIs could choose to do this only a few times per year, for example quarterly (low frequency). On the other hand, for efficient statistical and economic analysis and timely decision-making, a higher frequency may be required (Chamberlin, 2010).

Furthermore, low frequency data sources are more precise and describe the long-term movements better than high frequency sources, but the latter provide the only information on the short-term movements. By applying a temporal disaggregation technique one combines the strengths of both frequency types.

There are several variants of the Chow-Lin method. The difference between these variants is the estimation of the covariance matrix of the residuals, which in reality is often unknown. The residuals are defined as the difference between the actual monthly and estimated monthly values. Chow and Lin (1971) propose two assumptions to estimate the covariance matrix.

- There is no serial correlation amongst the residuals. This means that when the monthly values are estimated, the levels of the high frequency indicator are used.
- The residuals are serially correlated, which means that the changes and fluctuations of the high frequency indicators are taken in the monthly estimates instead of the level.

The unknown covariance matrix needs to be estimated and one can estimate this matrix based on the first or second assumption dependent on empirical evidence or presumption of the user. The assumption of no serial correlation in the residuals of monthly estimates is generally not supported by empirical evidence. Chow and Lin propose a method to estimate the covariance matrix under the second assumption that the residuals follow a first order autoregressive AR(1) process. We will not elaborate here, as it is out of the scope of this text, but this covariance matrix is eventually used to estimate high frequency values. A slightly other covariance matrix will cause different high frequency values, although the additivity constraint will still hold independent of the variants used. In short, if the user presumes that the residuals are serially correlated which implies that the fluctuations are more important than the levels, one should use an AR(1) model for the residuals. If not, the residuals need

not be modelled and so the Chow-Lin method boils down to Ordinary Least Squares (a normal regression).

3. Preparatory phase

4. Examples – not tool specific

4.1 Example: The Chow-Lin Regression based approach for the multivariate case

To illustrate the Chow-Lin method for the multivariate case, we use a dataset obtained from Statistics Netherlands. Our dependent variable will be the Dutch GDP measured quarterly from 2005 till 2012. We will use three monthly indicators as explanatory variables where all three have higher frequency (monthly); index of manufacturing, inflation rate and the unemployment rate, as one may assume a correlation between these variables and GDP.

Hence we have three times more indicator observations than GDP observations. Before we continue this example, we want to stress that we do not use this technique at Statistics Netherlands. This example is purely for illustration of the Chow-Lin method.

In Figure 2 the plots of the GDP and the other monthly economic indicators are shown. The results of the regression used to derive the monthly GDP estimates are stated in Table 2. The unemployment rate and index of manufacturing are able to explain a lot more than the inflation as the latter is statistically not significant.

Table 2. Results regression

	Estimate	Standard Error	t-statistic
<i>Intercept**</i>	36751.07	5176.65	7.099
Inflation	1177.68	728.31	1.617
Unemployment *	-1246.98	560.83	-2.223
Index of Manufacturing**	158.83	43.25	3.672

*Significance level $\alpha = 0.05$, ** $\alpha = 0.01$

Figure 3 presents the monthly GDP estimates from the Chow-Lin regression approach, where we choose to use an AR(1) model to allow for serial correlation in the residuals as proposed by Chow and Lin (1971). In this example all three explanatory variables displayed in Table 2 are used, although inflation could have been left out as it is insignificant.

For comparison purposes, the original quarterly series (from Figure 1) is also plotted in Figure 3. It is clear that both frequencies approximately have the same behaviour but the monthly estimates show more detail in terms of fluctuations in the short term behaviour.

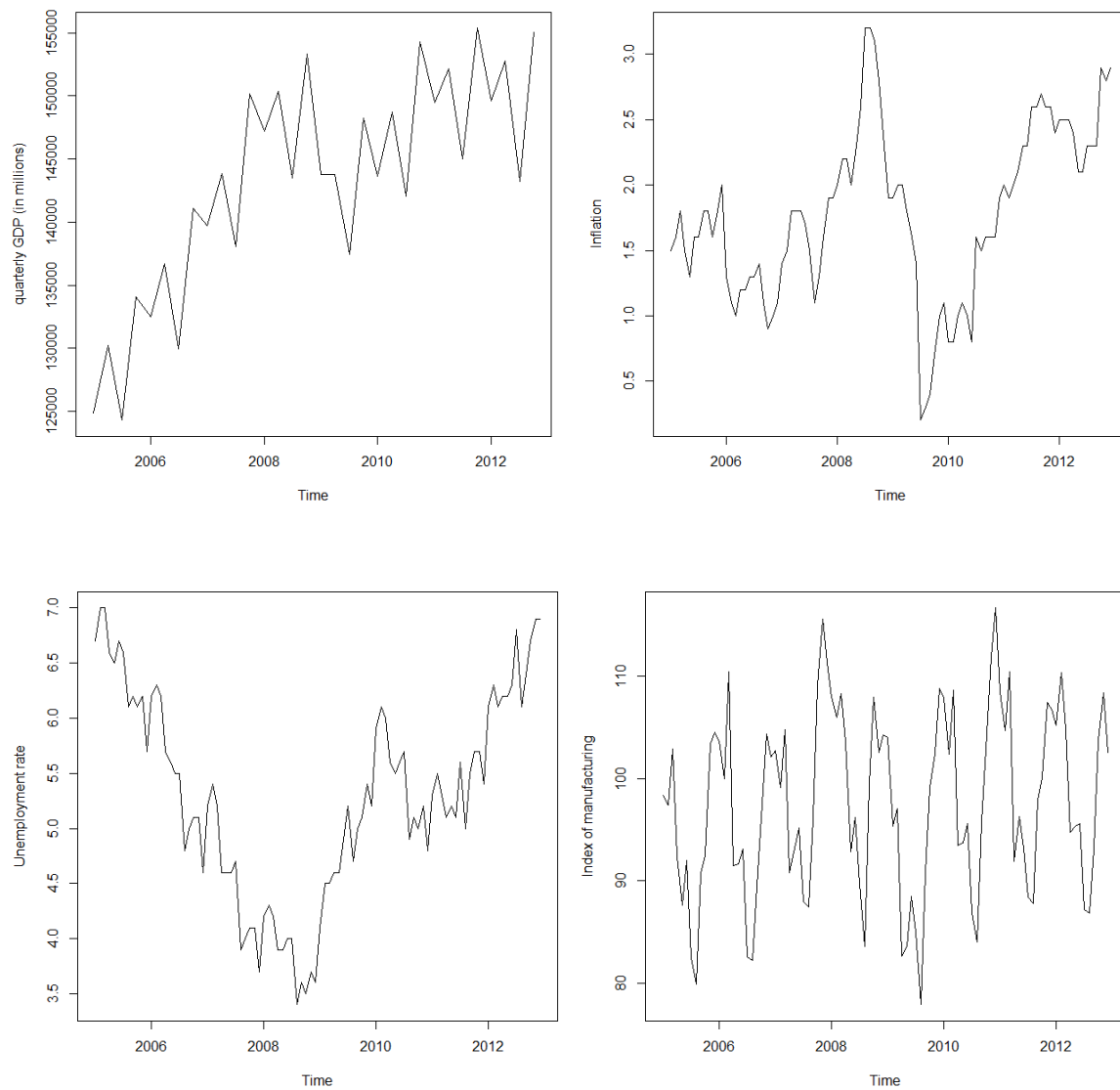


Figure 2. Plots of the low frequency (GDP) and high frequency series.

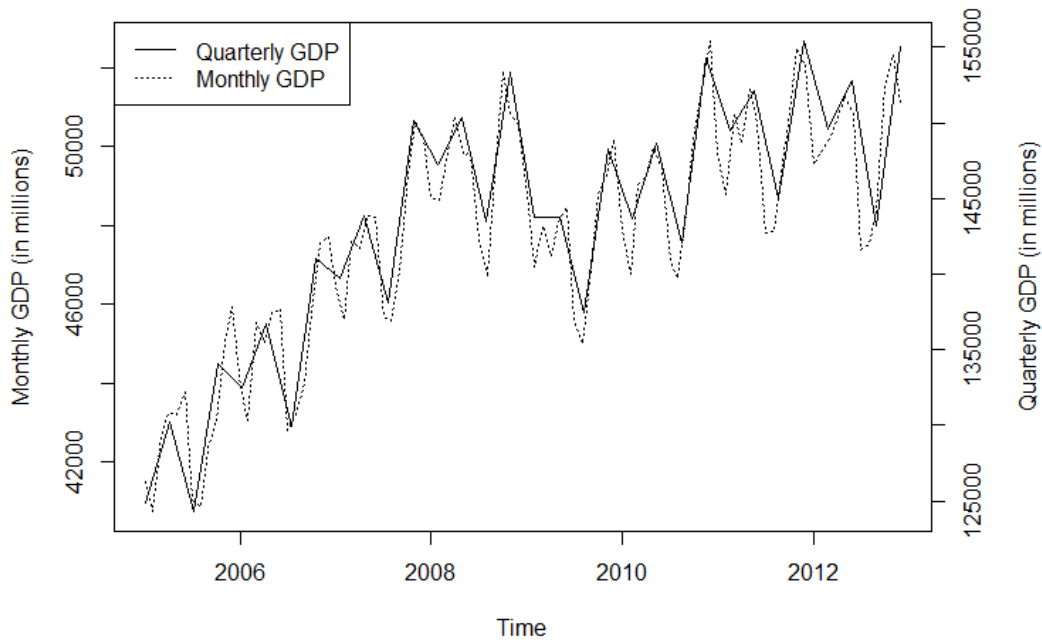


Figure 3. Monthly GDP estimates

4.2 Example: The Chow-Lin Regression based approach for the univariate case

Here we look at the univariate case of the Chow-Lin method. For this example we only use one high frequency indicator series, consumer credit, to estimate the values of the GDP month-to-month changes (low frequency series). The original GDP series is measured in terms of quarter-to-quarter changes (in percentages). In Figure 4, the GDP (left) and the consumer credit (right) are plotted.

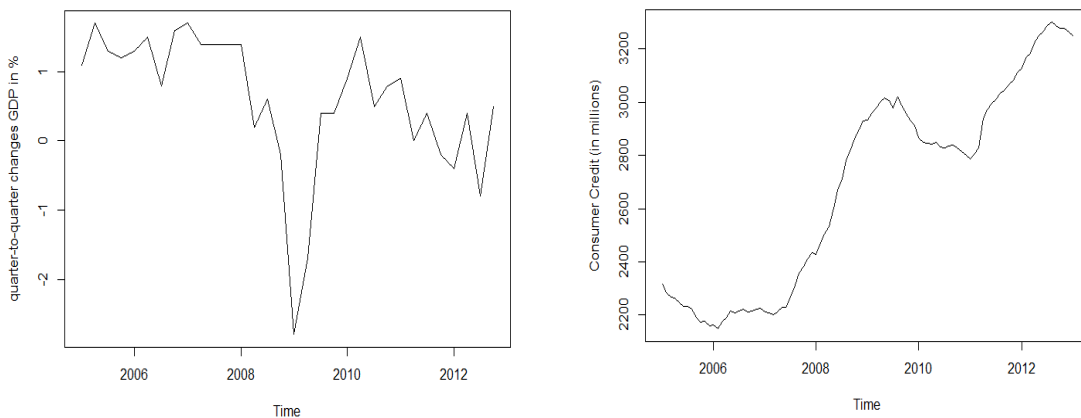


Figure 4. Plots of the GDP (left) and the consumer credit (right).

In Table 3, the result of the univariate regression is shown. Once again, we model the serial correlation in the residuals with the means of an AR(1) model. From the results stated in Table 3 we can see that the consumer credit has an significant effect on the GDP. With a R^2 of 0.27, consumer credit explains a reasonable part of the GDP's movement.

Table 3. Regression results of the univariate regression

	Estimate	Standard Error	t-statistic
<i>Intercept**</i>	1.7171	0.4421	3.884
Consumer Credit**	-0.0006	0.0002	-3.461

**Significance level $\alpha = 0.01$

When we make a plot of the estimated high frequency GDP and compare this to the original quarterly GDP series, it is clear from Figure 5 that the monthly GDP has the same fluctuation pattern as the quarterly GDP series. The only difference is that the estimated monthly GDP is more smooth.

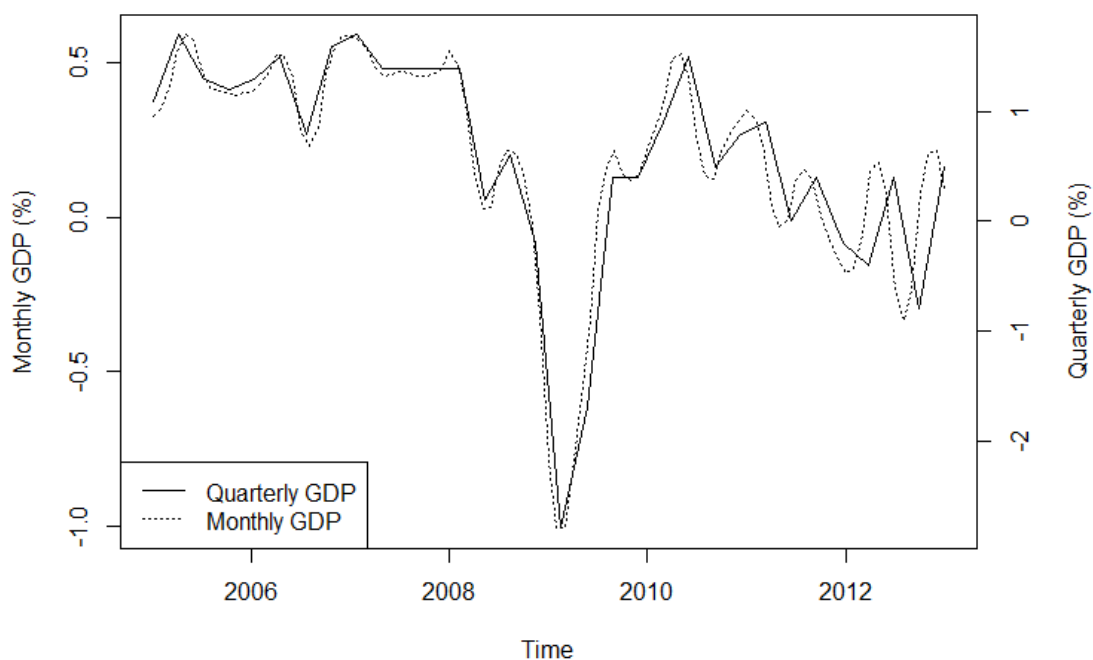


Figure 5. Comparison of the estimated monthly GDP with the quarterly GDP

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Barcellan, R. and Buono, D. (2002), *ECOTRIM interface user manual*. Eurostat.
- Chamberlin, G. (2010), Methods Explained: Temporal disaggregation. *Economic and Labour Market Review* **4**, 106–121.
- Chow, G. and Lin, A. (1971), Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. *The Review of Economics and Statistics* **53**, 372–375.
- Chow, G. and Lin, A. (1976), Best linear unbiased estimation of missing observations in an economic time series. *Journal of the American Statistical Association* **71**, 719–721.
- Dagum, E. B. and Cholette, P. A. (2006), *Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series*. Springer, New York.
- Denton, F. T. (1971), Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization. *Journal of the American Statistical Association* **66**, 99–102.
- Fernández, R. (1981), A methodological note on the estimation of time series. *The Review of Economics and Statistics* **63**, 471–478.
- Islam, M. R. (2009), Evaluation of different temporal disaggregation techniques and an application to Italian GDP. *BRAC University Journal* **VI**, No. 2, 21–32.
- Litterman, R. (1983), A random walk, Markov model for the distribution of time series. *Journal of Business and Economic Statistics* **1**, 169–173.
- Mitchell, J., Smith, R. J., Weale, M. R., Wright, S., and Salazar, E. L. (2005), An Indicator of Monthly GDP and an Early Estimate of Quarterly GDP Growth. *Economic Journal, Royal Economic Society* **115**, F108–F129.
- Rizk, M. (2010), Temporal Disaggregation of the Quarterly Real GDP Series: Case of Egypt. Research Department Working Paper, Central Bank of Egypt.

Specific section

8. Purpose of the method

The method is used for temporal disaggregation which disaggregates a low frequency time series to a higher frequency series.

9. Recommended use of the method

1. This method can be applied to any dataset that contains variables with different frequencies and where the user needs a dataset with the same frequency for all variables.
2. The method is useful if one wants to combine the strengths of the short-term (high frequency) and the long-term data (low frequency).
3. The method is more applicable for flows series as these are used for economic statistics.
4. Application of this method is only possible if besides the low frequency series, other high frequency indicators are available.

10. Possible disadvantages of the method

1. The results rely on a covariance matrix, which is often unknown in practice and needs to be estimated on the basis of assumptions.

11. Variants of the method

1. Based on the Chow-Lin method these variants are expansions of the Chow-Lin method. The only difference is the computation of the covariance matrix, which is used for the monthly estimates.

1.1 Fernandez (1981)

1.2 Litterman (1983)

12. Input data

1. A low frequency time series (i.e., quarterly).
2. One or more high frequency indicator time series (i.e., monthly).
3. The value for serial correlation in the covariance matrix; if the user knows this beforehand or assumes a value. If not specified, the value for serial correlation can be estimated.

13. Logical preconditions

1. Missing values
 1. High frequency indicator may not contain missing values.
 2. Low frequency series may contain missing values. When the low frequency value for the last period is missing, the Chow-Lin method can be used to forecast that missing value (Mitchell et al, 2004).
2. Erroneous values

1. Erroneous low frequency input values will be preserved in the output.
2. Erroneous high frequency input values can result in invalid high-frequency output series. But the high-frequency series that is obtained will still be consistent with the low-frequency data.
3. Other quality related preconditions
 1. One quarterly period covers a whole number of monthly periods. As for annual periods; these cover a whole number of sub-annual periods.
 2. For the short-term movements to be incorporated in the estimated target series, proper high frequency indicators should be used in the sense that they have a (statistical) relationship with the low frequency series.
4. Other types of preconditions
 - 1.
 - 2.

14. Tuning parameters

15. Recommended use of the individual variants of the method

1. Different time series are called co-integrated, if they do not drift apart during time. For example, consumption and income are co-integrated, if consumption is roughly a constant proportion of income over the long term. If the target series and the high frequency series are not co-integrated, the Fernandez (1981) or Litterman (1983) variant is preferred (Rizk, 2010).

16. Output data

1. Ds-output1 = a dataset with estimated high frequency values derived from the original low frequency series and high frequency indicators.

17. Properties of the output data

1. The estimated high frequency data add up to the original low frequency values. In that sense, the output is consistent.
2. The short-term movement of the high frequency indicators is incorporated.

18. Unit of input data suitable for the method

Processing full datasets.

19. User interaction - not tool specific

1. Before execution of the method, the input datasets must be specified.
2. During operation no user interaction is needed.
3. After use of the method the quality indicators should be inspected.

20. Logging indicators

- 1.

21. Quality indicators of the output data

1. The consistency of the high frequency estimates should be checked, i.e., the values should add up to the original low frequency values. Furthermore, the (short-term) behaviour is of particular interest. The short-term behaviour of in- and output time series should somewhat have the same behaviour. This can be checked fast and easy by graphical means.

22. Actual use of the method

1. Intensively used by National Statistical Institutes, especially in France, Italy, Belgium, Portugal and Spain (Barcellan and Bueno, 2002).

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Macro-Integration – Main Module

24. Related methods described in other modules

1. Macro-Integration – Denton’s Method

25. Mathematical techniques used by the method described in this module

1. Linear regression with linear constraints

26. GSBPM phases where the method described in this module is used

1. GSBPM phase 6.2 “Validate Outputs”

27. Tools that implement the method described in this module

1. Matlab
2. R
3. ECOTRIM – Eurostat has made an application program named “ECOTRIM” for the use of several temporal disaggregation methods.

28. Process step performed by the method

Temporal disaggregation

Administrative section

29. Module code

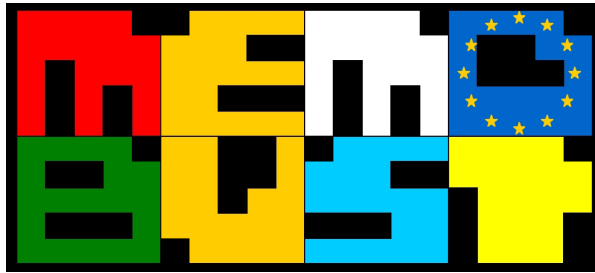
Macro-Integration-M-Chow-Lin

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-05-2013	first version	Feysel Negash	CBS
0.2	04-10-2013	review comments Roberto Iannaccone adopted	Feysel Negash Jacco Daalmans	CBS
0.3	30-10-2013	review comments Editorial Board adopted	Jacco Daalmans	CBS
0.3.1	31-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	
Print date	



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Asymmetry in Statistics – European Register for Multinationals (EGR)

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Outward FATS	3
2.2 Inward FATS	4
2.3 Key population characteristics	4
2.4 Organisation of the statistical production processes.....	4
2.5 Frame population methodology.....	4
2.6 On frames, target populations, frame populations and survey populations.....	5
2.7 Frames and frame populations in a multi-user environment	5
2.8 EGR methodology on FATS frame populations	7
2.9 EGR benefits	9
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	11
Administrative section.....	12

General section

1. Summary

All over the world, globalisation is seen as the predominant agent of change and the main policy challenge. At the heart of this complex and somewhat blurry concept, however, lie businesses and their ever-increasing drive to expand their activities across national borders, most notably by establishing foreign affiliates. Europe plays a key role in this. The EU has become a very important destination for foreign companies and their affiliates, and European businesses are among the most active around the world. A host of crucial policy challenges flow from this, not least the issue of outsourcing jobs and keeping European firms competitive. Consequently, there is a huge and ever-growing demand for data on these developments. Foreign Affiliates Statistics (FATS) statistics are particularly useful because they help explain how businesses are expanding internationally and what the consequences are for the European Union.

Every EU/EFTA country has to compile Outward and Inward FATS statistics. Within a country a National Statistical Institute or a National Central Bank is appointed as compiler. Some countries produce also intra-EU data for Outward FATS. In theory intra-EU statistics compiled by Outward FATS of country A for EU/EFTA country B should be equal to Inward FATS statistics of country B.

Example from praxis:

Outward FATS of country A produces the following figure: multinational enterprise groups controlled by a resident legal entities in country A are controlling 1101 enterprises resident in country B. Inward FATS of country B should produce the following figure: 1101 enterprises resident in country B are controlled by a legal entity in country A. However country B publishes 505 enterprises, a difference of 596 enterprises.

These differences or asymmetries can have different causes. This paper is dealing with one type of cause: differences in frame populations and how register methodology, in this case frame population methodology of the EuroGroups register (EGR) can contribute to reduce/eliminate these kinds of differences.

2. General description

2.1 Outward FATS

The **Outward FATS target population of statistical units** is composed of all foreign enterprises (see also the module “Statistical Registers and Frames – The Statistical Units and the Business Register”) located in extra-EU countries or intra-EU countries that are controlled by an institutional unit resident in an EU Member State. ‘Foreign enterprise’ shall mean an enterprise not resident in the compiling country over which an institutional unit resident in the compiling country has control.

The **Outward FATS target population of reporting units** differs from the population of statistical units. To identify the relevant target population of reporting units and to unambiguously associate the statistical units with them: the ‘Ultimate Controlling Institutional unit approach’ (UCI) is applied.

2.2 *Inward FATS*

The **Inward FATS target population of statistical units** comprises all enterprises and all branches¹ under foreign control. Statistical data have to be allocated to the country of residency of the ‘Ultimate Controlling Institutional unit’ (UCI). The Inward FATS target population is a subset of the target population of Structural Business Statistics (SBS). Target populations of statistical and reporting units are equal in inward FATS, as data are collected directly from enterprises and branches on which information is needed.

2.3 *Key population characteristics*

Two kinds of statistics are compiled: Inward and Outward FATS. The reference period is the calendar year. Presently the scope of Inward FATS is restricted to enterprises resident in EU/EFTA country classified in B to N and PQRS of NACE rev.2. Member States are obliged to compile extra EU outward FATS data and are not obliged to compile intra-EU statistics. Nonetheless, given the users interest in this information, Member States are asked to compile these data on a voluntary basis.

- Inward and Outward FATS use the same statistical unit: the enterprise
- For the EU area the populations of enterprises for Inward and Outward FATS are overlapping.
- The Inward FATS population of enterprises is a sub-population of the Outward FATS population of enterprises
- The Inward FATS population of enterprises is a sub-population of the SBS target population of enterprises
- The ‘Ultimate Controlling Institutional unit’ (UCI) is a common concept applied to the population of reporting units and to define for enterprises the country of foreign control (=country of residency of the ‘Ultimate Controlling Institutional unit’ (UCI)).

2.4 *Organisation of the statistical production processes*

FATS statistics are produced by National Statistical Institutes or National Central Banks in 31 EU+EFTA countries. Every organisation is designing and implementing its own statistical production process.

2.5 *Frame population methodology*

The challenge is the definition and implementation of a methodology (called: frame population methodology) which guarantees that the survey populations used in statistical production processes on globalisation:

- a) don’t have double counting nor have ‘gaps’
- b) are synchronised in case of sub- or overlapping populations used in different statistical activities
- c) have identical classifying characteristics like NACE and country code (activity and geographical breakdown)

¹ Branches under foreign control are considered as quasi enterprises.

- d) are based on a common view on units, e.g., statistical unit enterprise or ‘Ultimate Controlling Institutional unit’ (UCI).

Frame population methodology is guideline system consisting of rules, procedures and tools for the creation, maintenance and use of frame populations.

2.6 On frames, target populations, frame populations and survey populations

According to the Memobust glossary: ‘Population is the total membership or population or “universe” of a defined class of people, objects or events. There are two types of population, viz, target population and survey population. A target population is the population outlined in the survey objects about which information is to be sought and a survey population is the population from which information can be obtained in the survey. The target population is also known as the scope of the survey and the survey population is also known as the coverage of the survey. For administrative records the corresponding populations are: the “target” population as defined by the relevant legislation and regulations, and the actual “client population”.’

The Outward FATS frame population of reporting units for statistical reference year T consists of ‘Ultimate Controlling Institutional units’ (UCIs) resident in the EU, as registered in the EGR and referring to 31 December of year T.

The Inward FATS frame population for statistical reference year T consists of foreign controlled enterprises active in reference year resident in the EU, classified in section B to N and PQRS of NACE rev.2 and as registered in the EGR.

2.7 Frames and frame populations in a multi-user environment

A statistical system consists of different statistical activities aimed at describing (a part of) the same target population (see also the module “Statistical Registers and Frames – Survey Frames for Business Surveys”). To allow integration and to secure coherence and consistency of statistical frame populations should be shared, not only on the country level but also on, e.g., EU level.

Part of a frame population methodology is agreement on the *frame* in which the creation and dissemination of frame populations among different users/statistical activities takes place. Generally statistical business registers are appointed performing the function of *frame*.

The role of national business registers should be enhanced as a basic infrastructure element where the national statistical authorities should identify and maintain the statistical units for business statistics and should be used as a source of information for the statistical analysis of the business population and its demography, for the definition of population frames of surveys and for establishing the link to administrative data.

Draft regulation on European business statistics (FRIBS) – version 15 May 2012

Statistics on globalisation or even statistics in a globalised world require adding a supra national dimension to the frame population methodology, which means not only that a supra national *frame* is needed but also harmonisation of national frame population methodologies and additional rules, procedures and tools.

The EuroGroupsRegister (EGR) aims becoming the supra national frame for statistics on globalisation among which FATS statistics. The EGR is integrating data from relevant sources with the objective to compile frame populations for statistics on globalisation, building a European statistical business register of multinational enterprise groups

Having a system of national statistical business registers and a European statistical business register of multinational enterprise groups, unambiguous rules on the roles of and the relationships between these registers are needed.

The role of the EGR register is complementarily to the national statistical business registers.

The complementarity approach means that:

1. The national statistical business registers are responsible for the frame populations of national enterprises facilitating the national integration and coherence of data collected by the national statistical activities. The EGR frame population methodology considers the national business registers as the *authentic store* for national frame populations of enterprises.
2. The EGR is responsible for the population of multinational enterprise groups controlled by ‘*Ultimate Controlling Institutional units*’ (UCIs) and responsible for the links with the national enterprises facilitating the supra national integration and coherence of data collected by the national statistical activities (vertical integration). The EGR considers itself as the *authentic store* for the population of ‘*Ultimate Controlling Institutional units*’ (UCIs) and the attribute ‘*country of the UCI*’ of national enterprises belonging to for the population of multinational enterprise groups.

The EGR as the frame for statistics on globalisation consist of a network of the central EGR register and national statistical business registers.

The complementarity rule implies rules on data flows. Changes on frame populations must be first processed in the *authentic store* before other data stores are updated. This rule is called: ‘**single flow principle**’. For example: a proposal for change of a NACE code of an national enterprise has to be first processed in the national statistical business register before the change reaches the EGR. The other way around: a change in the ‘*country of the UCI*’ of a national enterprise has to be first processed in the EGR before it is applied in the national statistical business register.

The business population is a very dynamic one. Statistical production processes need stable frame/survey populations during a production cycle. Changes in the population during the phase of data collection or during later phases can have serious complicating consequences, methodological as

well as organisational. There exists a high interest in **freezing** survey populations once a statistical production process has started.

This requires rules, procedures and tools dealing with *frame errors*: **frame population error procedure**. Frame errors are mistakes in the frame population due to time lags in information flows, erroneous information, misinterpretation of information etc. Basic rules of this procedure are:

- a) all statistical activities apply agreed rules in dealing with frame errors, e.g., an erroneous NACE code is kept in the statistical outcome except when rule b) is applicable;
- b) frame errors considered as ‘significant for the quality of the statistical output’ are undergoing a process of validation and acceptance. The outcome has to be implemented by all statistical activities involved.

2.8 EGR methodology on FATS frame populations

2.8.1 Outward FATS

The objective of EGR 2.0 is to provide by EU+EFTA country Master Outward FATS frame populations of reporting units (called: Master National Outward FATS frame population of reporting units) for reference year T in April T+1. [Table 1]

Table 1. Descriptive coordinated characteristics of the National Outward FATS frame population of reporting unit's reference year T.

	Characteristic	Explanation
1	Frame reference year	Reference year of the frame population
2	EGR ID of the UCI	A meaningless ID assigned by the EGR system to a legal unit which is defined as Ultimate Controlling Institutional Unit and assigned as reporting unit for Outward FATS to be applied for the period of the frame reference year. In the EGR version 1.0 it is called EU_LEU_ID. This number will stay in EGR version 2.0.
3	Name of the UCI	Legal name of a legal unit which is defined as Ultimate Controlling Institutional Unit.
4	EGR ID of the Global Enterprise Group	A meaningless ID assigned by the EGR system to global enterprise groups. In case of changes in the structure of a group (merging, take-over etc.) the assigned of new ID or the continuation of an ID will be based on the methodology for economic demographic statistics
5	Name of the Global Enterprise Group	The name is included because it is used in the communication between staff.
6	Date in population (month/-year)	The annual population is defined as a volume amount: all groups active in whole or part of the reference year. The date defines from which month the unit has to be included in the frame population.
7	Date out of population (month/-year)	This date defines the last month of inclusion in the frame population

The EGR can and (depending on the user needs) will provide more information, e.g., attributes on global enterprise group, which can be used for stratification of samples or defining thresholds: employment, turnover and assets by reference year and information on links to enterprises and legal units.

2.8.2 Inward FATS

The objective of EGR 2.0 is to provide by EU+EFTA country Master Inward FATS frame population of enterprises (called: Master National Inward FATS frame population of enterprises) for reference year T at the end of March T+2.

The population of enterprises is defined according to the SBS criteria:

- Active in reference year T (= volume population)
- Resident in compiling country
- Classified in B to N and PQRS of NACE rev.2
- Belonging to SBS frame population of compiling country for reference year T.

Table 2. Descriptive coordinated characteristics of the National Inward FATS frame population of enterprises.

	Characteristic	Explanation
1	Frame reference year	Reference year of the frame population
2	EGR ID of the Enterprise	The ID is needed in electronic data exchange.
3	NSA ID of the Enterprise	ID assigned by a NSA to an Enterprise
4	Name of the Enterprise	The name is included because of its use in the communication between users.
5	Date in population (month/-year)	The annual population is defined as a volume amount: all enterprises active in whole or part of the reference year. The date defines from which month the unit has to be included in the frame population.
6	Date out of population (month/-year)	This date defines the last month of inclusion in the frame population
7	EGR ID of the Global Enterprise Group	A meaningless ID assigned by the EGR system to global enterprise groups. In case of changes in the structure of a group (merging, take-over etc.) the assigned of new ID or the continuation of an ID will be based on the methodology for economic demographic statistics
8	Country of the UCI	Country of the Ultimate Controlling Institutional unit defined in accordance with the FATS Recommendations Manual
9	NACE code	According to NACE Rev 2
10	Institutional sector code	According to ESA2010
11	Size class	Employment classes according to SME definition, used in SBS

The EGR can and will provide more information, e.g., address state of activity, employment, turnover, link to global enterprise group and links to legal units and units in administrative registers. This kind of information will be the most topical information available. For example: updates after March T+2

are possible for attributes like employment, whether the enterprise was active in the reference year or turnover.

The critical milestones are April T+1 for Outward FATS and March T+2 for Inward FATS. The disseminated frame populations of those dates are called: 'Master frame populations'. The 'master' versions contain the content serving the coordination of statistics and should be used as the only reference on statistical units and their characteristics in the production of statistical output.

Final disseminated frame populations are the product of data collection, data processing and data analysis (including validation)². To produce a final population two or more iterations are needed. Final versions of frame populations are called Master version of frame populations. The frame populations produced in the iterations before the production of a master population are called initial and intermediate versions of frame populations. The initial and intermediate frame populations serve data quality management (see next section).

Regarding the National Outward FATS Frame population of reporting units an initial version for reference year T will be provided by the EGR in September year T. An intermediate version will be provided in January/February year T+1.

Regarding the National Inward FATS Frame population of enterprises an initial version for reference year T will be provided together with the Master Outward Frame population of reporting units in April T+1. At least one intermediate version will be provided in the period September T+1/March T+2 for validation after which a master version is published.

The content of a master frame population (as defined in Tables 1 and 2) will be in principle 'frozen'. Changes in the content of units and characteristics as described in Tables 1 and 2 give serious complications in statistical production process. For example a change in a the UCI and its residency implies a change in a national Outward FATS frame population which implies a change in the survey populations of compiling countries/statistical authorities (e.g., sending a new questionnaire, deleting data, grossing-up procedures etc.). The frame population error procedure provides rules for dealing with changes after the dissemination of Master frame populations. Accepted changes will be processed in the Master frame populations.

2.9 EGR benefits

Official European statistics should be credible. One of the requirements is that statistical figures published by Eurostat and Member States are consistent: '*tell the same*'. The example below shows that this is not always the case.

Outward FATS of country A produces the following figure: multinational enterprise groups controlled by a resident legal entities in country A are controlling 1101 enterprises resident in country B. Inward FATS of country B should produce the following figure: 1101 enterprises resident in country B are controlled by a legal entity in country A. However country B publishes 505 enterprises, a difference of 596 enterprises.

² ESSnet EGR is using the GSBPM, version 4.0 – April 2009 as standard terminology for business processes needed to produce EGR output.

An EGR based on a EGR frame population methodology offers a solution provided that the MS strictly follow the rules and procedures agreed. Critical parts of these rules are:

1. FATS statisticians of country A and B accepts the (country of) UCI as registered in the EGR.
2. FATS statisticians of country A and B accept the EGR population of enterprises in country B.

Moreover the EGR contains identifying information on enterprises in country B. As country B is responsible for data collection on enterprises producing SBS statistics the EGR offers the opportunity for sharing data collected which not only contributes to consistency but also reduces response burden (country does not need to collect data on enterprises in country B anymore).

FATS statistics is one statistical activity which benefits from the EGR. Other statistics like Foreign Direct Investment (FDI) and foreign trade statistics can gain quality by using the EGR as provider of coordinated frame populations.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Eurostat (2012), *Foreign AffiliaTes Statistics (FATS) recommendation manual*, version 2012.

ESSnet EuroGroups Register (2013), FATS frame population, ESSnet EGR view, version 1.3.
<http://egr.istat.it/>

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
2. Statistical Registers and Frames – Survey Frames for Business Surveys
3. Statistical Registers and Frames – The Statistical Units and the Business Register
4. Derivation of Statistical Units – Derivation of Statistical Units

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

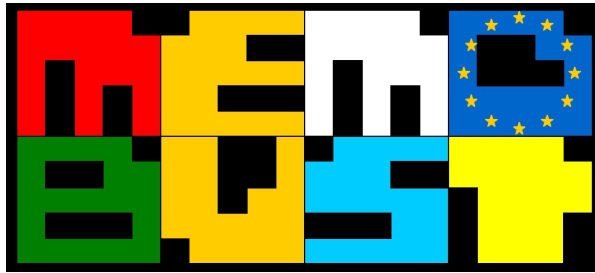
Macro-Integration-T-Asymmetry in Statistics - EGR

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-04-2013	initial version	Harrie van der Ven	CBS
0.2	30-10-2013	review results processed	Harrie van der Ven	CBS
0.3	14-03-2014	EB review processed	Harrie van der Ven	CBS
0.3.1	14-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:26



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Seasonal Adjustment – Introduction and General Description

Contents

General section.....	3
1. Summary	3
2. General description.....	4
2.1 Main objectives and general description	4
2.2 Benefits and costs	9
2.3 Users' perspective	14
3. Design issues	17
4. Available software tools.....	17
5. Decision tree of methods	17
6. Glossary.....	18
7. References	18
Interconnections with other modules.....	20
Administrative section.....	21

General section

1. Summary

A common practice nowadays for a National Statistical Institute (NSI), when dealing with systems of time series collected on sub-annual basis, is to perform seasonal adjustment (SA) in order to help users to interpret published statistics. By separating the non-seasonal part from the seasonal and calendar effects a user is likely to obtain a refined picture about the underlying movement from the time series observations. Hence, the SA-procedure eliminates the estimated seasonal and calendar effects from the original time series and obtain the SA estimates. Such estimates are likely to reveal what is new in a time series, which is a crucial issue related to seasonal adjustment. Hence, SA may be viewed as an aid in decision making, usually used for comparisons between different regular periods in time (month-to-month, quarter-to-quarter, etc.) but also for forecasting purposes and for model-building. For example, SA of macroeconomic indicators is useful for policy makers and other users because of the need for understanding repetitive fluctuations in economic activity (business-cycles) as well as the short-term and the long-term movements in time series. These effects are in a SA-procedure regularly expressed in terms of a unified trend-cycle component (see, e.g., Statistics Canada, 2009; ABS, 2008).

Since SA is a modelling procedure which transforms the original data in order to obtain the estimates a natural question is how reliable these estimates are. Further issues usually associated with SA are reliability and quality with respect to benefits and costs associated with the procedure in question. Some other issues, such as revisions, outlier treatment, aggregation and data presentation are also common to different domains of statistical production which necessitates standardised, coherent and consistent treatment of SA-procedures.

A NSI should also take care about the needs of both the internal and external users, which typically implies shifting focus from a pure methodological aspect to some other (perceived) quality aspects. Balancing between these two aspects is recommended since statistics should be of high quality but also easily interpretable for users.

A vast and very detailed literature about issues related to SA is already available to the public. See, e.g., IMF (2001), ECB (2003), Dagum and Cholette (2006), European Communities (2001) etc. The websites of some prominent statistical offices and developers of statistical software offer detailed information about the related procedures (e.g., Statistics Canada, 2009; ABS, 2008; U.S. Census Bureau, 2012; Bank of Spain, 2012; Koopman and Lee, 2010). Also, the European Statistical System (ESS) developed a set of guidelines on seasonal adjustment (Eurostat, 2009) and the new software Demetra+ (Eurostat, 2012). Although the ESS Guidelines provide a set of recommendations for the best practices, this document by its nature does not give a comprehensive introduction to seasonal adjustment for the non-specialists and typical users at a NSI.

Hence, in this and related modules the main focus is put on aggregating information about SA from different sources and experiences in order to assist the users at NSIs with relevant easy-to-read information and references to the more detailed technical and methodological description.

2. General description

2.1 Main objectives and general description

Intra-annual (monthly and quarterly) macroeconomic indicators represent a key tool for several people: policy makers, business managers, journalists, economists, statisticians, etc. Most of such indicators exhibit a (dominant) seasonal pattern obscuring and dwarfing other components of greater economic relevance to understand the economic phenomena. As a consequence, the seasonal fluctuations should be filtered out through the *seasonal adjustment*, a technique aimed at estimating the seasonal component and removing it from the observed time series. Figure 1 shows two examples of seasonal series (raw or unadjusted series) together with their corresponding SA series: the Italian industrial production index and the Italian labour force. Although both series are seasonal, the graphs reveal their different features. In particular industrial production shows more regular and larger seasonal fluctuations than labour force.

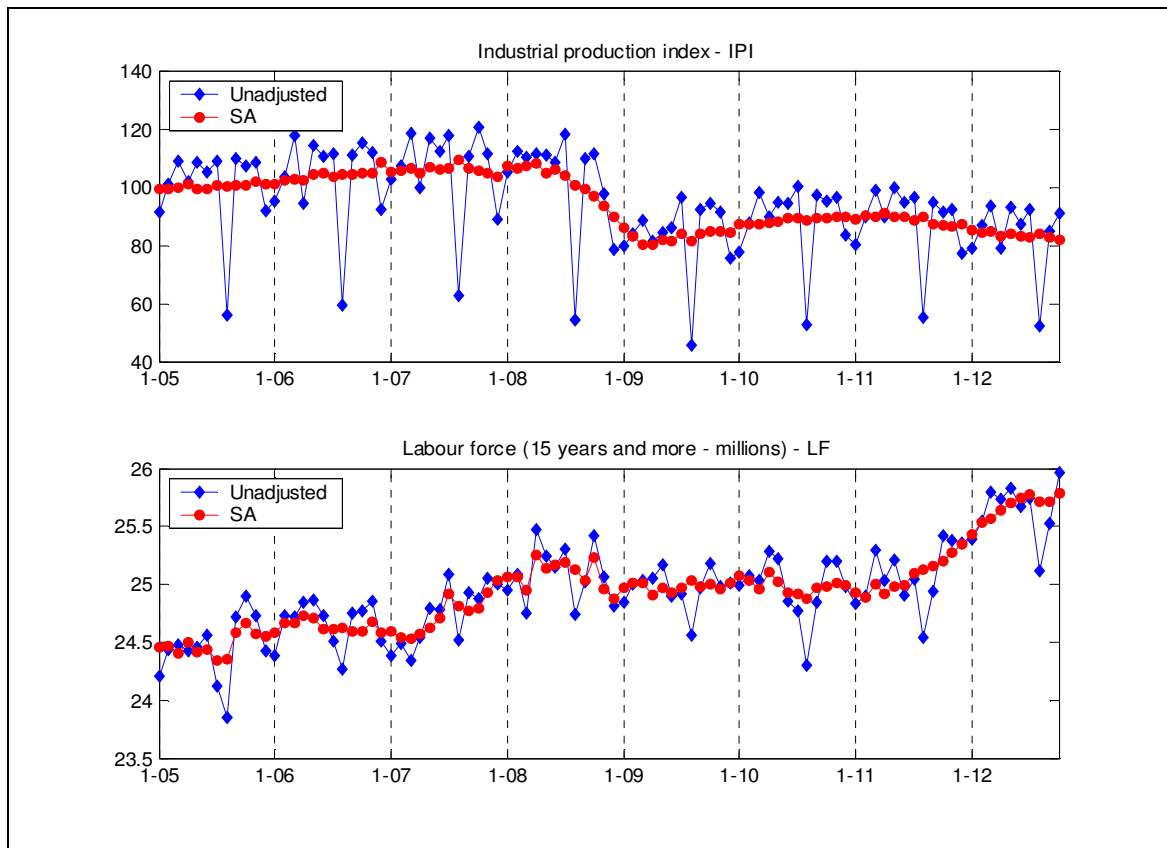


Figure 1: Italian industrial production index and labour force. Unadjusted and seasonally adjusted data (by Tramo-Seats for Linux).

From an analytical point of view seasonally adjusting a time series, X_t , means to decompose it in four different components:

1. trend component T_t that shows the long-term tendency;
2. seasonal component S_t that represents intra-year fluctuations which recur every year to the same extent (short-term regular variations);

3. cyclical component C_t that indicates the medium and long term fluctuations containing the long-term irregular variations. The cyclical component is worth examining only in case of very long time series. As a general practice we assume that it is included in the trend component that it is referred to as cycle-trend;
4. irregular component I_t that contains the random effect that we cannot predict.

Depending on the relations among these components, different decomposition models can be considered:

- a. the additive model

$$X_t = T_t + S_t + C_t + I_t$$

where the differences between the observed data and the cycle-trend (called seasonal differences) are supposed to be nearly constant in the same periods (months or quarters) of different years;

- b. the multiplicative model

$$X_t = T_t \times S_t \times C_t \times I_t$$

where the ratios between the observed data and the cycle-trend (called seasonal-irregular ratios) are supposed nearly constant in similar periods of different years;

- c. the log-additive model

$$\ln(X_t) = \ln(T_t \times S_t \times C_t \times I_t) = \ln(T_t) + \ln(S_t) + \ln(C_t) + \ln(I_t)$$

that can be used to specify an additive model on the logarithm of the time series.

There are some other decomposition models but these three are the most commonly used. Figure 2 shows two examples of deterministic time series built both summing up and multiplying a linear trend and a deterministic seasonality (both shown in the upper panels). Their sum is displayed in the lower left-hand panel, while their product is represented in the lower right-hand panel: in the former case the seasonal amplitude is constant around the trend, in the latter case the seasonality amplifies with the trend.

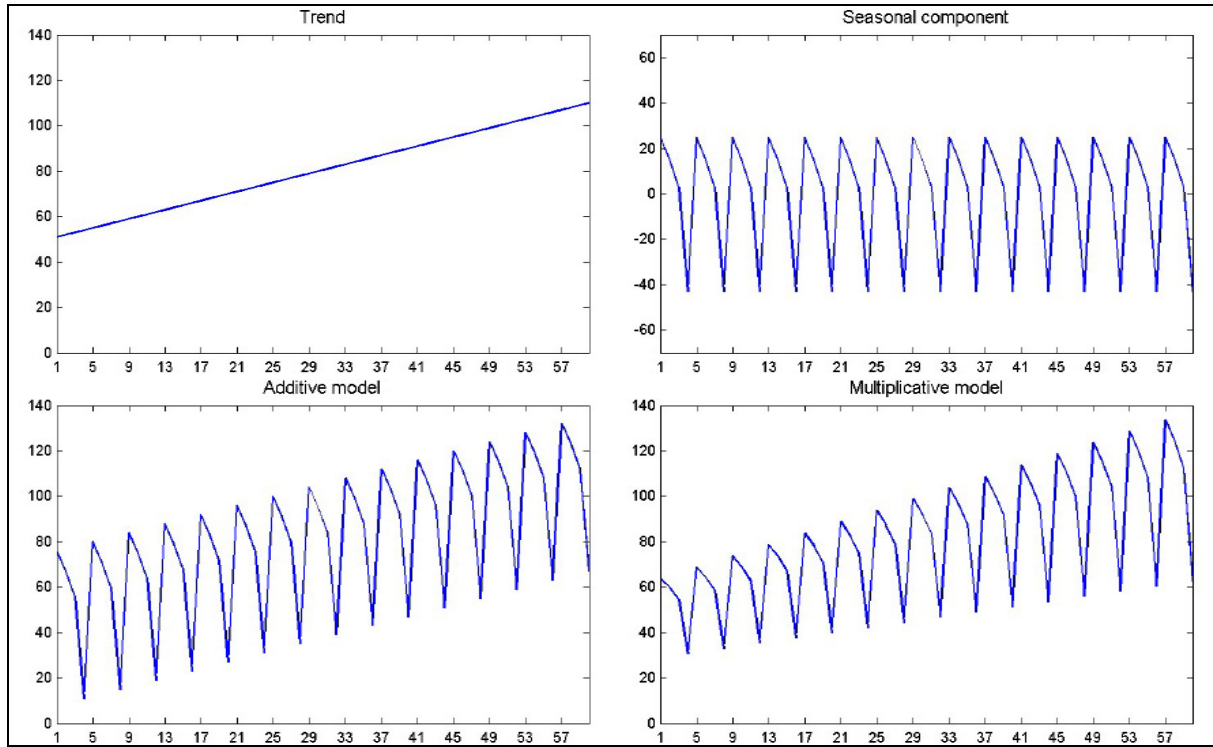


Figure 2: Additive and multiplicative models using a linear trend and a deterministic seasonal component.

According to the decomposition model used, the seasonally adjusted time series A_t , which contain neither calendar effects nor seasonal component, can be formulated in two alternative ways:

$$A_t = X_t - S_t = T_t + C_t + I_t \quad \text{additive model}$$

$$A_t = X_t / S_t = T_t \times C_t \times I_t \quad \text{multiplicative model}$$

It is worth noting that in the additive decomposition additive components have the same scale as the original series and the expected value of the irregular component is 0, while in the multiplicative or log-additive decompositions only the trend (and consequently the SA series) is expressed in the original scale and the expected value of the irregular component is 1.

Time series may be affected by the composition of calendar (*calendar effects*) or may contain atypical observations which do not follow the usual pattern of the time series (*outliers*). The former are always included in the seasonality and therefore removed from the SA series. As far as outliers are concerned, when they are assigned either to the irregular or to the trend, they are visible in the SA series, while when they are assigned to the seasonal component they are removed from the SA series.

Both calendar effects and outliers are generally dealt as deterministic components that are described below.

Calendar effects

The calendar effect component is a part of the time series which represents calendar variations, such as trading/working days, moving holidays and other calendar-related systematic effects that occur not the same way from year to year.

a) Trading/working day effect

Although the trading day and working day could be distinguished, we will use these as synonyms. Since the number of trading days may be different both in consecutive periods and in the same period of different years, it cannot be managed as an ordinary seasonal effect.

b) Holiday effect

The number of working days depends also on the holidays, which do not fall on weekends. As the national holidays vary from nation to nation, it is recommended to consider national calendar including national holidays to build country specific regression variables (or regressors), avoiding the use of standard regressors.

c) Easter effect and other moving holiday effect

There are some holidays which do not fall on a fix date. For example, Easter may be either in March or in April. Moreover, Easter may have one-week or more time duration before and/or after Sunday.

d) Bridging effect

Bridging days are days lying between a public holiday and a weekend. They are counted in purely calendar terms as full working days, but because of their particular date, they could be considered as holidays to offset overtime already worked or for long weekends.

e) Leap-year effect

There is an additional day in every four year which may affect the time series.

Figure 3 represents trading-day, leap-year and Easter effects drawn from an additive decomposition. The overall calendar effect is the sum of the represented effects, derived multiplying the regressors by their respective parameters of the regression model estimated on the unadjusted time series.

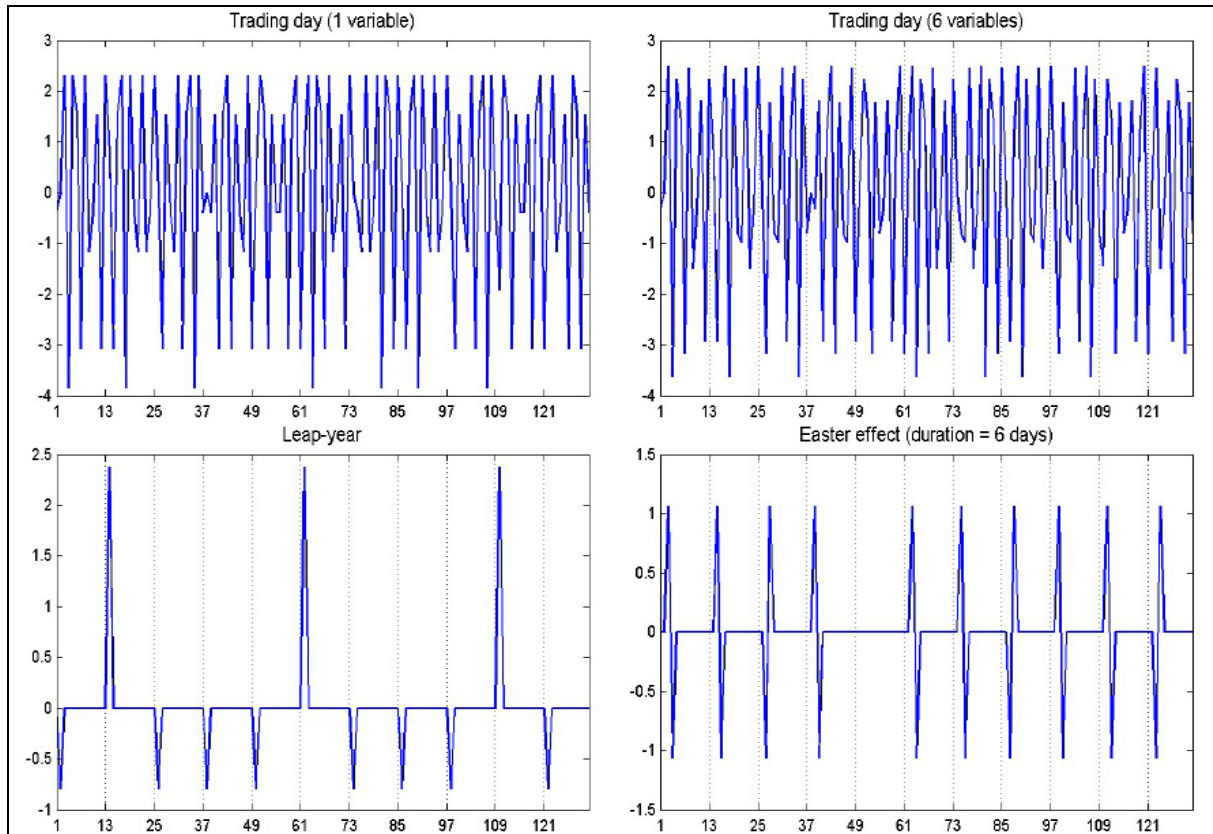


Figure 3: A representation of trading-day, leap-year and Easter effects in an additive decomposition.

Outliers

Outliers are data which differ greatly from the tendency. Typically these are caused by a one-off economic or social event. The most known type of outliers are:

1. the additive outlier which influences only one observation (it is included in the irregular component);
2. the transitory change that affects several observations, but reduces gradually (exponentially) until the time series returns to the initial level (it is included in the irregular component);
3. the level shift, which represents a step, that is a permanent change in the time series level (it is included in the trend component).

The left-hand panels of figure 4 represent the above outliers. Their effects on the time series are highlighted in the right-hand panels through red lines.

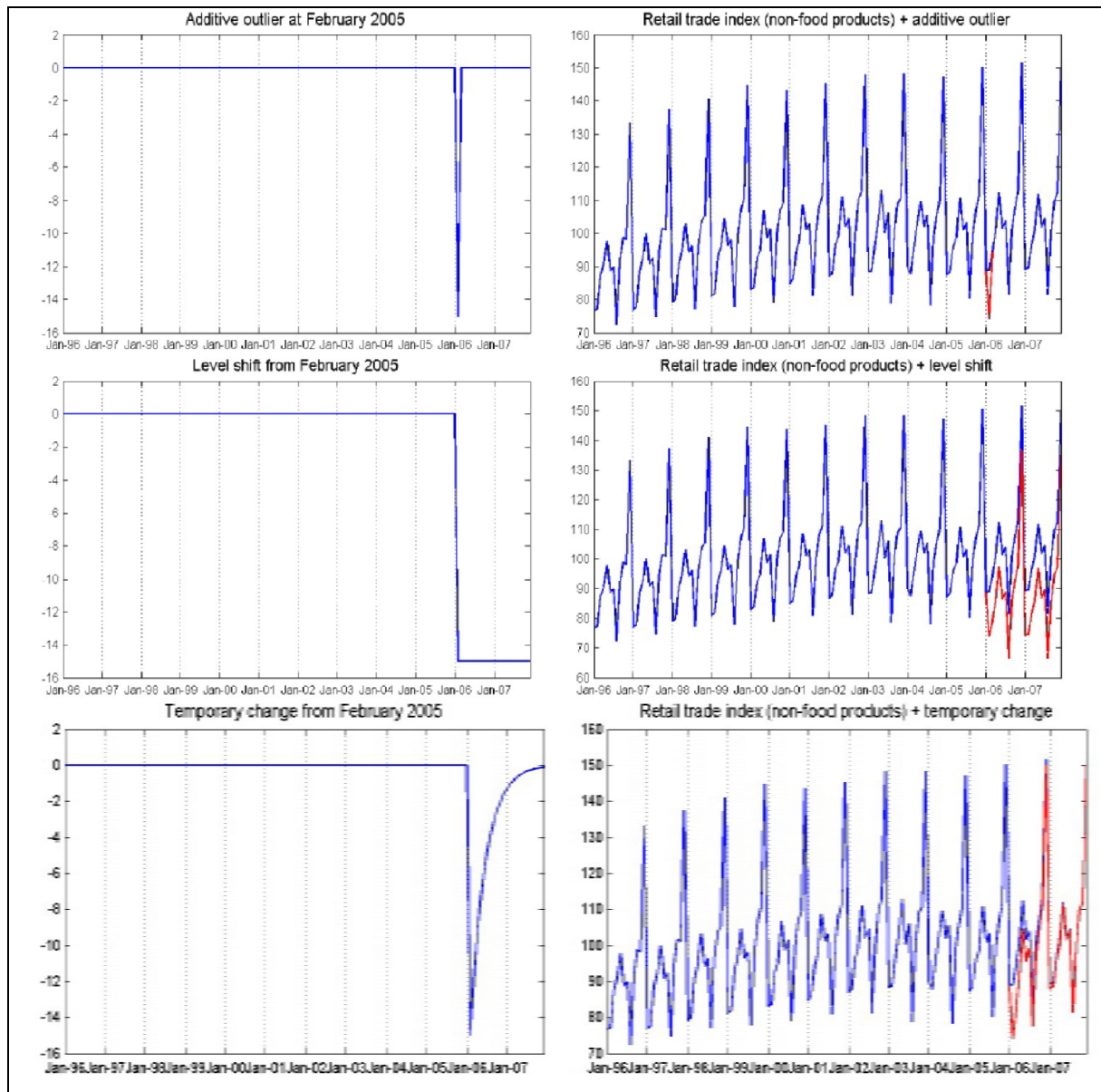


Figure 4: A representation of additive outlier, level shift and transitory change.

There are also other less known types of outliers which can be detected and treated: ramp outlier, innovational outlier and seasonal outlier.

More details on calendar effects and outliers can be found in the module “Seasonal Adjustment – Seasonal Adjustment of Economic Time Series”.

2.2 Benefits and costs

From the previous section it has been seen that the main aim of seasonal adjustment is to filter out systematic seasonal fluctuations from time series, due to the noneconomic causes such as weather, calendar events and timing decisions. Generally SA data are preferred to unadjusted data since they are more easily interpreted because of their comparability between adjacent periods. Figure 3 confirms that, displaying month-on-month (m-o-m) growth rates calculated on both the unadjusted and the SA series of figure 1. Moreover, the two panels on the right-hand side, where m-o-m growth rates

calculated on SA are represented, stress a typical feature of SA data: their volatile profile due to the presence of the irregular (and unpredictable) component overlapping the more smoothed trend-cycle component.

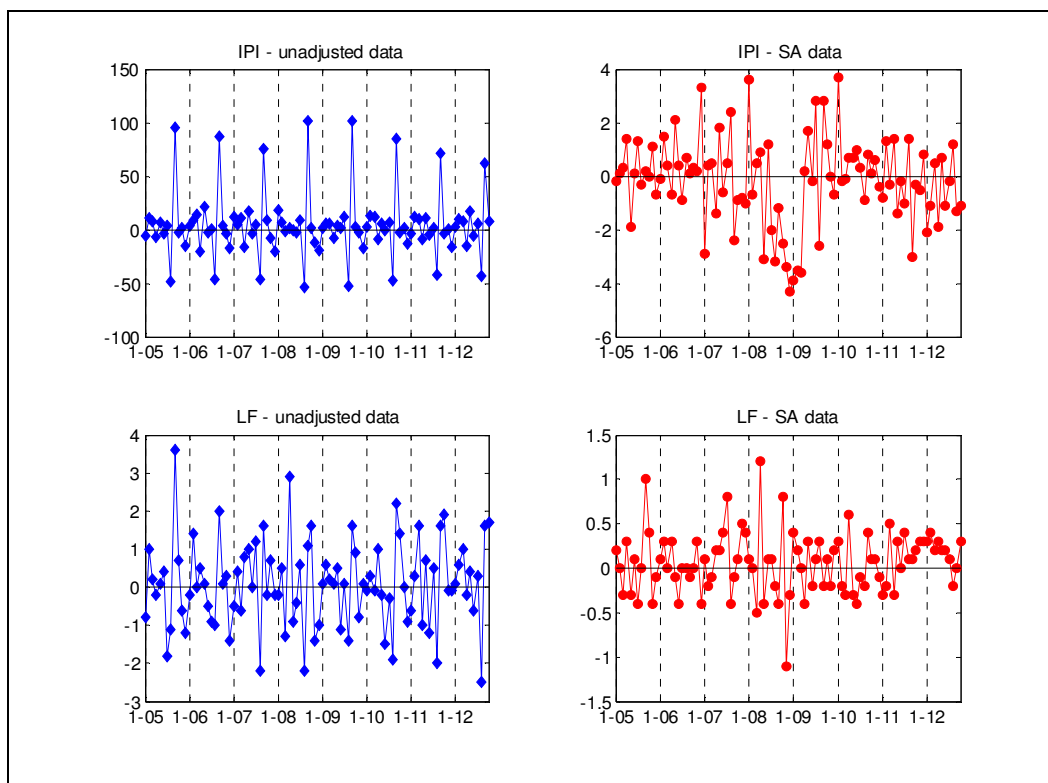


Figure 5: Month-on-month growth rates of Italian industrial production and labour force, calculated on unadjusted and SA data.

One simple way of removing seasonal fluctuations and understanding the recent movement of economic indicators is to calculate year-on-year (y-o-y) growth rates (i.e., applying seasonal differences on the log-transformed data) on the unadjusted data. However, inaccurate conclusions could be drawn utilising y-o-y growth rates on unadjusted data: firstly, time series do not show *regular* seasonal fluctuation, on the contrary they are often featured by a *moving* or an *evolutive* seasonality; secondly, y-o-y growth rates depend on the dynamic of two consecutive years and turning points in the data are shown up with some delay.¹ Seasonal adjustment allows to overcome both of these drawbacks: the first one is very intuitive, the second one needs some further details. To this end the example of the Italian index of industrial production is considered, both in calendar adjusted (i.e., unadjusted data with calendar effects removed) and SA form. Moreover, it is quarterly aggregated in order to have a smooth time series. Calendar adjusted data and SA data are presented in the upper

¹ It is worth noting that here the focus is not put on the debate concerning the calculation of y-o-y growth rates either on unadjusted data or on SA data. At this regards, useful references can be found on the handbook on data and metadata reporting and presentation (OECD, 2007). On the contrary, y-o-y growth rates on unadjusted data are presented as a very simple tool used to remove seasonality and to read the recent movement of economic indicators.

panel of figure 6, while the respective y-o-y and q-o-q growth rates are displayed in the lower panel. The main message conveyed by the latter is the two quarter delay in detecting the turning point of industrial production through the y-o-y growth rates, both at the third quarter 2009, the beginning of the expansion phase (highlighted through the grey area in the first panel), and at the second quarter 2011, that is the beginning of the new recession phase.

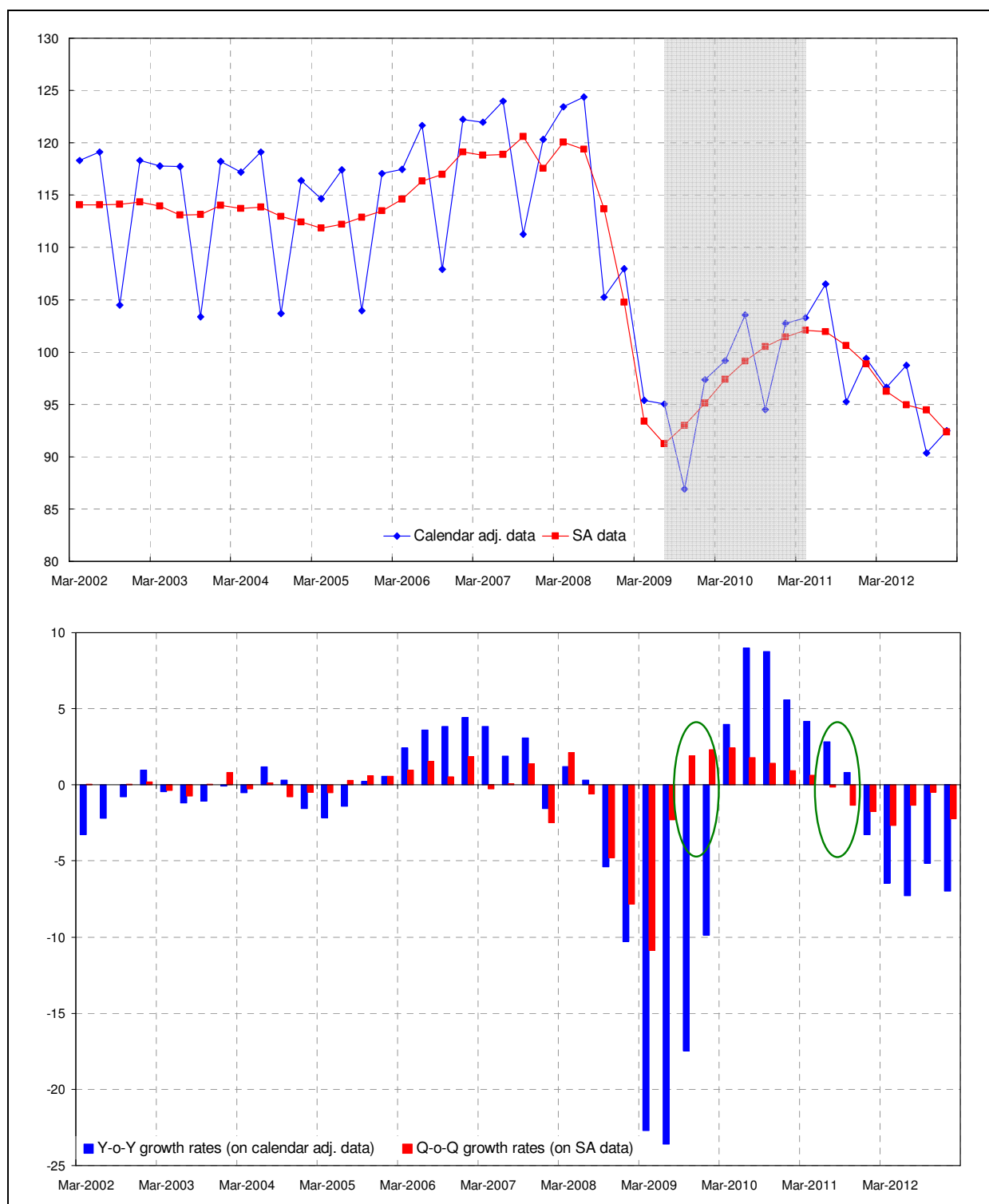


Figure 6: Italian industrial production index, quarterly aggregated.

It should be stressed that, although a time series is seasonally adjusted, unadjusted data remain useful. They represent the base to understand particular phenomena or events (introduction of a new classification, change of base year, introduction of new statistical methods, strikes, introduction of a new tax, ...) and to take them into account in the model of seasonal adjustment.

As already sketched in the previous section, seasonal adjustment also includes the elimination of calendar effects. Data adjusted for calendar effects, generally achieved as a by-product of the seasonal adjustment, are often required by European regulations and reported in the press releases (useful information on reporting unadjusted, calendar adjusted and seasonally adjusted data, together with the corresponding growth rates can be found in OECD (2007)). Table 1 contains unadjusted and calendar adjusted index of industrial production, together with the number of working days (Monday to Friday) net of Italian holidays falling in working days and the y-o-y growth rates calculated on both unadjusted and calendar adjusted data. It can be seen that the calendar adjustment affects y-o-y growth rates only when the compared months have a different number of working days (light blue and yellow rows of the table).

Table 1: Italian industrial production index. Unadjusted and calendar adjusted data, number of working days and y-o-y growth rates.

Period	Unadjusted		Calendar adjusted		Working days		y-o-y (%)	
	2010	2011	2010	2011	2010	2011	Undjusted	Cal. adj.
Jan	77.8	80.4	81.7	81.9	19	20	3.3	0.2
Feb	87.6	89.8	88.0	90.2	20	20	2.5	2.5
Mar	98.1	99.0	94.7	97.7	23	22	0.9	3.2
Apr	89.8	89.8	86.7	90.1	21	20	0.0	3.9
May	94.6	99.6	95.6	97.6	21	22	5.3	2.1
Jun	94.4	94.8	93.3	93.7	21	21	0.4	0.4
Jul	100.4	96.3	100.6	99.4	22	21	-4.1	-1.2
Aug	52.7	55.2	51.6	54.1	22	22	4.7	4.8
Sep	97.4	94.8	95.4	92.9	22	22	-2.7	-2.6
Oct	95.1	91.5	98.2	94.5	21	21	-3.8	-3.8
Nov	96.3	92.4	95.2	91.3	21	21	-4.0	-4.1
Dec	83.4	77.1	78.3	76.9	22	20	-7.6	-1.8
Year	89.0	88.4	88.3	88.4	255	252	-0.7	-0.1

When the observed phenomenon depends on the number of worked days of each month (quarter), calendar effects have to be estimated and removed in order to improve both temporal comparisons and quality of seasonal adjustment. Calendar adjustment is part of the pre-treatment of the series performed before the decomposition and the seasonal adjustment.

Generally NSI and other official producers of SA data expend many efforts to produce carefully SA data and to make them available to the general public for several further purposes (modelling and forecasting, trend-cycle decomposition, turning points detection, business cycle analysis, ...). This is due to several reasons.

- a) A precise and rigorous definition of seasonality does not exist and, consequently, several methods and procedures have been developed to deal with seasonal macroeconomic indicators. Moreover, different procedures or different models/options, within the same procedure, give almost always different seasonally adjusted data.

- b) Due to the specific *filters* used to remove the seasonal component (i.e., two-sided moving averages), when new unadjusted data become available, the seasonal adjustment performed on the longer series revises the seasonally adjusted data previously released, especially at the end of the series. Figure 7 shows the seasonally adjusted data of the Italian industrial production index released from July 2012 to December 2012.

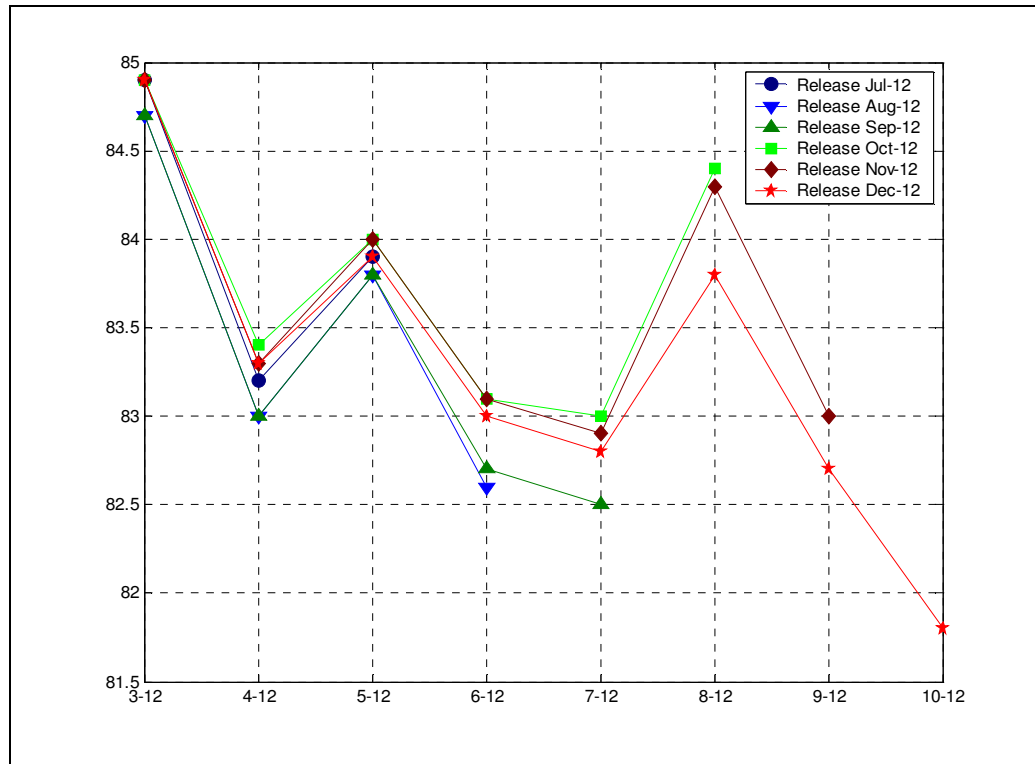


Figure 7: Italian industrial production index. Last releases of seasonally adjusted data.

- c) Once a method and a procedure were chosen to approach the SA of a new domain, there are several issues that have to be dealt with:
- the number of time series to be treated (when not regulated by the European regulations): the whole domain or the most relevant indicators. Generally indicators at more disaggregated breakdown are more irregular and volatile and, therefore, more difficult to treat adequately;
 - the choice between direct and indirect approach to seasonally adjust vertical/horizontal aggregates. In fact they may be seasonally adjusted through a seasonal adjustment procedure (direct approach) or aggregating the seasonally adjusted disaggregated components (indirect approach). Coherence is fulfilled by the latter, but problems of residual seasonality in the aggregates may arise;
 - the treatment of outliers (breaks, unusual movements, extraordinary events) especially at the end of time series;
 - the presentation of seasonally adjusted data and the respective metadata (procedure, model options, ...).

- d) Events affecting contemporaneously the domains of short-term business statistics, quarterly national accounts, business surveys, ... (e.g., the 2008-2009 crisis) should require consistent treatments and solutions in order to release seasonally adjusted data of good quality, to avoid misleading results that may confuse users and undermine the credibility of the producer of seasonally adjusted data.

Dealing with all these issues is time and human resources consuming and requires a good knowledge of unadjusted data. As a consequence it is recommended that the statisticians who compile the indicators should also be in charge of the seasonal adjustment, with the assistance of specialists to handle important and/or crucial indicators.

2.3 *Users' perspective*

One of the main goals of the SA-procedures is to produce time series data which can relatively easily be interpreted by the users. Hence, the transparency should be a crucial issue for a NSI aiming to fulfil the users' needs. The term *user* might be interpreted in different ways in different offices. Thus, some distinction has to be made in this context to clarify what kind of users this module is considered with.

In the context of official statistics, the users of SA data and related procedures might be divided into the two main categories: internal users and external users. The internal users are usually employees in a NSI with certain level of responsibility in a process of production of business statistics but they are usually not specialists in SA. This module is mainly oriented to these users.

The external users, on the other hand, are typically researchers, business analysts, journalists or governmental policy makers. Common to all these users is that they are not involved in a process of production of seasonally adjusted data in official statistics. Some of the external users might have more or less influence on NSI with respect to certain quality aspects of seasonally adjusted data. However, the point of view of these external users will not be treated in this handbook except for very general remarks.

General to all these users is that they expect good quality of seasonally adjusted estimates. However, what some users may experience as a good quality might be different from the quality interpretation from the point of view of a specialist. Furthermore, different users put attention to certain aspects of data while others are more interested in economic interpretation of the results. Here, we summarise some of the most relevant issues related to the "perceived" quality and put those issues in a context of a NSI. The main idea is to bridge gaps between these two concepts by recommending strategies for satisfying the main users' needs while keeping *statistical* quality at a satisfactory level. More general aspects of statistical quality are discussed in some other modules in this handbook. Here, we focus on the specific topics concerned with seasonal adjustment and related issues.

The users of seasonally adjusted series are typically interested in the following issues:

- e) Interpretability of seasonally adjusted time series:
 - Economic or other relevant interpretation.
 - Outlier interpretation.
- f) Different aspects of consistency and coherence.
- g) End-point analysis for forecasting purposes.
- h) Business-cycle analysis and early detection of turning-points.

- i) Revisions: nature of revisions- distinction between the natural source of revision (from the original unadjusted series) and the part of the revision due to specific SA-procedure.

2.3.1 *Interpretability*

Seasonally adjusted time series are expected to reflect basic properties of the original time series. Hence, the users would like to have seasonally adjusted estimates that reveal the “true” development in an economic variable. An informed user may have an *internal* knowledge about the “true” properties of a certain time series variable. This knowledge (read: information) may imply the expected future development in a specific direction. Furthermore, this user might have strong reason to believe that the future development of the corresponding seasonally adjusted estimates would follow the expectations and preferably lie within an expected (prediction) interval. Any strong divergence from these expectations would imply questioning the quality of SA.

From a point of view of a producer of official statistics, this kind of situation would lead to a thorough investigation about the source of such a deviation from the expected results. Sometimes, the estimation process within a seasonal adjustment procedure produces results that are not wrong from the statistical point of view but the same results might be interpreted as erroneous by the users. It is well-known that SA might induce some spurious marginal effects on a seasonally adjusted estimate, especially at the end of time series.

A transparent communication between specialists, internal users and external users is important in any case. The recommended action for a producer is to make an attempt to meet the external users’ requirements if this would not result in a significant departure from the quality requirements. Otherwise, if there is no possibility to make any changes, it is important to clarify the actual cause of the discrepancy between the expected results and the actual outcome. Also, motivation for further actions has to be understandable from the users’ point of view.

2.3.2 *Consistency and coherence*

The term consistency is usually related to the users’ needs for internal coherence within a system of seasonally adjusted time series where some pre-defined inter-relationships have to be preserved. The nature of these relationships naturally originates from the raw data. The systems of time series are usually classified by certain attributes based on, e.g., artificial accounting constraints as in the systems of national accounts or by trade-group classifications as in the retail trade. In addition, there are natural classifications due to the different categories such as gender, regions or provinces, part-time or full-time employment in the system of labour force series. All sub-categories must add up to the marginal totals which in turn aggregates to the grand totals. These constraints are called the *cross-sectional aggregation* constraints meaning that all the original relationships between different categories are preserved for each period of time.

Quite often there are also *temporal constraints* that some or all series in a system are required to satisfy. This means that the seasonally adjusted yearly totals created from the aggregated higher frequency (monthly or quarterly) seasonally adjusted series, must add up to the corresponding annual benchmark. This annual benchmark is usually formed as a yearly total from the higher frequency unadjusted series.

The users would like to have consistencies in all directions in such systems of seasonally adjusted time series. This is important because of the necessity to explain the results in an easy way when communicating with external users and the public. Thus, in order to satisfy consistency restrictions the experts usually implement some kind of *reconciliation* or *benchmarking* technique. These techniques are mathematical tools to achieve aggregation consistency (summability) and the temporal consistency (benchmarking). See, e.g., Dagum and Cholette (2006) for more details about the reconciliation and benchmarking issues.

In some cases the users are interested in consistency in terms of coherence which is not as straightforward as aggregation. Coherence might be loosely interpreted as a more general form of relative correspondence between a set of mutually connected time series from different sources. This requirement is often more difficult to satisfy than consistency within a system of time series from one source. The internal users have to be aware of limitations of the SA methods and hence be able to explain nature of inconsistencies to the external users.

2.3.3 *End-point analysis and forecasting*

Seasonally adjusted time series are generally used by economic agents, policy makers, researchers and others who are interested in extracting information from the data. Usually, the main interest is in short-term prediction because of the nature of seasonally adjusted time series. Seasonal effects have impact on the higher frequency time series (within a year, usually monthly or quarterly frequency). This implies that the growth rates of seasonally adjusted estimates from one year to another should coincide with the corresponding growth rates from the original time series. Any discrepancy should be addressed to changes in seasonal variation from year to year (*moving seasonality*) or to the issues related to modelling.

Consequently, the users are likely to prefer stable seasonal patterns in order to extract other relevant information which is generally not stable and hence not repetitive. The main idea is to use seasonally adjusted data to reveal the news in time series in order to understand “where we are now” and “where we are going to”. According to ESS Guidelines (Mazzi and Calizzani, 2009, p. 6), this is the ultimate goal of SA. The extraction of the news in time series is closely connected to short-term forecasting purposes.

However, the most common SA-procedures are by default very sensitive to any instability at the end points. As a consequence, they actually fail to make reliable forecasts unless very strong assumptions are fulfilled. For example, the presence of outliers at the end of time series is likely to magnify uncertainty in the forecasts. However, there is no reliable statistical technique to identify presence and nature of these outliers. See, e.g., IMF (2001, p. 135) for a discussion about end-point problems.

Hence, the users of SA have to have enough knowledge about the end-point problems in order to communicate with the users of official statistics.

2.3.4 *Business cycle analysis and detection of turning points*

This issue is closely connected to the previous topic and the concept of “news”. Changes in trend or in the business cycle are related to the needs for understanding the past and the future. For example, the economic agents and policy makers try to explain effects of some political measures in the past using the trend-cycle analysis. They are interested in timing of business-cycles, i.e., in locating periods of

recessions and expansions in economy. They are indeed focused on identifying the turning points in the future too in order to modify course of future actions.

Hence, the users who run business statistics in a NSI would like to understand the nature of results from a SA-procedure in order to be able to explain their effects when necessary. In the case when a deep technical explanation is needed it is advisable to contact a SA expert.

2.3.5 Revisions

Small revisions in seasonally adjusted time series are of particular interest to the users. Large revisions usually imply questioning the data quality and relevance of the chosen seasonal adjustment methods. For an official statistics producer it is important to put effort to identify sources of revisions. In some cases revisions may be reduced by certain strategies which focus on increasing stability of the estimates. This is especially important if revisions do not originate from the actual revisions in published raw data. This issue is rather technical and requires a good communication between the users and the experts.

3. Design issues

4. Available software tools

The users of official statistics usually prefer a software tool which is user-friendly and stable. Furthermore, the use of some of the conventional and internationally accepted software tools and methods is typically favoured by the users. Recommendations from Eurostat are also important to the users since the NSIs usually distribute some seasonally adjusted data to this statistical institution.

In the ESS Guidelines for Seasonal Adjustment (Eurostat, 2009) two main (A-classified) methods are proposed, TRAMO-SEATS (Gomez and Maravall, 1996, 2001a, 2001b), X-12-ARIMA/X-13-ARIMA-SEATS (USCB 2011, 2012a, 2012b, 2013). The National Bank of Belgium in cooperation with Eurostat has recently developed two open source software platforms Demetra+ and JDemetra+. These platforms are based on the two leading algorithms mentioned above, TRAMO-SEATS and X-13-ARIMA-SEATS. Demetra+ and JDemetra+ are freely downloadable from the Eurostat's home page (Eurostat, 2012).

The Structural Time Series Models (B-classified) are also recommended as an alternative to the previous two methods, under certain conditions. Hence, the users are likely to favour one of the proposed methods, typically TRAMO-SEATS or X-12-ARIMA.

Some other preferences, e.g., those related to the chosen software platform, might occur depending on the available IT-architecture or specific subject-matter issues, which may vary from one NSI to another.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- ABS (2008), *Time Series Analysis: Seasonal Adjustment Methods*. Australian Bureau of Statistics. Available from:
<http://www.abs.gov.au/websitedbs/d3310114.nsf/51c9a3d36edfd0dfca256acb00118404/c890aa8e65957397ca256ce10018c9d8!opendocument/> [Accessed 20 March 2013]
- Bell, W. R. and Hillmer, S. C. (1984), Issues Involved with the Seasonal Adjustment of Economic Time Series. *Journal of Business and Economic Statistics* **4**, 291–320.
- Bank of Spain (2012), *Statistical and Econometrics Software*. Banco de España. Available from:
http://www.bde.es/bde/en/secciones/servicios/Profesionales/Programas_estadi/Programas_estad_d9fa7f3710fd821.html/ [Accessed 20 March 2013]
- Dagum, E. B. and Cholette, P. A. (2006), *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*. Springer.
- ECB (2003), *Seasonal adjustment*. European Central Bank publication. Available from:
<http://www.ecb.int/pub/pdf/other/statseasonaladjustmenten.pdf/> [Accessed 20 March 2013].
- European Communities (2001), *Seasonal Adjustment of European Aggregates: Direct versus Indirect Approach*. Office for Official Publication of the European Communities, Luxembourg.
- Eurostat (2009), *ESS Guidelines on Seasonal Adjustment*. European Commission, Luxembourg. Available from:
http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-RA-09-006/ [Accessed 20 March 2013].
- Eurostat (2012), *Demetra+*. Available from: <http://www.cros-portal.eu/content/demetra/> [Accessed 20 March 2013].
- Gomez, V. and Maravall, A. (1996), *Programs TRAMO and SEATS: Instructions for the user* (beta version: June 1997). Banco de Espana, Servicio de Estudios, DT 9628.
- Gomez, V. and Maravall, A. (2001a), Automatic modeling methods for univariate series. In: D. Peña, G. C. Tiao, and R. S. Tsay (eds.), *A Course in Time Series Analysis*. John Wiley and Sons, New York, NY.
- Gomez, V. and Maravall, A. (2001b), Seasonal adjustment and signal extraction in economic time series. In: D. Peña, G. C. Tiao, and R. S. Tsay (eds.), *A Course in Time Series Analysis*. John Wiley and Sons, New York, NY.
- Granger, C. W. J. (1978), Seasonality: causation, interpretation and implications. In: Zellner, A. (ed.), *Seasonal Analysis of Economic Time Series*, U.S. Department of Commerce, U.S. Bureau of the Census, Washington D.C., 33–46.
- Hylleberg, S. (ed.) (1992), *Modelling Seasonality*. Oxford University Press, Oxford, New York, Toronto.

- IMF (2001), Seasonal Adjustment and Estimation of Trend-Cycles. *Quarterly National Accounts Manual— Concepts, Data Sources, and Compilation*, International Monetary Fund Publication, Ch. VIII. <http://www.imf.org/external/pubs/ft/qna/2000/textbook/> [Accessed 20 March 2013]
- Koopman, S. J. and Lee, K. M. (2010), *STAMP*. <http://www.stamp-software.com/> [Accessed 20 March 2013]
- OECD (2006), *Glossary of statistical terms*.
<http://stats.oecd.org/glossary/detail.asp?ID=2398/> [Accessed 20 March 2013].
- OECD (2007), *Data and Metadata Reporting and Presentation Handbook*.
- Stuckey, A., Zhang, X. M., and McLaren, C. H. (2004), *Aggregation of Seasonally Adjusted Estimates by a Post-Adjustment*. Methodological Advisory Committee, November 2004, Australian Bureau of Statistics. Available from:
http://www.uow.edu.au/~craigmc/abs_agg_2004.pdf [Accessed 21 March 2013]
- Statistics Canada (2009), *Seasonal adjustment and trend-cycle estimation*.
<http://www.statcan.gc.ca/pub/12-539-x/2009001/seasonal-saisonnal-eng.htm/> [Accessed 20 March 2013].
- U.S. Census Bureau (2011), *X-12-ARIMA Reference Manual (Version 0.3)*.
- U.S. Census Bureau (2012a), *X-12-Arima Seasonal Adjustment Program*. United States Census Bureau. <http://www.census.gov/srd/www/x12a/> [Accessed 20 March 2013]
- U.S. Census Bureau (2012b), *X-13-Arima-Seats Seasonal Adjustment Program*. United States Census Bureau. Available from: <http://www.census.gov/srd/www/x12a/> [Accessed 20 March 2013]
- U.S. Census Bureau (2013), *X-13ARIMA-SEATS Reference Manual (Version 1.1)*.
<http://www.census.gov/srd/www/x13as/>
- Willeboordse, A. (ed.) (1998), *Handbook on the Design and Implementation of Business Surveys*. Office for Official Publications of the European Communities, Luxembourg.

Interconnections with other modules

8. Related themes described in other modules

- 1.

9. Methods explicitly referred to in this module

1. Seasonal Adjustment – Seasonal Adjustment of Economic Time Series

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 6. Analyse - 6.1 Prepare draft outputs

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

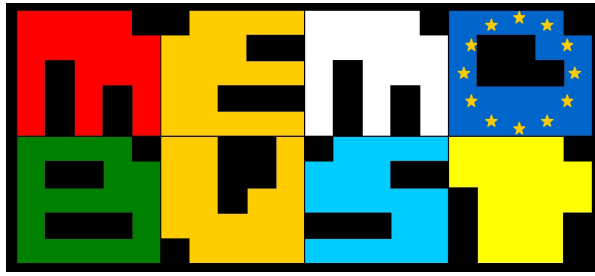
Seasonal Adjustment-T-Introduction

15. Version history

Version	Date	Description of changes	Author	Institute
0.0.1	26-11-2012	first draft: Module 1.3.User's perspective	Suad Elezović	SCB (Sweden)
0.0.2	01-02-2013	second draft: module chapters compiled by HU	Anna Ciammola, Attila Lukacs, Suad Elezović	ISTAT (Italy), KSH (Hungary), SCB (Sweden)
0.0.3	19-03-2013	changes according to reviewers' comments; glossary included	Suad Elezović	SCB (Sweden)
0.0.4	13-06-2013		Anna Ciammola	ISTAT (Italy)
0.0.5	20-09-2013	changes according to reviewers' comments	Suad Elezović	SCB (Sweden)
0.0.6	04-10-2013	updated ch. 4 + minor changes in references	Suad Elezović	SCB (Sweden)
0.1	22-11-2013	version submitted to Editorial Board	Suad Elezović	SCB (Sweden)
0.1.1	10-12-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:28



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Seasonal Adjustment of Economic Time Series

Contents

General section.....	3
1. Summary	3
2. General description of the method	4
2.1 Pre-treatment	4
2.2 Decomposition in TRAMO-SEATS and X-12-ARIMA	14
2.3 STS model based decomposition.....	23
2.4 Step by step seasonal adjustment	30
3. Preparatory phase	39
4. Examples – not tool specific.....	39
5. Examples – tool specific.....	40
6. Glossary.....	40
7. References	40
Specific section.....	44
Interconnections with other modules.....	45
Administrative section.....	47

General section

1. Summary

Seasonal adjustment, which consists in the estimation and the removal of the seasonal variation from time series has a long tradition, documented in Zellner (1978), Hylleberg (1992) and, more recently, Bell, Holan and McElroy (2012). Since both seasonally adjusted series and seasonal component are unobserved components and, consequently, a given time series has an unknown composition, many methods and procedures have been proposed and implemented to perform the seasonal adjustment. In addition to the ARIMA (AutoRegressive Integrated Moving Average)-model based method, implemented in TRAMO-SEATS (Maravall, 2012) and to the moving average based method implemented in X-12-ARIMA (U.S. Census Bureau, 2012), seasonal adjustment may be performed by some more or less conventional methods, such as Structural Time Series (STS) models, Bayesian seasonal adjustment, signal-extraction methods, different non-parametric (like spline-based) methods etc. Although the two main-stream procedures, TRAMO-SEATS and X-12-ARIMA, are generally recognised and accepted as the leading procedures in a process of production of seasonally adjusted data in official statistics, it is still important to study the alternatives in order to encourage diversity in development of seasonal adjustment.

An outline of the several seasonal adjustment procedures used at the National Statistical Offices of the European Union is given in Fischer (1995). Although this document might look outdated it still contains interesting comparisons among several methods used in different national institutes in Europe. This document emphasises advantages of TRAMO-SEATS and X-12-ARIMA over the comparing methods DAINITIES, BV4, SABL, X-11 UK version and X-11-ARIMA. Note that some of the methods described in the document are no longer in use.

Since the time of publication of the mentioned document several new methods for seasonal adjustment have been proposed in the available literature. These methods arise because of the need to deal with some issues that ARIMA-model based methodologies have difficulty tackling. Real time signal extraction is one such methodology based on the Direct Filter Approach (Wildy, 2008), implemented in the R-package signal extraction (R Development Core Team, 2012) created by the same author. The author claims that this method has certain advantages over the ARIMA-model based methods with respect to the turning-point detection and other relevant timing issues.

Non-parametric methods such as STL allegedly generate robust estimates of the time series components not distorted by aberrant observations (outliers). See Cleveland (1990) and R-package STL for more details (R Development Core Team, 2012). Although robust to outliers, the STL-method has some disadvantages in official statistics. This procedure does not have full functionality needed to produce seasonally adjusted estimates in a way relevant to a government statistical agency. Furthermore, the development of this method seems to be stagnated during recent years.

Bayesian seasonal adjustment, originally proposed by Akaike (1980), has been developed and implemented in several software-platforms, such as R-package TIMSAC and SAS procedure TSBAYSEA (SAS Institute, 2009). However, such a methodology has not yet attracted attention of the national statistical institutes, due to its complexity and the required theoretical background necessary to deal with the Bayesian framework.

One of the alternative modelling frameworks, the STS-models, is recommended as a substitute to the two main methods in the ESS (European Statistical System) guidelines on seasonal adjustment (Eurostat, 2009), if certain conditions are satisfied. The use of some other alternative methods falls under the category “to be avoided”.

The ESS guidelines on seasonal adjustment aim to achieve harmonisation of the member state’s national practices by promoting the idea of best practices in seasonal adjustment. Although the guidelines work towards a unified framework for seasonal adjustment within the ESS, they are not supposed to put limitations on the use of other methods. Under appropriate circumstances some less conventional models might offer innovative solutions to certain re-occurring problems that the national statistical institutes (NSI) have to deal with in their daily work with seasonal adjustment.

The main focus of this module is put on description of the decomposition based on ARIMA models, on moving averages and on STS-models, while the other classes of models are not treated. Section 2 is organised as follows. Sections 2.1 and 2.2 describe the two main stages of the seasonal adjustment of a given time series through the most widespread procedures, i.e., TRAMO-SEATS and X-12-ARIMA (X-13-ARIMA-SEATS): the pre-treatment and the decomposition. In particular, section 2.1 deals with the pre-treatment of time series required by both procedures before the decomposition and section 2.2 gives an overview of the decomposition based on moving averages (or *ad hoc* filters) and ARIMA models. Section 2.3 presents the STS model based approach, highlighting features that make it an appealing tool for seasonal adjustment. Finally, referring to TRAMO-SEATS and X-12-ARIMA, section 2.4 details the seasonal adjustment process of time series, distinguishing and describing eight steps.

2. General description of the method

2.1 Pre-treatment

The most widespread procedures of seasonal adjustment, TRAMO-SEATS and X-12-ARIMA, require the pre-treatment of time series aimed at adjusting the original series for special effects before the decomposition. Usually these effects refer to calendar effects, outliers, particular events known a-priori and so forth and the adjustment is carried out through reg-ARIMA models. These are presented in this section, while a different approach is considered in section 2.3.

2.1.1 Reg-ARIMA models

ARIMA models, as discussed by Box and Jenkins (1976), represent a practical way of dealing with moving features of seasonal time series. A general multiplicative seasonal ARIMA model for a time series Y_t can be written

$$\phi(B) \Phi(B^s) (1 - B)^d (1 - B^s)^D Y_t = \theta(B) \Theta(B^s) a_t \quad (1)$$

where

- Y_t may be replaced by deviations from its mean, $Y_t - \mu$;
- B is the backshift operator, such that $BY_t = Y_{t-1}$;
- s is the seasonal period ($s = 12$ for monthly data, $s = 4$ for quarterly data, ...);

- $\phi(B) = (1 - \phi_1 B - \dots - \phi_p B^p)$ is the non-seasonal AutoRegressive (AR) polynomial of order p and $\Phi(B^s) = (1 - \Phi_1 B^s - \dots - \Phi_P B^{sP})$ is the seasonal AR polynomial of order P ;
- $(1 - B)^d$ and $(1 - B^s)^D$ imply, respectively, the non-seasonal differencing of order d and the seasonal differencing of order D (generally $d = 0, 1, 2$ and $D = 0, 1$);
- $\theta(B) = (1 - \theta_1 B - \dots - \theta_q B^q)$ is the non-seasonal moving average (MA) polynomial, $\Theta(B^s) = (1 - \Theta_1 B^s - \dots - \Theta_Q B^{sQ})$ is the seasonal MA polynomial;
- $a_t \sim WN(0, \sigma^2)$ is a white noise process with mean zero and variance σ^2 .

In order to build ARIMA models the so called Box-Jenkins approach is used. It consists of an iterative scheme containing three stages: *i*) model identification, i.e., the selection of a tentative model, in particular the selection of the degree of regular/seasonal differencing and the orders of the stationary AR and invertible MA polynomials; *ii*) estimation of $(p + P + q + Q)$ parameters of the AR and MA polynomials and of the white noise variance; *iii*) diagnostic checking, mainly based on model residuals, assumed to be normally, identically and independently distributed (n.i.i.d.), on statistical significance of parameters and on in-sample and out-of-sample forecast performance (useful references are Box and Jenkins (1976), Harvey (1989), Hendry (1995)).

In model identification, it is important to employ the smallest possible number of parameters for an adequate representation of the time series. This principle of parsimony is particularly important in time series analysis, because variables in a time series model are usually autocorrelated and cross correlated. If a model is not reasonably parsimonious, such correlations may lead to spurious relationships in the model.

Since this model building process is often complex and time consuming, the choice of the ARIMA model is often based either on information criteria¹ such as AIC (Akaike Information Criterion), its corrected version AICC, BIC (Bayesian Information Criterion) and others or on automatic procedures. As far as information criteria are concerned, they are expressed in terms of the maximum of log likelihood and a penalty function depending on the number of parameters. The use of these information criteria implies that if different models produce similar maximum values of the log likelihood, the model with fewer parameters should be preferred. On the contrary, an additional parameter should be added in the model only when the maximum value of the log likelihood increases substantially. Although the choice of the best information criterion is not an easy task, all information criteria share the general principle of parsimony.

With reference to the automatic procedure for ARIMA model identification, it is worth emphasising that model identification is the most important step in the model building process influencing parameters estimates, forecasting and decomposition. The availability of a powerful automatic procedure for model identification in the most widespread procedures used for seasonal adjustment (Gomez and Maravall, 2001) TRAMO-SEATS and X-12-ARIMA (X-13-ARIMA-SEATS), has greatly simplified the seasonal adjustment, allowing a massive treatment and decomposition of many seasonal time series and enhancing the overall quality of data.

¹ Useful suggestions for a proper use of the information criteria to compare several (reg-)ARIMA models can be found on the X12 user manual (Census Bureau, 2013).

A particular class of ARIMA models is the *airline model*, so called because applied to a series of airline passengers in Box and Jenkins (1976):

$$(1 - B)^d (1 - B^s)^D Y_t = (1 - \theta B)(1 - \Theta B^s) a_t. \quad (2)$$

It is a parsimonious model providing a good fit for many seasonal macroeconomic time series. Its parameters can be given a structural interpretation (see Kaiser and Maravall, 2001):

- a) the trend behaviour becomes more and more stable when $\theta \rightarrow 1$;
- b) the seasonal behaviour becomes more and more stable when $\Theta \rightarrow 1$.

Anyway, attention should be paid when, estimating an airline model, its parameter estimates are near the non-invertibility region (e.g., estimates of θ and/or Θ are -0.99). In fact, two reasons can explain this result: either trend/seasonality are practically deterministic or the model is overdifferenced. Testing for the significance of a linear trend or seasonal dummies determines the correct explanation.

Before considering a time series appropriate for ARIMA models, several prior treatments (adjustments) are generally needed in order to:

- remove special effects such as working/trading day, Easter effects and other national holidays (*calendar effects*);
- correct outliers;
- deal with special events known a-priori through intervention variables.

These pre-adjustments are implemented using a regression ARIMA model (hereinafter reg-ARIMA model), also called time series regression model or dynamic regression model (Pankratz, 1991).

A reg-ARIMA model can be written as

$$Z_t = \sum \beta_i X_{i,t} + Y_t \quad (3)$$

where Z_t is the (observed) time series, the $X_{i,t}$ are regression variables observed concurrently with Z_t , the β_i are regression parameters and $Y_t = Z_t - \sum \beta_i X_{i,t}$, the time series of regression errors (hereinafter called *linearised series*), is assumed to follow the ARIMA model in (1). The expressions (1) and (3) define the general reg-ARIMA.

In the reg-ARIMA model written in (3), the regression variables $X_{i,t}$ affect the dependent series Z_t only at concurrent time points, i.e., model (3) does not explicitly consider lagged regression effects $X_{i,t-1}$. Moreover, regression variables are deterministic variables, whose future values can be exactly predicted with a null forecast error. Lagged and stochastic effects can be included in the reg-ARIMA models implemented in the most recent releases of TRAMO-SEATS.

In order to include regression variables in the model, user knowledge about the time series being treated is required. Some variables that are frequently used are generated by the programs used for the seasonal adjustment, while other specific variables needed to deal with specific effects/abrupts in time series can be created by the user. Next section deals with three main groups of regression variables: calendar variables, outliers and intervention variables.

2.1.2 Regression variables

A. Regression variables for calendar effects

Many economic time series, such as production, sales and turnover, are an aggregation of unobserved daily values and are compiled each month. These time series may contain two kinds of calendar effects: the *trading day effect* (or day-of-week effect) and moving holidays (e.g., Easter) that are set according to a lunar calendar.

The trading day effect results from a combination of an underlying weekly periodicity in the unobserved daily data along with how many times each day of the week occurs in a given month. For example, July 2013 began on a Monday, so there are five Mondays, Tuesdays and Wednesdays and four of each of the other days. In July 2011, there are five Fridays, Saturdays and Sundays and four of each of the other days. Thus, the weekly periodicity along with the different numbers of each weekday may considerably affect time series. This can be shown comparing the sample autocorrelation function of unadjusted data with the one of data adequately treated for trading day effects. In fact, when the time series being analysed is significantly affected by these effects, its sample autocorrelation function may be seriously distorted. Moreover, since the ARIMA model suggested by its profile is not a parsimonious and easily interpretable model, these effects must be properly accounted for before a meaningful analysis of the data can be conducted.

Methods used to deal with trading day effects are based on the counting of the number of specific weekdays in a given month t (i.e., the number of Mondays $W_{1,t}$, the number of Tuesdays $W_{2,t}$, ..., the number of Sundays $W_{7,t}$). These counts are then used as regression variables and the total trading day effects can be written as

$$td(\zeta_1, \dots, \zeta_7, W_{1,t}, \dots, W_{7,t}) = \sum_{i=1,7} \zeta_i W_{i,t} \quad (4)$$

with ζ_i , $i = 1, \dots, 7$ representing the effects due to Mondays, ..., Sundays (here ζ_i and $W_{i,t}$ play the same role as β_i and $X_{i,t}$ in equation (3)). To avoid multicollinearity and also to consider the non-seasonal part of the trading day effects (as required in seasonal adjustment), the trading day effects are constrained to vary around zero, i.e., their long run average is required to be null

$$1/n \sum_{t=1,n} \sum_{i=1,7} \zeta_i W_{i,t} = 1/n \sum_{i=1,7} \zeta_i \sum_{t=1,n} W_{i,t} = 0 \quad (5)$$

where $n = 12 \times 28$ because the calendar is periodic of 28 years (if only years not multiple of 400 are considered in the 28 year span). It follows that relation (5) is fulfilled for $\sum_{i=1,7} \zeta_i = 0$, yielding $\zeta_7 = -\sum_{i=1,6} \zeta_i$, and therefore

$$td(\zeta_1, \dots, \zeta_7, W_{1,t}, \dots, W_{7,t}) = \sum_{i=1,6} \zeta_i W_{i,t} - \sum_{i=1,6} \zeta_i W_{7,t} = \sum_{i=1,6} \zeta_i (W_{i,t} - W_{7,t})$$

$$TD(\zeta_1, \dots, \zeta_6, D_{1,t}, \dots, D_{6,t}) = \sum_{i=1,6} \zeta_i D_{i,t} \quad (6)$$

with $D_{i,t}$ representing the *contrast* variables built using the variable for Sunday, $W_{7,t}$. The use of Sunday in (6) to build contrast variables is usual in the literature. However, in a more general approach each day of the week could be used, depending on the features of the economic activities/domains being considered (see Attal-Toubert and Ladiray, 2011).

Additionally, another regression variable can be included to model the length of the months, namely the leap year variable $LY_t = \sum_{i=1,7} W_{i,t} - l_m$, where

$$l_m = \sum_{i=1,7} W_{i,s} \text{ for } m = \text{January, March, ..., December} \quad (7)$$

$$l_m = 1/n \sum_{t=1,n} \sum_{i=1,7} W_{i,s} = 28.25 \text{ for } s = \text{February and } n = 4,$$

is the average length of months. In particular, LY_t is not null only for the months of February (0.25 when the month t is a February with 28 days and -0.75 when the month t is a February with 29 days).

Sometimes in the reg-ARIMA estimation stage, some trading day parameters may not be statistically significant. In these cases, it is important not to eliminate the insignificant parameters, because the whole set of variables has to be completely retained or completely removed. On the contrary, the effect due to leap year, when statistically insignificant, may be omitted.

There is a more parsimonious representation of the effects due to the composition of calendar based on one regression variable. It is supposed that Monday to Friday have similar effects, while Saturday is treated as contrast variable along with Sunday. Its final representation is:

$$WD(\zeta, D_t) = \zeta D_t = \zeta (\sum_{i=1,5} W_{i,t} - 5/2 \times \sum_{i=6,7} W_{i,t}). \quad (8)$$

Usually (6) and (8) are referred to as *trading day* effects and *working day* effects, respectively.

As far as the calendar adjustment for working/trading days is concerned, two aspects deserve to be stressed: the one refers to the treatment of the national (civil or religious) holidays falling on working/trading days (point of view of data producers); the other concerns the interpretation of working-day adjusted data when they are disseminated to users (point of view of data users).

1. Among the several methods existing to adjust for trading-day and holiday effects in monthly economic time series, two methodologies are widespread among NSIs (Roberts *et al.*, 2009): one based on the U.S. Census Bureau's X-12-ARIMA method and one developed by Eurostat and suggested in the ESS guidelines on seasonal adjustment (Eurostat, 2009).
 - a. According to the U.S. Census Bureau's X-12-ARIMA method, fixed national holidays falling on a particular date or on a particular working/trading day of a given month are expected to have fixed effects (not affecting other months) and, consequently, to be absorbed by the seasonal component of the series. There is no need to include further regressors for these holidays in the reg-ARIMA model.
 - b. According to Eurostat's method, fixed national holidays falling on trading/working days are included in the above mentioned regressors and treated as Sunday. These regressors, corrected for fixed holidays and called country specific regressors, are expressed as:

$$\begin{aligned}
& (\# \text{ Mon}_t - \# \text{ hol}_{\text{Mon},t}) - (\# \text{ Sun}_t + \# \text{ hol}_{\text{Mon},t}) \\
& \dots \\
& (\# \text{ Sat}_t - \# \text{ hol}_{\text{Sat},t}) - (\# \text{ Sun}_t + \# \text{ hol}_{\text{Sat},t}),
\end{aligned} \quad (9)$$

where $\# \text{ hol}_{\text{Mon},t}$ is the number of fixed holidays falling on Monday for the month t , or

$$(\# \text{ Mon}_t - \# \text{ hol}_{\text{Mon},t}) - (\# \text{ Sun}_t + \# \text{ hol}_{\text{Mon},t}) \quad (10)$$

where $\# \text{ hol}_t$ is the number of fixed holidays falling on Monday, Tuesday, ..., Friday for the month t . The main drawback of these country specific regressors is that they show a seasonal pattern. The first panel of figure 1 represents the autoregressive spectrum of an example of the regressor described in equation (4): spectral peaks are evident at both calendar frequencies (vertical pink lines) and seasonal frequencies (vertical dotted red lines). As stressed in the guidelines on seasonal adjustment, regression variables related to calendar effects have to remove only the non-seasonal part of these effects, since the seasonal part will be removed in the decomposition stage. Since the variables described in equations (3) and (4) show seasonality, the non-seasonal part of the day-of-week composition of the month/quarter can be estimated by the deviation of the number of working/trading days from their long-term monthly/quarterly average, i.e., removing monthly or quarterly averages (computed on a calendar whose length is a multiple of 28 years).

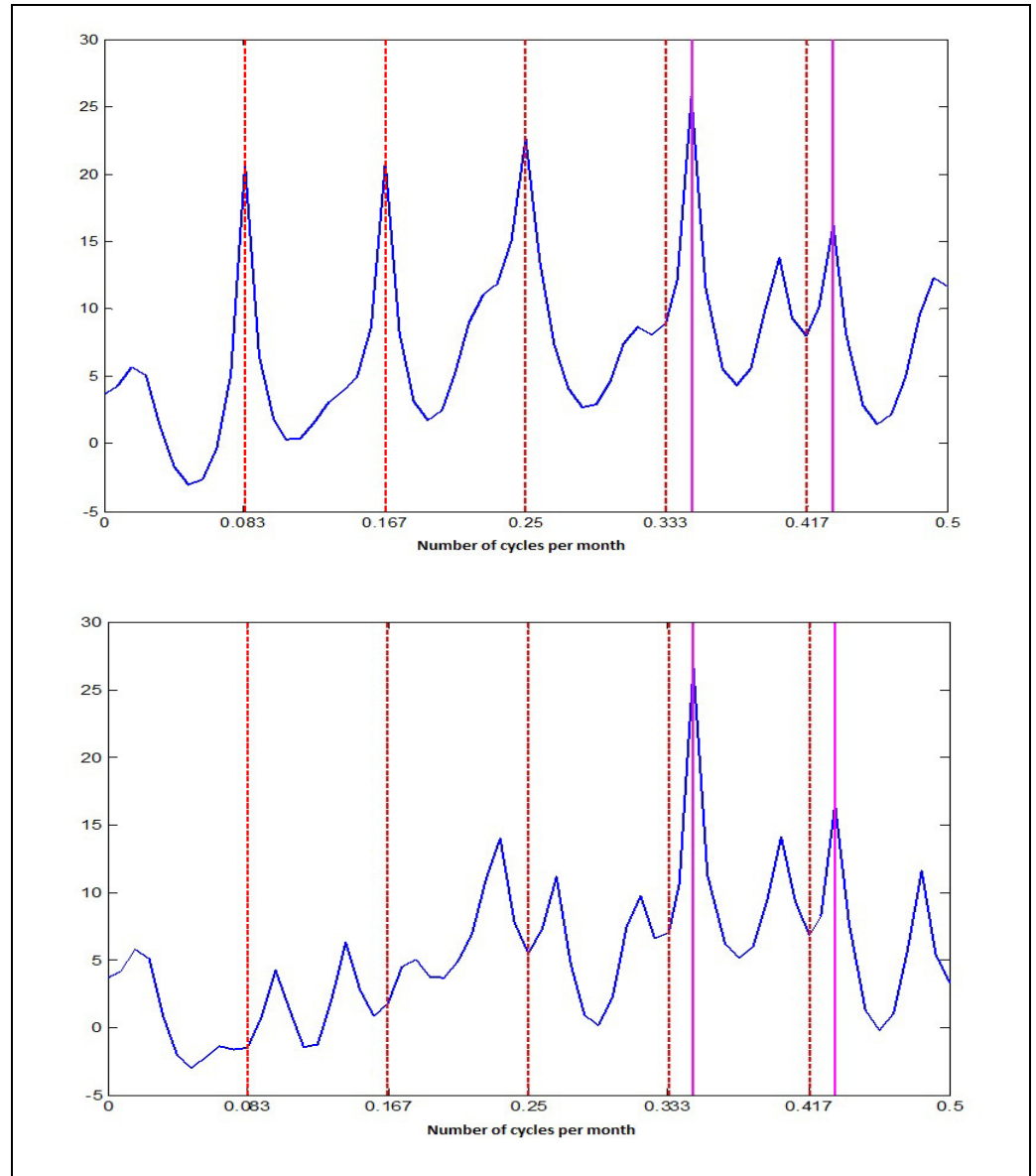


Figure 1: Autoregressive spectrum of regressor described in equation (10) (upper panel) and of its deseasonalised version (lower panel).

2. Another issue concerning the interpretation of calendar adjusted data (here the adjustment based on one regressor is considered) refers to comparison of y-o-y growth rates computed on both unadjusted and calendar adjusted data (in particular index number) when the same period, month or quarter, of the year y and $y-1$ have the same number of working days. In this case, in fact, their equality is expected. However, there may be cases where the equality is not fulfilled, in particular when the additive model is used. In fact, the additive adjustment for calendar adjustment is not proportional to the data level with the consequence that smallest data are overadjusted. Moreover the size of the difference between the two types of y-o-y growth rates (in case of additive model) depend on the size of the unadjusted y-o-y growth rates: the larger they are, the larger the difference is. This is shown in figure 2 where the difference between the y-o-y growth rates calculated on the unadjusted and the calendar adjusted data is reported on the vertical axis. It depends on the levels of data to be calendar adjusted (here the index numbers are considered) and on the size of the y-o-y growth rates of unadjusted data (in the figure they are displayed in percentages). For the multiplicative model, the light blue surface, intersecting the vertical axis at value zero, shows that when a period (month or quarter) has the same number of working days for two consecutive year (y and $y-1$) y-o-y growth rates on working day adjusted data are equal to y-o-y growth rates on unadjusted data (their difference is null as expected). On the contrary, for the additive model, small values (levels) are overadjusted and differences between unadjusted and working-day adjusted y-o-y growth rates are larger. This is emphasised when unadjusted low levels are associated with large (absolute) y-o-y growth rates. This situation is very common with time series featured by an important seasonal component with very small values in at least one period.

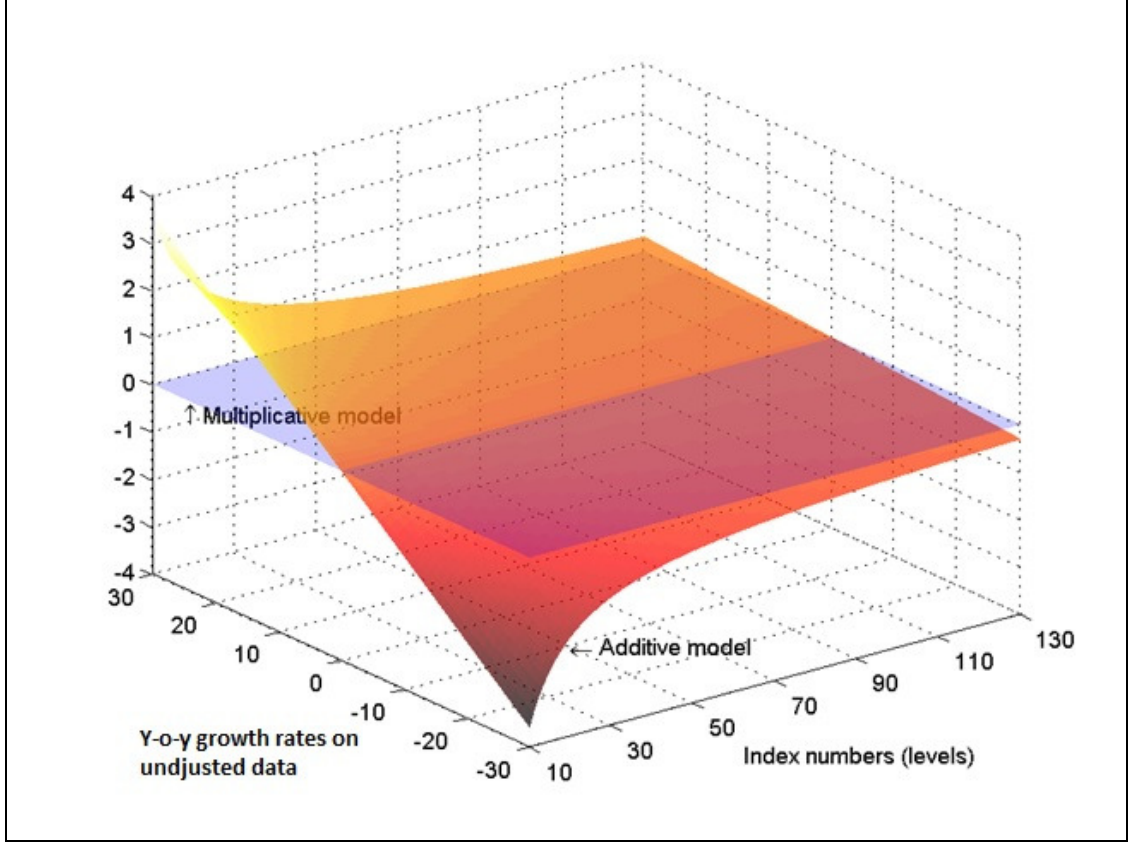


Figure 2: Differences between y-o-y growth rates computed on unadjusted and working day adjusted (reported on the vertical axis).

There are holidays that need a different correction because they are set according to the lunar calendar and, therefore, may fall in different days/months. One of them is Easter holiday. It represents a mobile holiday that may fall between the 22nd March and the 25th April, whose effects refers to the days/weeks before the holiday. An example is represented by the sales turnover that generally increases before Easter. For Christmas holiday, sales turnover also increases before the holiday, but it does not require a specific treatment because it falls in the same date every year.

Adjusting a time series for the Easter effects, therefore, requires a specific variable:

$$E_{\gamma,t}^* = (1/\gamma) \times n_t \quad (11)$$

where γ is the length (the number of days) of the Easter effect before Easter Sunday and n_t is the number of the γ days before Easter falling in month t . For example, if Easter falls the 4th April, under the hypothesis that the effects of the holiday lasts $\gamma = 6$ days, then this variable is null except in March and April, when it is 2/6 in March and 4/6 in April. In February it is nonzero only when $\gamma > 22$.

The deseasonalised and actual used version of this variable is obtained by removing the long run monthly averages of $E_{\gamma,t}^*$ computed on a long period (in X-12/X-13 a 500 year period of the Gregorian calendar is considered).

$$E_{\gamma,t} = E_{\gamma,t}^* - 1/T \sum_{t=1,T} E_{\gamma,t}^* \quad (12)$$

where T is the number of years in the period considered to calculate the averages.

The preceding paragraphs are based on three assumptions. Firstly, time series are available at monthly frequency. However, trading/working day effects can be found also in quarterly series, although they are not very common because the calendar composition of quarters does not change over time as that of months. In these cases, regressors are built counting the days of the week over the quarters. Secondly, time series are compiled aggregating daily values (flow series). If the series instead are compiled using the values at the end of the month (stock series), then different regression variables have to be used for an adequate adjustment of calendar effects (see Bell, 1984a and 1995, and Findley and Monsell, 2009). Thirdly, calendar effects are modelled through deterministic regression variables. In the ARIMA model based decomposition, Maravall and Pérez (2012) propose a stochastic trading/working day component (when the ARIMA model contains a regular AR polynomial, whose complex root has an associated frequency approximately equal to the theoretical trading/working day frequency)

B. Outliers

Macroeconomic time series are often subject to external events or abrupt changes such as introduction of new laws/regulations, sales promotions, strikes, recording errors and so forth. When these events are unexpected and their timing is unknown (e.g., recording errors), they are referred to as outliers, i.e., unusual observations that have a substantial impact on the time series and, consequently, on their analysis. Although several methods have been proposed for detection and adjustment of outliers, usually an automatic approach is used based on an iterative procedure (for details see Chen and Liu, 1993 and Gomez and Maravall, 2001a).

There are several reasons for outlier detection and adjustment in time series analysis (Pankratz, 1991):

- a. understanding the time series under study;
- b. discovering spurious observations such as recording errors;
- c. simplifying the structure of the model and improving parameter estimates;
- d. improving the forecasting performance.

All these motivations may have moderate to substantial impact on the seasonal adjustment of time series, in particular the improvement of the estimation of components (especially in an ARIMA model based approach) and the reduction of the revision size for seasonally adjusted data (when new observations are added).

In this section four types of outliers are presented, while their allocation to the different components is considered in section 3.

1. Additive outlier (AO)

An additive outlier is an event that affects a time series for one period only, $t = t_0$. It can be represented through a *pulse* function:

$$P_t(t_0) = 1 \quad \text{for } t = t_0, \quad P_t(t_0) = 0 \quad \text{for } t \neq t_0.$$

The reg-ARIMA model for the time series is

$$Z_t = \omega_{AO} P_t(t_0) + Y_t$$

where the value ω_{AO} , to be estimated, represents the deviation from the “true” value of Y_t and Y_t is assumed to follow the ARIMA model in (1).

2. Level shift (LS)

A level shift is an event that affects a time series permanently from a period $t = t_0$ onward. It can be represented by a *step* function:

$$S_t(t_0) = -1 \quad \text{for } t < t_0, \quad S_t(t_0) = 0 \quad \text{for } t \geq t_0.$$

The reg-ARIMA model for the time series is

$$Z_t = \omega_{LS} S_t(t_0) + Y_t$$

where the term $\omega_{LS} S_t(t_0)$ adjusts for the level of the time series Z_t in first part, adapting it to the one of second part.

3. Temporary change (TC)

A temporary change is an event that has an initial impact on the time series at $t = t_0$ and whose effect decays exponentially according to a factor $\delta \in (0,1)$, called dampening factor (i.e., the rate of decay back to the previous level of the time series):

$$T_t(t_0) = \delta^{t-t_0} \quad \text{for } t \geq t_0, \quad T_t(t_0) = 0 \quad \text{for } t < t_0.$$

The reg-ARIMA model for the time series is

$$Z_t = \omega_{TC} T_t(t_0) + Y_t.$$

4. Seasonal outliers (SO)

A seasonal outlier is an event that affects one period (month or quarter) of a time series permanently from time $t = t_0$ onward (Kaiser and Maravall, 2003). It can be represented by the following function (assuring null annual averages):

$$SO_t(t_0) = \begin{cases} 1 & \text{for } t < t_0 \text{ and } t \text{ same month/quarter as } t_0 \\ 0 & \text{for } t \geq t_0 \\ -(s-1)^{-1} & \text{otherwise.} \end{cases}$$

where s is the seasonal period ($s = 12$ for monthly data, $s = 4$ for quarterly data).

The reg-ARIMA model for the time series is

$$Z_t = \omega_{SO} SO_t(t_0) + Y_t$$

where the term $\omega_{SO} SO_t(t_0)$ adjusts for the level of the month/quarter of the time series Z_t in first part and slightly modifies the level of the other months/quarters. As requirement of seasonal adjustment, the annual sums of the variable $SO_t(t_0)$ are always null. In fact, in the decomposition of a time series the SO are assigned to the seasonal component and, therefore, have to be removed from the seasonally adjusted series without modifying the annual sums (or averages) of the unadjusted series.

There is another type of outlier, called innovational outlier (IO), which affects a time series from a period $t = t_0$ onward according to the ARIMA model of the process. It can be considered an AO

altering the white noise process a_t (see Chang, Tiao and Chen (1988) for further details and references). It is not considered here as it cannot be treated in the decomposition.

As far as outliers are concerned, the issue of detect an outlier at the end of a time series has to be stressed. In order to identify the type of an outlier some observations after the time of the occurrence of the event are needed. When the event occur at the end of the series under study, we are able to detect the outlier (unless its effects are moderate or negligible), but we cannot identify its nature (type). Although this inability affects neither the estimates of the model parameters, nor the estimated seasonally adjusted series (unless the detected outlier is a SO), it can seriously affect the estimation of the other components (i.e., trend and irregular) and the forecasting of both the unadjusted series and its components. As a consequence, attention should be paid when an outlier is detected at the end of the series. Some recommendations are listed below:

1. avoiding outliers at the end of the time series, unless they have a substantial impact on the parameter estimates;
2. if an outlier is detected at the end of the series, information should be collected to explain the reason of the outlier;
3. when an outlier at the end of the series is included in the model, its type should be checked as new observations become available.

As final remark, it is worth noting that all the outliers considered in this section can be detected automatically in the most recent releases of X-13 and TRAMO-SEATS. However, as far as the detection of SO is concerned, the plot of seasonal-irregular ratios computed on the preliminary components before adjusting for outliers may be very useful.

C. Intervention variables

As already said, macroeconomic time series are often subject to external events or abrupt changes such as introduction of new laws/regulations, sales promotions, strikes, recording errors and so forth. When these events are known (e.g., introduction of new laws/regulations) they are referred to as interventions. Intervention analysis is the method to incorporate such effects on the models. It is not considered in this section (an exhaustive treatment is presented in Box and Tiao, 1975).

2.2 Decomposition in TRAMO-SEATS and X-12-ARIMA

Completed the preliminary treatment aimed at removing the calendar effects, the outliers and other deterministic effects and estimating possible missing values, the resulting time series (the so-called *linearised series*²) are decomposed into the unobservable or latent components trend-cycle, seasonality and irregular. The most widespread procedures used by NSIs and other international agencies to produce official seasonally adjusted data are TRAMO-SEATS and X-12-ARIMA (they are also suggested by the ESS guidelines on seasonal adjustment). The former implement an ARIMA model-based decomposition, while the latter decompose a time series applying moving averages according to a recursive approach. Notwithstanding, these procedures have some common features: firstly, the models used are linear stochastic processes parametrised in the ARIMA-type format; secondly, to fulfil the previous assumption, the series needs some modification, called pre-treatment. So, assuming

² It is called linearised series because it can be assumed to be generated by a linear process.

an additive decomposition, the seasonal adjustment performed through the two approaches can be set in a unique framework described in figure 3. Given an observed time series, X_t , a reg-ARIMA model is estimated on it to derive: *i*) the regression effects, representing the deterministic part of the series; *ii*) the autocorrelated disturbance of the regression, modelled with an ARIMA model, representing the purely stochastic part of the series (i.e., the linearised series Y_t). This latter is decomposed, obtaining the stochastic components, hereinafter called simply components. The final components are derived summing up the regression effects to the components, according to their nature. Considering only the most frequent effects treated, the following rules are generally considered:

- 1) calendar effects and seasonal outliers are assigned to the seasonal component (so they do not appear in the seasonally adjusted series);
- 2) level shifts and ramp effects are assigned to the cycle trend;
- 3) transitory changes and additive outliers are assigned to the irregular component.

There is another practical reason to require pre-adjustment: filters used to estimate the components are two-sided filters involving past, present and future observations (and consequently past, present and future outliers or special effects/events), that is

$$S_t = \dots + v_{-2} Y_{t-2} + v_{-1} Y_{t-1} + v_0 Y_t + v_1 Y_{t+1} + v_2 Y_{t+2} + \dots$$

$$= (\dots + v_{-2} B^2 + v_{-1} B + v_0 + v_1 F + v_2 F^2 + \dots) Y_t = v(B, F) Y_t.$$

In order to avoid this, such effects are removed and, after the decomposition, they are re-assigned to the components.

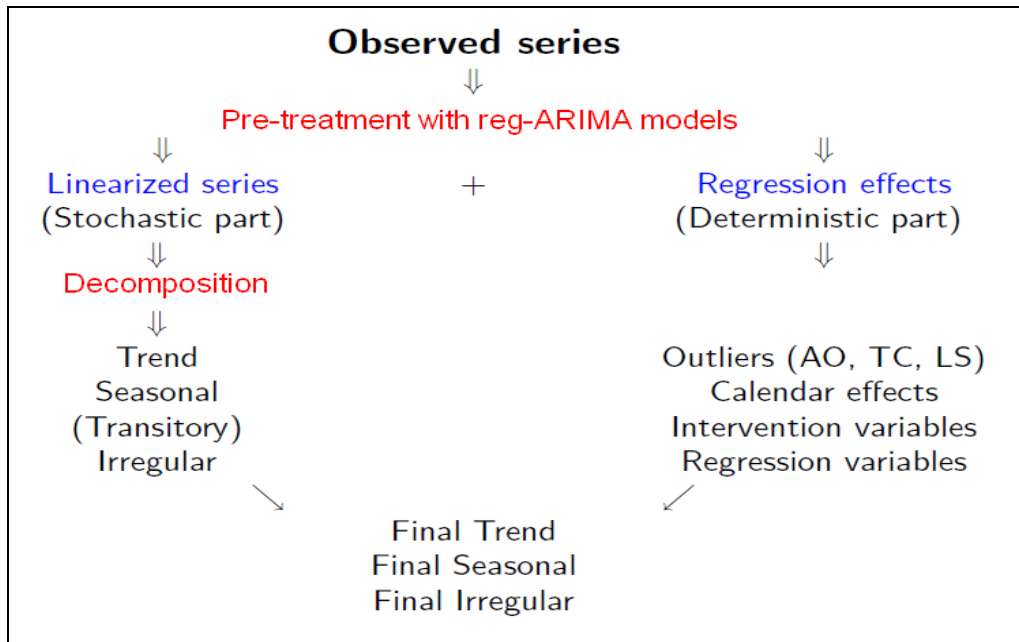


Figure 3: Pre-treatment and decomposition of a time series using reg-ARIMA models.

This section focuses on the decomposition of the linearised series (left side of figure 3): the approach based on moving averages (*ad hoc* filters) is presented in subsection 2.2.1, the ARIMA model based approach is described in subsection 2.2.2.

2.2.1 The moving averages based decomposition of X-12-ARIMA

The X-12-ARIMA decomposition can be viewed as sophisticated use of non-parametric smoothing based on filtering techniques. The two elements are combined into an algorithm that also takes into account extreme observations.

Moving averages

A time series can be smoothed by using the three-term simple (equal weights) moving average

$$P_t = (Y_{t-1} + Y_t + Y_{t+1})/3 \quad (13)$$

This method is called a 3x1 moving average. One may perform such smoothing twice to obtain a 3x3 moving average. That is, the smoothed series is calculated as a three-term simple moving average of a three-term simple moving average. Then, it follows that

$$P_t = (Y_{t-2} + 2Y_{t-1} + 3Y_t + 2Y_{t+1} + Y_{t+2})/9 \quad (14)$$

The centre of four successive observations is between two time periods. The mean of two four-term simple averages is, however, centred at a time period. This is called a 2x4 moving average and the expression is

$$P_t = (Y_{t-2} + 2Y_{t-1} + 2Y_t + 2Y_{t+1} + Y_{t+2})/8 \quad (15)$$

Equations (14) and (15) are examples of two five-term general moving averages. The set of weights is also called a filter and in this case the filter length is five. One may write these two filters in a more compact form as $[1,2,3,2,1]/9$ and $[1,2,2,2,1]/8$. Since we have symmetry this can also be written as $[1,2,\underline{3}]/9$ and $[1,2,\underline{2}]/8$ (centre underlined).

With a fixed filter length, the variance is minimal when the weights are equal. Of course, one can always reduce variance by increasing the filter length. The best filter reduces variance (eliminate noise) without losing too much relevant information. Accordingly, the filter length depends on the variability of the series.

At the beginning and the end of the series asymmetric filters can be used to solve the problem of non-available observations. An example of an asymmetric filter is $[-0.034, 0.116, 0.383, 0.534, 0, 0, 0]$. This filter is an asymmetric variant (Musgrave) of the 7-term Henderson filter. More details about Henderson filters can be found below.

The initial decomposition

Below we consider both the additive ($Y_t = T_t + S_t + I_t$) and the multiplicative model ($Y_t = T_t S_t I_t$).

For monthly data the initial estimate of the trend is found by using a 2x12 moving average. This 13-term filter is also known as a centred 12-term moving average. The weights are simply $[1/2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1/2]/12$. Thus

$$T_t = \left(\frac{1}{2}Y_{t-6} + Y_{t-5} + Y_{t-4} + \dots + Y_t + \dots + Y_{t+5} + \frac{1}{2}Y_{t+6} \right) / 12 \quad (16)$$

With this filter length all months are equally weighted. Therefore, a stable seasonal component will not affect this trend estimate.

We now calculate the so-called SI-ratios, denoted as SI (not necessarily S times I). Note that, within this X-12 framework, SI is not necessarily S times I and neither it is an abbreviation of "seasonal index".

$$SI_t = Y_t - T_t \quad (17)$$

for additive models or

$$SI_t = Y_t / T_t \quad (18)$$

for multiplicative models. So in the case of a multiplicative model, the SI-ratio is $S_t * I_t$ and it is the ratio Y_t / T_t . Preliminary seasonal factors are calculated by using the 3x3 moving average to each month.

$$\hat{S}_t = (SI_{t-24} + 2SI_{t-12} + 3SI_t + 2SI_{t+12} + SI_{t+24}) / 9 \quad (19)$$

To normalise these, averages over 12-month periods are calculated. That is, 2x12 moving averages are calculated.

$$\tilde{S}_t = \left(\frac{1}{2}\hat{S}_{t-6} + \hat{S}_{t-5} + \hat{S}_{t-4} + \dots + \hat{S}_t + \dots + \hat{S}_{t+5} + \frac{1}{2}\hat{S}_{t+6} \right) / 12 \quad (20)$$

The seasonal components are now found as

$$S_t = \hat{S}_t - \tilde{S}_t \quad (21)$$

for additive models and for multiplicative models as

$$S_t = \hat{S}_t / \tilde{S}_t \quad (22)$$

The final decomposition is an improvement of this initial estimate. The underlying idea is based on two elements: How the trend can be estimated from a time series without seasonality and how the seasonal component can be estimated from a time series without a trend.

Finding a trend when seasonality is not present

To find the trend when seasonality is not present is a question of smoothing the time series. The initial trend estimate used equal weights for most months. Curvature trends are, however, better fitted using different weights. In fact, one may use negative weights at the ends. This is the case for the so-called Henderson filters (Henderson, 1916) which is used by X-12-ARIMA to obtain the trend. These filters are constructed so that filtering of third degree polynomials leave the time series unchanged. Another criterion is that the sequence of weights should be as smooth as possible. Further details can be found

in Ladiray and Quenneville (2001). As noted in this reference, the coefficients of these moving averages may be calculated explicitly. For an order $2p+1$ average, by letting $n=p+2$, the coefficients for $i = -p, \dots, p$ can be written as

$$\frac{315[(n-1)^2 - i^2][n^2 - i^2][(n+1)^2 - i^2][3n^2 - 16 - 11i^2]}{8n(n^2 - 1)(4n^2 - 1)(4n^2 - 9)(4n^2 - 25)} \quad (23)$$

When using X-12-ARIMA it is possible to specify the filter length manually. Otherwise, the default automatic method will, in the case of monthly series, use 9, 13 or 23 terms. For quarterly series, the program will choose either a 5- or a 7-term Henderson moving average. Using the notation above, the weights of these two filters are $[-21, 84, 160]/286$ and $[-42, 42, 210, 295]/715$, respectively. The 13-term Henderson filter is $[-325, -468, 0, 1100, 2475, 3600, 4032]/16796$.

Finding the seasonality when a trend is not present

The seasonal component is found by applying moving averages to each month or quarter. One alternative is to calculate the simple average of all the values for each month or quarter. This is called a stable seasonal filter. When using other filters it is allowed the seasonal component to vary along time. The older versions of X-12-ARIMA used a 3×5 moving average as default. This is a 7-term filter with weights $[1, 2, 3, 3, 3, 3, 3]/15$. For monthly data this means that a seasonal average would be calculated as

$$P_t = (Y_{t-36} + 2Y_{t-24} + 3Y_{t-12} + 3Y_t + 3Y_{t+12} + 2Y_{t+24} + Y_{t+36})/15 \quad (24)$$

Other possible filters are, 3×1 , 3×3 , 3×9 and 3×15 . Note that the latter filter is simply $[1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3]/45$. By default, the filter type is selected automatically by the program.

The combined algorithm

As mentioned above, asymmetric filters can be used at the beginning and the end of the series. However, X-12-ARIMA still uses symmetric Henderson filters to calculate the trend at the end of the series. Forecasts from the reg-ARIMA modelling are then used in place of the unobserved values.

The algorithm for calculating the trend and seasonal components starts with initial estimates as described above. The whole algorithm involves several steps. One element is downweighting of observations with an extreme irregular component. Iterations are therefore needed. At each stage it is possible to obtain a seasonal adjusted series (seasonality not present) or a series based only on the seasonal and irregular components (trend not present). This way the final estimates are calculated according to text above (“*Finding a trend when seasonality is not present*” and “*Finding the seasonality when a trend is not present*”). For details, see Dagum (1980), Findley et al. (1998) and Ladiray and Quenneville (2001).

2.2.2 The ARIMA model based (AMB) decomposition of SEATS

In the ARIMA model based decomposition, filters depend on the time series features because they are derived from the ARIMA model estimated on the data. Moreover, it is possible to do inference on the estimated components (because their theoretical properties are known) and to derive forecasts for the components together with their confidence intervals.

In order to understand how a filter can depend on the time series features let us consider the following examples (drawn from Maravall, 2012):

1) $Y_t = a_t$, with $a_t \sim WN(0, \sigma_a^2)$

The time series is not seasonal, that is the seasonal component $S_t = 0$, so the filter applied to Y_t to derive $s_t = 0$ should be $u(B, F)=0$.

2) $(I + B + \dots + B^s)Y_t = w_t$, with w_t having an MA structure

The time series is the seasonal component, that is $S_t = Y_t$, so the filter applied to Y_t should be $u(B, F)=1$.

Figure 4 represents the steps of the AMB decomposition. Given the model for Y_t , firstly the models of the unobserved components are derived (if an acceptable decomposition exists), then the Minimum Mean Square Error (MMSE) estimators for components are computed and finally component estimates are derived.

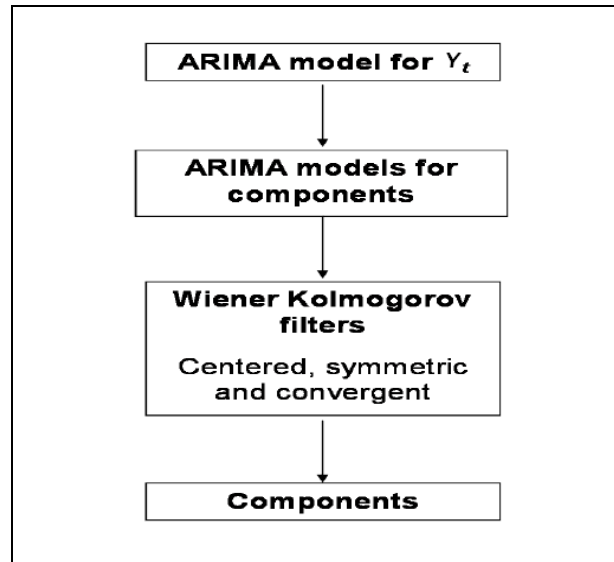


Figure 4: A representation of the AMB decomposition method.

I. Model for Y_t

Given an observed time series, the first step is the identification of the (multiplicative seasonal) ARIMA model:

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D Y_t = \theta(B)\Theta(B^s)a_t, \quad a_t \sim WN(0, \sigma_a^2),$$

Using a more compact notation, it can be re-written as

$$\phi_Y(B)Y_t = \theta_Y(B)a_t.$$

It is worth stressing that the previous model is, in general, invertible and non-stationary. In particular non-stationarity, $d, D > 0$, allows for evolving trend and seasonal component whose features change over time.

II. Decomposition of the model for Y_t

The AR polynomial $\phi_Y(B)$ is factorised, allowing the definition of the components of a given series. For example, if the AR polynomial $\phi_Y(B)$ is $\nabla\nabla_4$ (the product of the regular and the seasonal differencing operators), then it can be factorised as $\nabla\nabla_4 = (1-B)(1+B^4) = (1-B)^2(1+B+B^2+B^3) = \nabla^2 S$, where the factor ∇^2 implies the presence of the trend and the factor S (the annual aggregation operator) implies the presence of the seasonal component. Therefore, the series can be decomposed into trend, seasonality and irregular:

$$Y_t = T_t + S_t + I_t. \quad (25)$$

These components are assumed to follow ARIMA models

$$\phi_T(B)T_t = \theta_T(B)a_{T,t} \quad (26)$$

$$\phi_S(B)S_t = \theta_S(B)a_{S,t} \quad (27)$$

$$I_t = a_{I,t} \quad (28)$$

where $\phi_i(B)$ and $\theta_i(B)$, $i = T, S$, are finite polynomials in B of order p_i and q_i , respectively, having no common zeros and all zeros lying on or outside the unit circle.

As far as the ARIMA model for the trend is concerned (equation 24), generally $\phi_T(B)$ is non-stationary since it contains the regular differencing operator, either $\nabla = (1-B)$ or $\nabla^2 = (1-B)^2$, while the r.h.s. of the model allows the trend to evolve over time (i.e., a stochastic trend). The upper two panels of figure 5 compare a deterministic and a stochastic trend. With reference to the model for the seasonal component (equation 25), usually $\phi_S(B)$ is non-stationary and contains the annual aggregation operator, $S = (1+B+B^2+B^3)$ for quarterly series or $S = (1+B+B^2+\dots+B^{11})$ for monthly series; the r.h.s. of the model allows the seasonal component to evolve over time but preserving regular fluctuation locally. In the lower two panels of figure 5 a deterministic and a stochastic seasonal component are displayed.

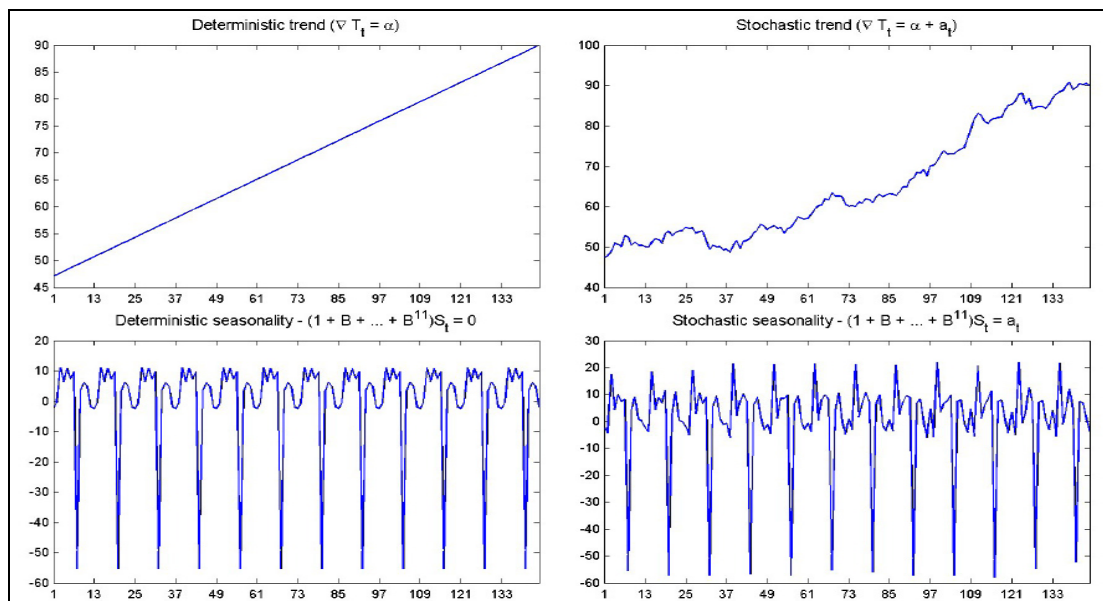


Figure 5: Deterministic and stochastic components.

In the representation (24-26) the following assumptions are fulfilled:

- 1) the variables $a_{T,t}$, $a_{S,t}$ and $a_{I,t}$ are mutually independent white noise processes, identically and independently distributed as $N(0, \sigma_T^2)$, $N(0, \sigma_S^2)$ and $N(0, \sigma_I^2)$;
- 2) the autoregressive polynomials $\phi_T(B)$ and $\phi_S(B)$ do not share common roots;
- 3) the moving average polynomials $\theta_T(B)$ and $\theta_S(B)$ have roots lying on and outside the unit circle and do not share unit common zeros.

The first assumption implies independent components and is based on the consideration that causes of the different components are not much related (e.g., weather causes seasonal fluctuations, while technology and investment cause the evolution of the trend); the second assumption implies that different components (generally non-stationary) are associated with different spectral peaks; the third assumption admits non invertible components and guarantees the invertibility of the model for Y_t .

Exploiting a different representation of the models for both Y_t and components

$$Y_t = \frac{\theta_Y(B)}{\phi_Y(B)} a_t, \quad T_t = \frac{\theta_T(B)}{\phi_T(B)} a_{T,t} \quad \text{and} \quad S_t = \frac{\theta_S(B)}{\phi_S(B)} a_{S,t}$$

from relation (25) the following identity can be derived

$$\frac{\theta_Y(B)}{\phi_Y(B)} a_t = \frac{\theta_T(B)}{\phi_T(B)} a_{T,t} + \frac{\theta_S(B)}{\phi_S(B)} a_{S,t} + a_{I,t}$$

Multiplying both sides for the factorisation of $\phi_Y(B)$, i.e., $\phi_T(B)\phi_S(B)$, the following identity is obtained

$$\theta_Y(B)a_t = \theta_T(B)\phi_S(B)a_{T,t} + \theta_S(B)\phi_T(B)a_{S,t} + \phi_T(B)\phi_S(B)a_{I,t}.$$

Assuming, in general, that $\theta_T(B)$ and $\theta_S(B)$ have the same order of $\phi_T(B)$ and $\phi_S(B)$, respectively, by equating the autocovariance function of both sides one can get a system of equations whose unknowns are the parameters of $\theta_T(B)$, $\theta_S(B)$ and the variances σ_T^2 , σ_S^2 and σ_I^2 . Two issues have to be stressed:

- a) some models do not admit a decomposition, because some components may have a negative spectra;
- b) if a model admits a decomposition, since the number of unknowns are greater than the number of equations, infinite decompositions exist and a choice must be made. This underidentification problem is solved through the *canonical decomposition*, i.e., the decomposition that maximises the variance σ_I^2 and, therefore, minimises the variances σ_T^2 and σ_S^2 . Minimising the latter variances means that the trend and seasonal component are made as stable as possible, remaining compatible with the model for Y_t , and their models became noninvertible.

III. Estimators for the components

The optimal estimators of the trend, seasonal and irregular component are computed as the MMSE estimators, that is as a conditional expectation of S_t given $\{Y_1, Y_2, \dots, Y_T\}$ (here S_t represents the more generic signal)

$$\hat{S}_t = E(S_t | Y_1, Y_2, \dots, Y_T)$$

Assuming the multivariate normal distribution, this conditional expectation is a linear combination of Y_1, Y_2, \dots, Y_T and it can be obtained through either the Kalman filter or the Wiener-Kolmogorov (WK) filter. The latter is considered because it is more useful for analysis.

a) *Historical or final estimators ($T \rightarrow \infty$)*

$$\hat{S}_t = v(B, F)Y_t = \left[v_0 + \sum_{j=1}^{\infty} v_j(B + F) \right] Y_t$$

where $v(B, F)$ represents the WK filter, that is shown to be centred in t , symmetric and convergent in B and F as it represents the autocovariance generating function of a stationary model.

b) *Preliminary estimators (finite realisation)*

$$\hat{S}_{t|T} = v^t(B, F)Y_{t|T}^e$$

where $v^t(B, F)$ is the truncated filter and $Y_{t|T}^e$ is the “extended” series, i.e., the series extended with forecasts and backcasts, with $Y_{t|T}^e = Y_t$ if $t \leq T$ and $Y_{t|T}^e$ is the forecast or the backcast if $t > T$ or $t < 1$. In the particular case $t = T$, $\hat{S}_{T|T}$ is called concurrent estimator.

Figure 6 shows some examples of WK filters to derive the historical estimates of the components.

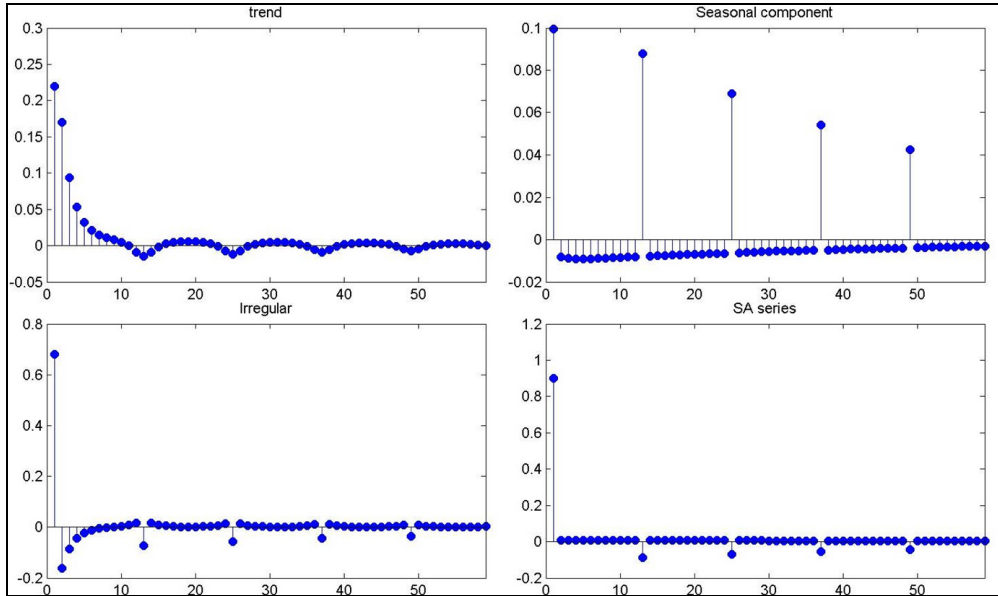


Figure 6: Examples of WK filters

IV. Computation of component estimates: some remarks

The application of the filters (derived from the ARIMA models) to the observed (linearised) series produces the estimates for the components. The comparison of their properties with the (theoretical) properties of estimators (available only in a model-based context) represents a useful diagnostic tool to assess the decomposition.

- 1) Convergence is an important property of filters since it allows us to truncate them when applied to the observed (linearised and extended) series. In many applications it is reached after three-five years, so with a time series of 20 years, the estimates for the central years (10-14 years) can be considered final.
- 2) Symmetry of filters requires the extension of the series with forecasts. As new observations become available, forecasts are replaced with new data and therefore previous component estimates are revised. The revision size depends on the forecast error: the better the series can be forecasted, the smaller the revisions in the preliminary estimates will be. This stresses the importance of the identification of the ARIMA model for the observed series, from which depend the properties of both the component and the estimators.
- 3) Generally, for stable components the convergence of the preliminary estimates to the final ones is slow, while for highly stochastic components the convergence is more rapid but with larger revision errors (trade-off between stability and convergence).

2.3 STS model based decomposition

Time series components in a STS-model

The biggest difference between a Structural Time Series model and an ARIMA-based model, such as TRAMO-SEATS or X-12-ARIMA is in the formulation of the unobserved components. While the components, such as trends and cycles, do not have a direct interpretation in an ARIMA-based model, in a STS model this interpretation is straightforward and direct.

An ARIMA-based decomposition requires a preparatory step including reg-ARIMA- or TRAMO procedures to clean the data from irregularities. In this step differencing of time series to achieve stationarity is almost always imposed resulting in the loss of degrees of freedom. However, for some very noisy series the stationarity can not be achieved in this way, not even if differencing is performed several times. Hence, applying this approach would result in a relatively bad estimates of the so called de-noised series (the error from reg-ARIMA procedure). As this de-noised series is the one to be decomposed into the seasonal effect, the trend-cycle and the irregular component, such an approach which would in turn lead to a large uncertainty in the estimated components.

The STS models on the other hand do not suffer from the stationarity issues since a time series Y_t to be decomposed is directly formulated as the sum of the above mentioned components. Hence, differencing to achieve stationarity is not necessary. Furthermore, the STS models do not require forecasting to obtain the end-point estimates which is an important advantage over the ARIMA-model based methodology. In principle, a univariate STS model may be viewed as a regression model where the explanatory variables are components from the classical decomposition model for a time series Y_t , as formulated here

$$Y_t = T_t + S_t + I_t, \quad t = 1, \dots, T, \quad (29)$$

where T_t is trend, S_t is seasonal component and I_t is irregular component. The explanatory variables are thus functions of time and the parameters are time-varying. As an STS-model may be expressed in many different and complex ways, the first step in the analysis is to find out which modeling alternative is most suitable for a particular time series or a set of time series. This issue is crucial and may appear similar to ARIMA-model based methodology. However, the difference between ARIMA-models and STS-models in this context is big: the STS-modeling framework does not need de-noising of original series in order to obtain a de-linearised series of errors to be decomposed into the basic time series components, which is needed for the reg-ARIMA (TRAMO) part of the ARIMA-model based procedures. Instead, the decomposition is applied directly to the original series, which is treated as a dependent variable. Hence, the pre-treatment step in the STS-models is reduced to find out a plausible modeling alternative within this framework. Once this choice is made the estimation of both the unobserved time series components and the possible other explanatory variables (e.g., calendar factors) is done in one single step. The decomposition is made as an integrated process through a *state space* form where the state of the system is represented through the unobserved components, such as trend-cycle and seasonal components (for details about the state space models see, e.g., Durbin and Koopman, 2001).

Basic Structural Model

Here is given a brief description of the basic STS-model and its components. See, e.g., Harvey (1990) for a more sophisticated description.

In its most basic form a STS-model may be formulated as follows

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \quad t = 1, \dots, T, \quad (30)$$

$$\varepsilon_t \sim i.i.d. \ N(0, \sigma^2),$$

where μ_t , γ_t and ε_t are the trend, seasonal factor and irregular component, respectively. The expression (30) is called the basic structural model (BSM). All components are stochastic and each one is modelled separately.

The random error ε_t is usually called the *irregular component* in the seasonal adjustment literature. In the model's basic form this component is assumed to be a purely Gaussian white noise process, as indicated in (30). This implies that this component is modelled as a sequence of independent, identically distributed zero-mean random variables. Anyhow, the normality property is not exclusive since the irregular component might be modelled in different ways through a more complex modelling alternative. For simplicity, we focus on the basic form of a structural model where the irregular component is modelled either as Gaussian white noise or an ARIMA process (for details see, e.g., Harvey and Shephard (1993)).

Extensions of BSM

Usually a BSM is extended to include the cyclical component and the predictor effects

$$y_t = \mu_t + \gamma_t + \psi_t + \sum_{j=1}^m \beta_j x_{jt} + \varepsilon_t, \quad t = 1, \dots, T, \quad (31)$$

where ψ_t is cycle while the regression term

$$\sum_{j=1}^m \beta_j x_{jt}$$

incorporates effects of fixed regression coefficients that are likely to have influence on the response variable y_t .

Interventions may be included as regression effects as dummy or pulse variables, as explained in Harvey (1990, pp. 397-399). This approach is similar to RegARIMA approach but also allows for some extensions, for example, treating of changes in seasonal pattern.

Modelling the trend component

As mentioned earlier, each unobserved component may be modelled in a different way. As trend-component is defined as the natural tendency of a series in the absence of any noise (seasonality, effects of exogenous variables and unexplained variation expressed in the irregular component) it is natural to start with determining a good general model by modelling the trend in an optimal way. Two most common models for the trend component are the random walk (RW) model and the locally linear time trend (LLT) model. The RW model may be described as a model where the trend movement depends on the variance of the error term:

$$\mu_t = \mu_{t-1} + \eta_t, \quad \eta_t \sim i.i.d. N(0, \sigma_\eta^2). \quad (32)$$

If this variance (σ_η^2) is zero then the trend is simply a constant.

The LLT model involves both the level and the slope in the trend representation:

$$\begin{aligned} \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, & \eta_t &\sim i.i.d. N(0, \sigma_\eta^2) \\ \xi_t &= \beta_{t-1} + \xi_t, & \xi_t &\sim i.i.d. N(0, \xi_t^2) \end{aligned} \quad (33)$$

In (33) the disturbances η_t and ξ_t are assumed to be independent of each other and also independent of the main error ε_t in (31). The stochastic slope β_t follows a random walk model.

Expansions of these two basic models for trend are possible but this is usually not needed.

Model for cyclical component

The cyclical component (cycle) is treated as either deterministic or stochastic, depending on how the model is specified. A cycle is usually represented by period, amplitude and phase. A deterministic cycle assumes time-invariant amplitude and phase during the consecutive fixed periods meaning that

the cyclical variations are repetitive and predictable. A model with stochastic cycle, on the other hand, is motivated by the fact that the cyclical variations usually vary over time influenced by random disturbances.

A deterministic cycle as a function of frequency λ , which is measured in radians, is expressed as a mixture of sine and cosine waves. This cycle depends on two parameters α and β , as shown here

$$\psi_t = \alpha \cos(\lambda t) + \beta \sin(\lambda t). \quad (34)$$

If t is measured on a continuous scale the amplitude will be $\omega = (\alpha^2 + \beta^2)^{1/2}$ and the phase is $\phi = \tan^{-1}(\beta/\alpha)$. This will lead to an equivalent formulation of the cycle in terms of the amplitude and phase as

$$\psi_t = \gamma \cos(\lambda t - \phi). \quad (35)$$

In most applications this pure deterministic form is not used. Instead, the cycle is usually built up recursively as a sum of cycles of different frequencies and amplitudes. This formulation leads to a stochastic cycle model of the following form

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos \lambda_c & \sin \lambda_c \\ -\sin \lambda_c & \cos \lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} v_t \\ v_t^* \end{bmatrix} \quad (36)$$

For more information about the properties of the specification and parameters in (36) see, e.g., Harvey (1990, pp. 38-39).

Modelling seasonal component

Seasonality in the context of STS-models is modelled in a way that allows for correction to the general trend of the series due to the periodic variations within a year. The simplest representation is a model of deterministic seasonality with the seasonal effect coefficients γ_t sum up to zero over a year. This model is described by the following expression

$$\gamma_t = \sum_{j=1}^s \delta_{jt} \gamma_j + \omega_t, \quad \omega_t \sim i.i.d. N(0, \sigma_\omega^2), \quad (37)$$

where s is the number of seasons in the year, ω_t is a random disturbance term with zero mean and variance σ_ω^2 and the dummy variable δ_{jt} is equal to one in season j , zero otherwise. This model may be extended to have coefficients that may also change over time which would lead to a model for stochastic seasonality. One such model is the model where each seasonal effect is modelled as a random walk process, as follows

$$\gamma_t = \gamma_{j,t-1} + \omega_{jt}, \quad \omega_{jt} \sim i.i.d. N(0, \sigma_{\omega_j}^2), \quad j = 1, \dots, s, \quad (38)$$

where the requirement that the seasonal components always sum to zero is accomplished by the restriction that the disturbance term sum to zero at each point in time.

Instead of dummy variables a model for seasonality may involve a set of trigonometric terms in a way similar to cycle representation in (36). A fixed seasonal pattern can also be modelled by a set of trigonometric terms at the seasonal frequencies $\lambda_j = 2\pi j / s$, $j = 1, \dots, [s/2]$, where $[s/2]$ implies rounding down to the nearest integer, leading to the following expression

$$\gamma_t = \sum_{j=1}^{[s/2]} (\alpha_j \cos \lambda_j t + \beta_j \sin \lambda_j t) . \quad (39)$$

A seasonal pattern in (39) may be allowed to evolve over time in a similar manner as the stochastic cycle in (36) which would lead to different extensions of the basic models for seasonality. See, e.g., Harvey (1990, pp. 40-42) or Harvey and Shephard (1993) for more details.

Modelling the Irregular Component

The structural dynamics of a response series y_t is captured by the previously explained components, such as trend, cycle, seasonal and regression effects. Hence the irregular component represents the unexplained remaining part in the series which corresponds to residual variation in an ordinary regression model. This residual variation might be treated in different ways from a very restrictive representation such as Gaussian white noise to far more complicated structures.

Statistical treatment: estimation, decomposition and diagnostic checking

In statistical sense the STS-models are usually treated through a general state space representation by using the Kalman filter algorithm. This generalisation allows treatment of both linear and non-linear form of a STS-model. Usually, a linear representation will be typical in the practical work since non-linear extensions are generally difficult to handle because of a huge variety of possible model specifications. An introduction to the linear and non-linear state space models is given in Durbin and Koopman (2001, Ch. 3 and Ch. 10, respectively).

The general linear Gaussian state space model for a time series y (or a set of time series \mathbf{y} with N elements) consists of a measurement equation and a transition equation, respectively:

$$y_t = Z_t \alpha_t + X_t \beta + \varepsilon_t , \quad (40.a)$$

$$\alpha_t = T_t \alpha_{t-1} + W_t \beta + R_t \eta_t . \quad (40.b)$$

See, e.g., Harvey and Shephard (1993, pp. 267-268) for details about (40). The observable variable y_t is related to a state vector α_t whose elements are not observable. However, the observations carry some information which can be estimated. This estimation is typically done by the Kalman filter, which is a recursive procedure for computing the optimal (in terms of the minimum mean square error) estimator of the state vector at time t . Hence, the state vector contains information about the unobserved components of time series y_t , such as seasonals, trend and irregulars. The estimation of all parameters is performed by the maximum likelihood method via the prediction error decomposition. Hence, the likelihood is evaluated by the Kalman filter using a numerical optimisation method for maximisation of likelihood.

See, e.g., Durbin and Koopman (2001, Ch. 2, 4 and 5) for details about the Kalman filter.

Since the state space form of structural models for time series is a model-based maximum likelihood approach it has many desirable statistical properties. As noted earlier, model selection does not rely on correlograms and related statistical devices in the way that the ARIMA model-based procedures do, which would imply differencing to obtain stationarity. This basically means that the variables and components are estimated in levels, which is an advantage in terms of interpretation of the estimated components.

The maximum likelihood approach within a state space framework provides a vehicle to make inference about the estimated components. It is relatively easy to make forecasts for each component with associated forecast errors since the mean square errors may be computed.

Hence, the most important step is actually the model selection with respect to modelling of each component in the general structural model. Harvey (1990, p. 13) discusses the most important criteria for a good modelling approach:

- a) Parsimony – a simpler model should be preferred to a more complicated one meaning that a model with a relatively small number of parameters should be preferred to a model with large number of parameters.
- b) Data coherence – the chosen model should provide a good fit to the data and the residuals should be approximately random.
- c) Consistency with prior knowledge – if there is any relevant information in economic theory or from any other relevant sources the model should be consistent with this information.
- d) Data admissibility – natural restrictions should be reflected in the model's ability to estimate and predict, e.g., model estimation for the variables that cannot be negative should not produce any negative value.
- e) Structural stability – good fit outside the sample is required.
- f) Encompassing – if a model is able to explain the results given by the rival formulation then it is said to encompass a rival formulation. This means that a rival model does not contain any information which could be used to improve the chosen model.

Once a plausible model is chosen the application of maximum likelihood and Kalman filter is straightforward but the technical details are less important in this context.

After estimation the diagnostic checking may be performed by using significant tests, usually based on three main assumptions concerning the residuals in the linear Gaussian state space models. These assumptions are independence, homoscedasticity and normality, which correspond to the general assumptions for a linear regression model. This is diagnosed by utilising the standardised prediction errors, as explained in, e.g., Commandeur et al. (2011, p. 9) or in Harvey and Koopman (1992).

Motivation for the use of STS- models

The two standard methods, TRAMO-SEATS and X-12-ARIMA, are widely used in official statistics and generally recommended by the European authorities. The reasons behind their popularity are natural. First of all these methods are relatively easy to interpret and implement in the statistical production since they are widespread across many well-supported IT-platforms. Furthermore, these methods have all necessary facilities that a modern seasonal adjustment procedure requires. Usually,

these two methods perform well in terms of short-run forecasting which makes them attractive to the policy makers.

However, when certain relatively strong assumptions for the underlying time series are not satisfied these modelling alternatives are likely to produce poor estimates of the related components. This is particularly true for the time series contaminated by many aberrant observations, the time series with strong moving seasonality or the data with other evident non-linearities. In some cases, the STS-models might be helpful since this framework may utilise varying coefficients for the moving seasonality problems or non-parametric methods (splines) to deal with non-linearities.

The usual assumption for a basic structural model is that the variance of each component is kept constant but it is possible to create a more complex extension which allows the trend component to be dependent on the business-cycle. The time-varying confidence intervals for seasonally adjusted estimates may be created in this way, as proposed in Koopman and Franses (2001).

One important issue regarding a standard seasonal adjustment from an ARIMA model-based procedure is how to treat calendar correction, especially with respect to estimation of moving holidays (such as Easter). Generally, the estimated parameters are held fixed as a result of ordinary least squares- or related estimation procedure. The state space framework permits these effects to vary over time which is practically impossible in the case with the competing methods. An example of a structural model involving the stochastic trading-day variations within a month is given in Dagum and Quenneville (1992).

Concerning the non-linear and non-Gaussian state space models, a detailed overview may be found in Durbin and Koopman (2001, Ch. 10). The STS-models within a state space framework can also tackle problems with temporal aggregation which is usually treated by a benchmarking procedure, as proposed in, e.g., Durbin and Quenneville (1997).

Furthermore, the structural state space models allows for a treatment of observations sampled at a higher frequency than monthly, meaning that the weekly, daily or even hourly observations can be treated within this framework. This is practically impossible with the two main competitors, TRAMO-SEATS or X-12-ARIMA. See, e.g., the study about the estimation of weekly seasonal pattern for the UK money supply in Harvey, Koopman and Riani (1997) or the estimation of hourly electricity data in Harvey and Koopman (1993). A treatment of different data irregularities, such as missing observations and observations at mixed frequencies, is illustrated in the study by Harvey and Cheung (2000) on the measurement of British unemployment.

Furthermore, if there is existence of complex relationships among different variables in a system of time series, a multivariate framework might be an alternative to a traditional univariate seasonal adjustment. Neither TRAMO-SEATS nor X-12-ARIMA have possibilities to treat multivariate time series. On the other hand, the univariate STS-models may relatively easily be extended to a multivariate framework. An overview of available software for state space models including an introduction into multivariate structural framework is given in Commandeur, Koopman and Ooms (2011). Different extensions to cope with more specific problems in a multivariate framework are described in, e.g., Koopman and Durbin (2000), Casals, Jerez and Sotoca (2002) and Birrell, Steel and Lin (2008).

The STS-models may also be extended to tackle the estimation problems with repeated overlapping sample survey. Pfeffermann (1991) proposes statistical treatment within this framework for estimation of population means based on rotating panel surveys when these surveys are overlapping. The proposed model allows for changes over time that might arise from an increase in sample size or a change in survey design. This framework permits a natural extension from a univariate STS-model to a multivariate STS-model, as described in Harvey and Shephard (1993). A monograph by Birrell (2008) gives a detailed description of a multivariate state space model tailored to the situation where a seasonally adjusted aggregate series is constructed by jointly modelling a set of sub-series.

Some conclusions

Obviously, the STS-models belong to a comprehensive framework suitable to almost any kind of time series analysis. Any ARIMA-model may be expressed in terms of a STS-model but the STS-models are much more than extensions of an ARIMA-modeling framework. They may involve State-Space approach, Bayesian approach, multivariate seasonal adjustment, non-linear models etc. The STS-models can handle data with irregular structure, the data with missing values and they are likely to be robust to different misspecifications.

Such flexibility may look attracting but these models have not been extensively used in official statistics. One reason for this is complexity of different modelling alternatives within this framework. Commonly, the methodological competence of an ordinary user at a national statistical office is rarely on a level required to understand the theoretical issues behind the procedures. Furthermore, the main-stream methods are likely to perform well for a large number of time series which quite naturally motivate for their use. And finally, complexity is not always easy to handle – not even for a specialist.

Anyhow, in some cases when the recommended procedures are not flexible enough to handle some deviations from the major assumptions they rely on, the STS-models might be helpful.

2.4 Step by step seasonal adjustment

The method of seasonal adjustment consists of several theoretical and practical issues which should be considered during the procedure in order to meet the expectations of experts and users. Although the modules “Seasonal Adjustment – Introduction and General Description” and “Seasonal Adjustment – Issues on Seasonal Adjustment” provide a comprehensive summary concerning the details, it is also necessary to describe the exact steps of the adjustment.

The following figure summarises the steps which are detailed in this subsection:

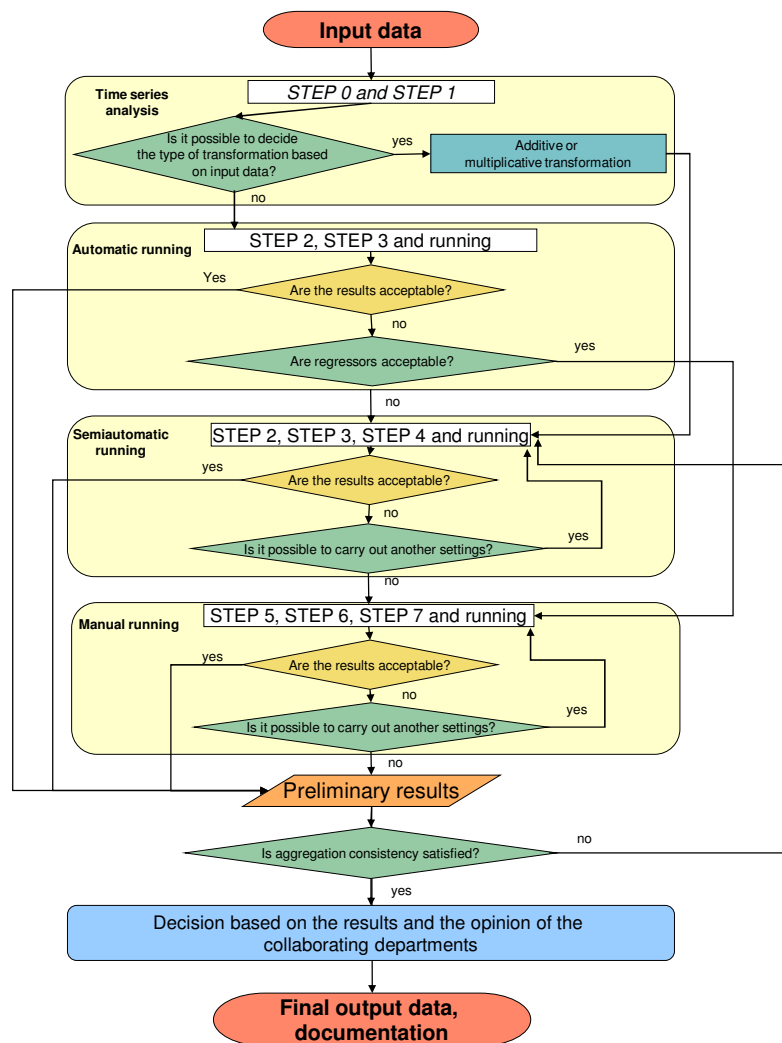


Figure 7: Steps of seasonal adjustment (source: HCSO)

STEP 0 – Examination of basic conditions and collection of expert information

Before the seasonal adjustment of a time series for the first time or the revision of the model and parameters, some basic properties of the given time series should be examined in order to achieve an adequate result:

- It is a software requirement (for both TRAMO-SEATS and X-12-ARIMA) for seasonal adjustment that the time series have to be at least 3 year-long (36 observations) for monthly series and 4 year-long (16 observations) for quarterly series. Naturally, these are minimum values; series can be longer for an appropriate adjustment. Series shorter than 3 years should not be seasonally adjusted by standard procedures, but in case of alternative, less standard procedures, it is possible (Hood, ECB (2003), EC (2005)). Special attention is necessary if series are 3-7 year-long as a result of instability problems. In this case, a general rule is to check the specification of the parameters several times per year (ESS guidelines). It is important to inform users about instability problems for short time series. However, if the time series is very long, the seasonal adjustment does not necessarily lead to higher quality because seasonality can change as time goes on. The sources of changes are the change in concepts,

definitions, methodology, legislative events, change of the weather, etc. If the series are not consistent for some reason, it might be better to shorten them for the purpose of identifying a more consistent seasonal pattern and to improve the decomposition. Another option in treating inconsistencies is to provide two separate time series, one for the latest period and one for an earlier period.

- Missing observation(s) in the time series should be identified. The identification is carried out, for example, via graphical analysis. Too many missing values in the given series lead to estimation problems in the adjustment. Thus, statisticians should substitute the missing observations with alternative data or statistical methods in the lack of original data.
- If series are part of an aggregate series, it should be verified that the starting and ending dates for all component series are the same.

If the aforesaid conditions hold, then preliminary expert information has to be collected about the

- calendar effects (trading/working day, leap year, moving holidays (e.g., Easter), national holidays)
- outliers
- seasonality
- methodological change of specialisation statistics
- methodological change of exterior factor (e.g., law, order)

Expert information is important, especially if the diagnostics of the adjustment are inconclusive (for example, outlier detection at the end of a time series) or in case of manual decomposition.

STEP 1 – Time series graphical analysis

Graphical analysis of the original time series provides useful information to the analyst because visual graphs help in identifying possible problems, quality issues and give relevant information to the process of seasonal adjustment.

Basic graphics

There are **basic graphs** by which possible problems in the data (such as outliers, zeros, negative values, missing observation(s) etc.), the structure of the trend-cycle or of the seasonal component are revealed or the presence of seasonality is examined.

Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend. The presence of seasonality is pre-condition of seasonal adjustment. Figure 8 illustrates a clear seasonal pattern.

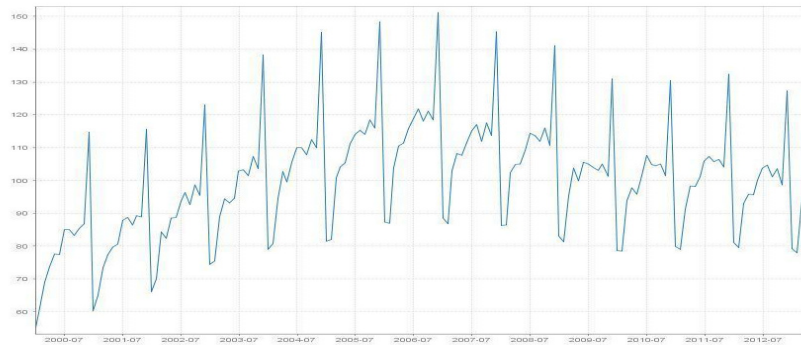


Figure 8: Hungarian monthly retail volume index, original series (source: HCSO)

Outliers could strongly affect the quality of the seasonal adjustment. The impact of these abnormal values could distort the estimation of components, therefore the seasonally adjusted series and the trend (the two most published and important data about seasonal adjustment) as well. In this part of the analysis, outlier identification and verification are carried out by addition of basic graphs and expert information. For example, if the graph of the original time series shows abrupt changes or there are data which do not fit in the past behaviour of the series, statistician should examine if these phenomena are valid (so they refer to the presence of outliers), or there are sign problems in the data, for example, captured erroneously. The two circled data do not fit in the past behaviour of the series. In such case, if there is an economic explanation behind the changes, data can be outlier.

The type of decomposition should be used automatically. Besides, there are situations when the diagnostics for choosing between decomposition schemes (models) are inconclusive. In this case one can choose to continue with the type of decomposition used in the past to allow for consistency between years, or if there is no experience about the past it is recommended to visually inspect the graph of the series.

- If the series has zero and negative value(s), or if the difference of the trend and the observed data is nearly constant in similar periods of time (months, quarters) irrespectively of the tendency of the trend, additive model is needed.
- If the series has a decreasing level with positive values close to 0, multiplicative model is considered.

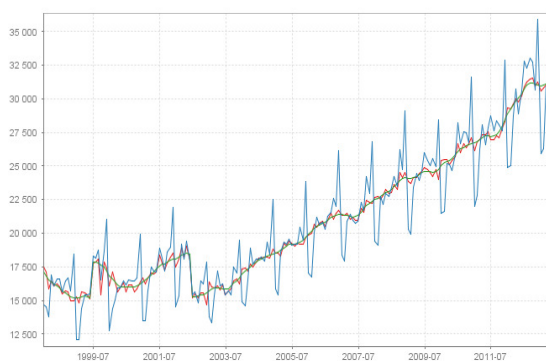


Figure 9: Additive decomposition

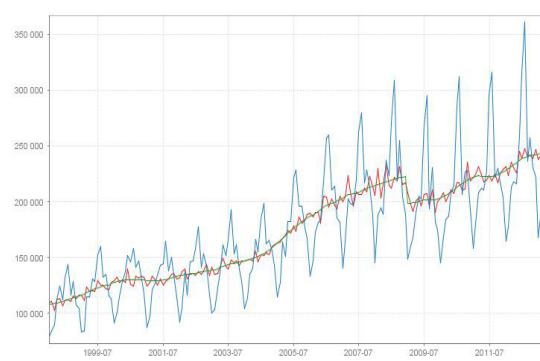


Figure 10: Multiplicative decomposition

Alternative approaches

Besides, there are more **sophisticated graphs**, such as spectrum or autocorrelograms which are two important tools of detecting seasonality and trading day effects in a time series. The peaks appearing in the spectrum indicate periodicity in the time series corresponding to the given frequency. Some frequencies are more important than others:

- *seasonal frequencies* show how many cycles of phenomenon are per year. For example, for monthly series the seasonal frequencies are (a whole period is represented by π): $\pi/6$, $\pi/3$, $\pi/2$, $2\pi/3$, $5\pi/6$, which are equivalent to 1, 2, ... cycle per year. Peaks at the seasonal frequencies indicate the presence of seasonality. Seasonality is a precondition for seasonal adjustment.

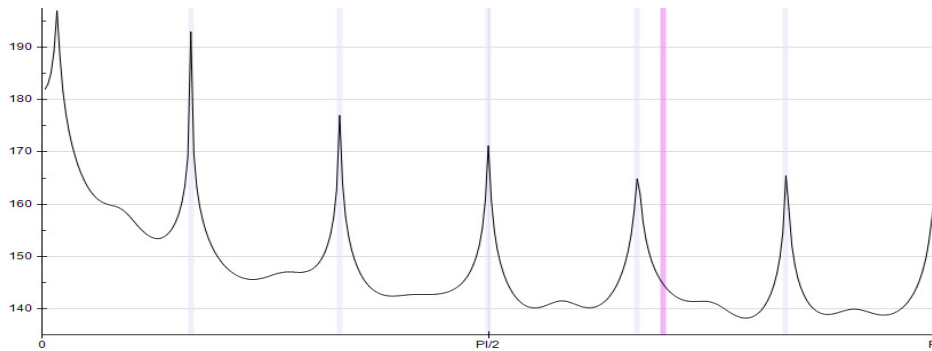


Figure 10: Auto-regressive spectrum of time series. Clear peaks at frequency $\pi/6$ and its multiples.

- peaks at trading days frequencies could occur due to inappropriate regression variables used in the model or the significant change of the calendar effect because the calendar effect cannot be modelled by fixed regression effect on the whole time series span.

Autocorrelation is the cross-correlation of a time series with itself. It is a mathematical tool for finding repeating pattern, to detect non-randomness in data, such as the presence of seasonality. In an autocorrelogram only positive and statistically significant autocorrelation at seasonal lags is important because of the concept of seasonal fluctuation. Figure 11 shows autocorrelogram of monthly time series. It is clear to see the significant autocorrelation at seasonal lags (12 and its multiples). In contrast with autocorrelogram, the partial autocorrelogram does not give reliable information about the presence of seasonality; its usefulness is to identify the ARIMA model.

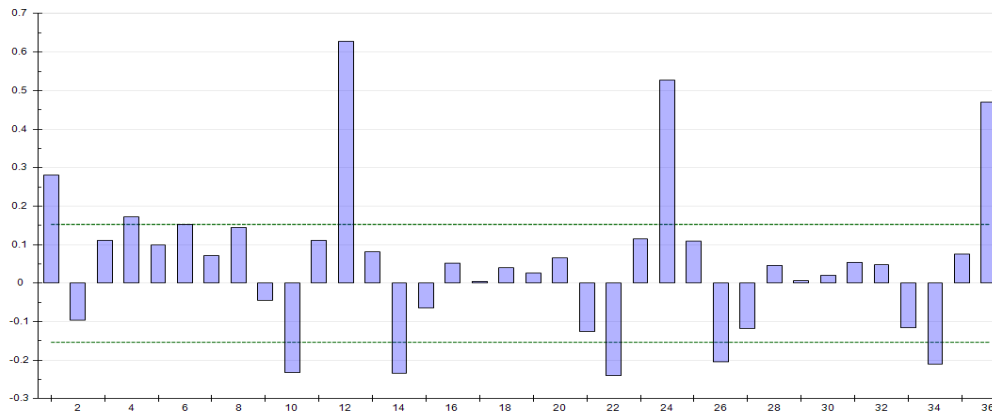


Figure 11: Autocorrelogram

The time series graphical analysis can be carried out by Eviews, R, SAS, Demetra+, JDemetra+, etc.

STEP 2 – Transformation

When the variance of the given time series is not constant, the series should be transformed in order to achieve stationary autocovariance function; hence, it stabilises the variance of the original time series. There are several ways of the transformation such as taking the logarithm, square root or differencing. The most commonly used is taking the logarithm. Log-transformation is offered by both TRAMO-SEATS and X-12-ARIMA. These software operate with automatic test which helps the user to choose between transformation types:

- no transformation → additive model is considered;
- log-transformation → log-additive model is used.

Confirm the results of the automatic choice by looking at the graphs of the series as it is described in STEP 1.

STEP 3 – Calendar adjustment

Calendar adjustment can be executed in a number of ways. One can distinguish between proportional and regression methods for adjustment. Under proportional approach the effects of trading days are estimated by counting the proportion of them on the month/quarter. Under the regression approach the effects of trading days are estimated in a regression framework. If possible, the proportional approach should be avoided – especially in case of model-based methods. The most recent and widely used seasonal adjustment tools (TRAMO-SEATS, X-12-ARIMA, X-13-ARIMA-SEATS) perform calendar adjustment by regression method, called reg-ARIMA. Under the reg-ARIMA approach it should be determined which regression effects (trading/working day, leap year, moving holidays) and national holidays are plausible for the series.

If an effect is not plausible for the series or the coefficients for the effect are not significant, then regressor should not be fit for the effect, it should be eliminated. Exception can be made in case of trading day regressors (see 2.1.1.).

If the coefficients for the effects are marginally significant then it should be determined if there is a reason to keep the effects in the model. For example, if there are some kinds of economic explanations behind the effects, they should be retained.

It is important to distinguish between seasonal and non-seasonal component of calendar effects, since the seasonal part of calendar effects is eliminated by the seasonal adjustment filters under the decomposition procedure of time series (see STEP 6). Therefore, under calendar adjustment within the pre-treatment of seasonal adjustment only the non-seasonal part of the effects has to be dealt with.

Seasonal adjustment approaches of the Demetra software family (TRAMO-SEATS, X-12-ARIMA, X-13-ARIMA-SEATS) automatically create appropriate calendar regression variables depending on the chosen specification. However, the user may need to change the automatic options, for example, for chaining two calendars for two different time periods or modifying the calendar regression variables to match the national holidays which differ from the previous options of the used software. Sometimes the automatic test does not indicate the need for trading day regressor, but if there is a peak at the first trading day frequency of the spectrum of the residuals, then one may fit a trading day regressor manually.

STEP 4 – Outliers

The presence of these abnormal values distort of the seasonal and calendar components because seasonal adjustment methods are usually based on linear models (e.g., reg-ARIMA). Therefore, outliers should be identified and removed before seasonal adjustment is carried out. Besides, they give information about some specific events (like strikes, etc.), so valid outliers should be reintroduced after the adjustment.

There are two possibilities to identify outliers. The first is when we identify series with possible outlier values by looking at graphs of the original series and any available information (economic, social, etc.) about the possible cause of the detected outlier, as in STEP 1. Since seasonal adjustment is carried out by software in practice, this direction is in service as an additional opportunity generally to the cheque of the automatic outlier detection. Therefore, the second possibility what we use is automatic outlier detection and correction. Outlier detection is always carried out automatically when time series are seasonally adjusted for the first time.

Outlier coefficients may be statistically non-significant when time series are already seasonally adjusted and reg-ARIMA models are revised (generally once in a year). In this case, the user has to decide whether to keep them in the model. There are criteria, for example, coherence with past decisions, based on which we may come to our decision..

The reliability of the seasonal adjustment depends on the number of outliers. A large number of outliers relative to the length of the series could result in over-specification of the regression model. Furthermore, it signifies if there is a problem related to weak stability of the process, or if there is a problem with the reliability of the data (for example, data captured erroneously). Shortening the time span or changing the critical value of the statistical tests may help in better modelling of outliers.

It is important to stress the treatment of outliers at the end of the series. For example, the change of the type of outlier later may lead to large revisions. In this case expert information is especially important because the type of outliers at the end of the series are uncertain, as real extraordinary economic effects are often unknown, and there is no information on what happens after the latest outlier appears. For instance, the level shift is indistinguishable from an additive outlier in this case, since we do not know how the level of the series will behave. Therefore, to collect external information on the event in question is very useful. It would help to define the type of outliers at the end of the series.

STEP 5 – ARIMA model

In the most widely used software – TRAMO-SEATS and X-12-ARIMA – seasonal adjustment are based on ARIMA-model methodologies. Automatic model identification usually produces satisfactory models. But there are cases when results are not plausible. Therefore, manual identification may be justified. Another situation when different ARIMA models could fit in the same series. In this case it is recommended by most of the statisticians to choose the simplest model with the smallest number of parameters with a satisfactory fit. This is better than a high-order model. During manual procedure, it is advisable to identify the not significant high-order ARIMA model coefficients and reduce the order of the model, taking care not to skip lags of autoregressive models. For moving average models, it is not necessary to skip model lags whose coefficients are not significant. Before choosing an MA model with skipped lag, the full-order MA model should be fitted and skip a lag only if that lag's model coefficient is not significantly different from zero.

Another situation in which manual identification may be justified is when automatic identification produces a model which, while satisfying the tests, still has some unsatisfactory features. For example, some individual significant correlation at fairly low lags, although the combined test on the serial correlations of the residuals passed. In this case it could be worth adding an extra coefficient at the appropriate lag to the AR or MA component. If the extra coefficient is significant and the significant serial correlation has been removed, the extra term may be justified.

The model identification statistics, particularly the BIC and the AIC, are useful tools in confirming the global quality of fitting statistics. The application of information criteria may help in choosing among different models.

STEP 6 – Decomposition

The last step of the pre-treatment procedure for seasonal adjustment is to decompose the original time series into different components: trend-cycle, seasonal and irregular component. Depending on the nature of seasonality components, several different schemes can be connected. The most frequently used schemes (models) are the following:

- *additive decomposition* (Figure 9), when the magnitude of seasonal effects does not change as the level of the trend-cycle changes. Also, any series with zero or negative values are additive. In this case, components are linked additively.
- the *multiplicative decomposition* implies that as the trend of the series increases, the magnitude of the seasonal spikes also increases (Figure 10). For multiplicative decomposition, components are linked through multiplication. The decomposition scheme of the most economic time series is multiplicative.
- *log-additive* scheme is to specify an additive model on the logarithm of the time series. Based on this fact, one of its main advantages is that the multiplicative model can be transformed to additive model, which is more manageable. Therefore, multiplicative and log-additive model are frequently considered identical.

Before the decomposition of a time series, some modifications should be carried out on it. It is required to determine and remove deterministic effects such as outliers or the non-seasonal part of calendar effects, because the adjustment is distorted in case of their presence. After removing the

deterministic part we get the purely stochastic part of the series. This is the autocorrelated disturbance of the deterministic part. It is decomposed by filters based on linear stochastic models. We get the final component of the series, if regression effects (deterministic part of the series) are reintroduced in the components according to their nature.

STEP 7 – Quality diagnostics

The procedure of seasonal adjustment is very complex, so the accurate monitoring of the results before disclosure and disseminated is very important. A wide range of quality measures are available to ensure the best quality.

Quality diagnostics of seasonal adjustment can divide into three main parts. The three issues are the following:

- model adequacy and diagnostics on the model residuals
- residual trading day effects and seasonality in both the seasonally adjusted series and the irregular component
- stability analysis

In the first part of monitoring the results it can be examined if the model used for the adjustment is adequate. At ARIMA modelling, the principal tool for assessing model adequacy is the widely-used Ljung-Box statistic, built from autocorrelations of residuals. Of particular interest are autocorrelations at low lags, say 1 to 4, and at seasonal lags 12 and 24. Low Ljung-Box p-values (below .05) at lags 12 and 24 result from one or more high residual autocorrelations and indicate model inadequacy. Monitoring of the seasonal moving average parameter is also important. When it is close to -1, the seasonal factors are highly stable; when it is close to 0, the factors tend to change rapidly. Series graphs and knowledge of the series can help assess how much movement in the seasonal is desirable.

After seasonal adjustment, we can check for residual seasonality and residual calendar effects using spectral graphics of the decomposed seasonally adjusted series and the irregular component. Peaks at seasonal frequencies of the adjusted series mean that the filters used in the decomposition are not well adapted to the series or to a large part of it. Peaks at the trading day frequencies could indicate that the regression variables of the model do not suit well the series or that the calendar effects change too much to be captured by the fixed regression effects applied for the whole duration of the series. If remaining seasonality is present one has to reconsider the model specification, the regression variables or the time span used for modelling.

Careful assessment of the seasonally adjusted data includes analysis about the stability of the seasonal component. The software reports several stability diagnostics such as statistical tests or graphical diagnostics. Revision history and sliding spans are the most commonly used stability diagnostics.

Revision history analyses what kinds of revisions are caused by adding new observations at the end of the series. It presents charts both for the seasonally adjusted and trend-cycle series. On Figure 12 each circle depicts the initial adjustment when this point is the last observation. The curve presents the final results. The closer the initial observation dots to the curve based on all available observations, the better the quality. Revision history table is also available in this part. This table presents the differences between the first estimates and the last estimates for the last four years. If some

observations are exceeded the given critical limit, it should be examined whether the adjustment is unsatisfactory or these abnormal values are, in fact, outliers.

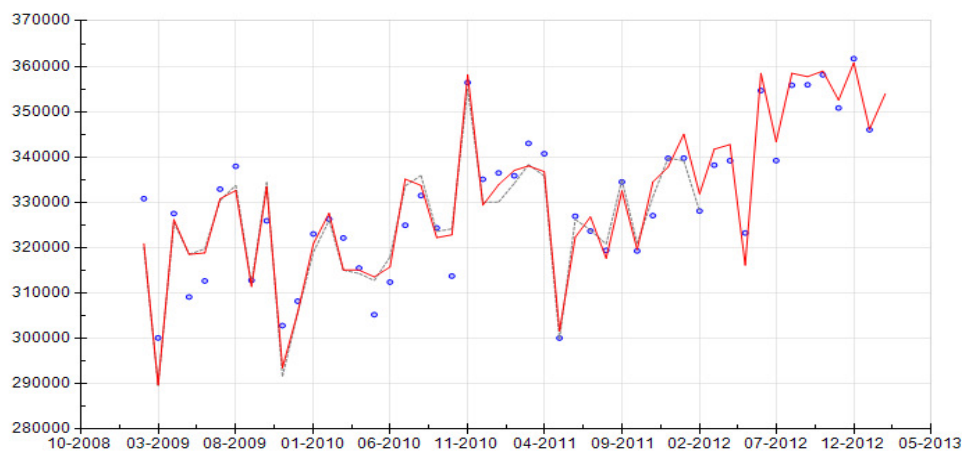


Figure 12: Revision history

Another very important tool for stability analysis is the sliding spans. It is particularly useful for a series with a large number of outliers or changes in seasonality. It depicts period-to-period changes. The results are stable if one cannot consider values exceeding a three per cent threshold. Any larger value is unstable. Figure 13 shows stable seasonal factors since none of the values exceeds three per cent.

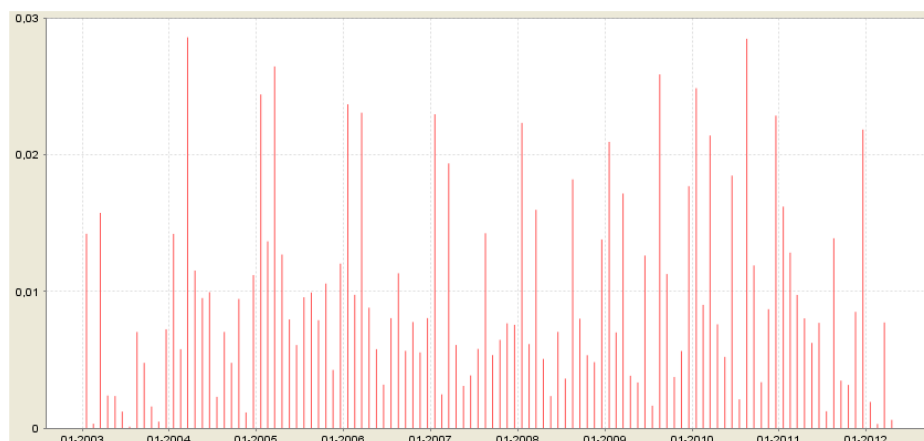


Figure 13: Sliding spans analysis

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Akaike, H. (1980), Seasonal Adjustment by a Bayesian Modeling. *Journal of Time Series Analysis* **1**, 1–13.
- Attal-Toubert, K. and Ladiray, D. (2011), Trading-Day Adjustment as a Practical Problem. *Proceedings of the 58th World Statistical Congress, 2011, Dublin*.
<http://2011.isiproceedings.org/papers/650329.pdf>
- Bell, W. R. (1984), Signal Extraction for Nonstationary Time Series. *Annals of Statistics* **12**, 646–664.
- Bell, W. R. (1984a), *Seasonal Decomposition of Deterministic Effects*. U.S. Bureau of the Census, Statistical Research Division, Report Number: Census/SRD/RR-84/01.
<http://www.census.gov/srd/papers/pdf/rr84-1.pdf>
- Bell, W. R. (1995), *Correction to Seasonal Decomposition of Deterministic Effects*. U.S. Bureau of the Census, Statistical Research Division, Report Number: Census/SRD/RR-95/01.
<http://www.census.gov/srd/papers/pdf/rr95-01.pdf>
- Bell, W. R. and Hillmer, S. C. (1984), Issues involved with the Seasonal Adjustment of Economic Time Series. *Journal of Business and Economic Statistics* **2**, 291–320.
- Bell, W. R., Holan, S. H., and McElroy, T. S. (eds.) (2012), *Economic Time Series: Modeling and Seasonality*. CRC Press, New York.
- Bell, W. R. and Martin, D. E. K. (2004), Computation of Asymmetric Signal Extraction Filters and Mean Squared Error for ARIMA Component Models. *Journal of Time Series Analysis* **25**, 603–625.
- Birrell, C., Steel, D. G., and Lin, Y. X. (2008), *Seasonal Adjustment of Aggregated Series using Univariate and Multivariate Basic Structural Models*. Centre for Statistical & Survey Methodology, University of Wollongong, Australia.
- Birrell, C. (2008), Efficiency gains for seasonal adjustment by joint modelling of disaggregated series. *University of Wollongong Theses Collection*, University of Wollongong, Australia.
- Box, G. E. P., Hillmer, S. C., and Tiao, G. C. (1978), Analysis and modeling of seasonal time series. In: Zellner, A. (ed.), *Seasonal Analysis of Economic Time Series*, U.S. Dept. of Commerce - Bureau of the Census, Washington, D.C., 309–334.
- Box, G. E. P. and Jenkins, G. M. (1970), *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.
- Box, G. E. P. and Tiao, G. C. (1975), Intervention Analysis with Applications to Economics and Environmental Problems. *Journal of the American Statistical Association* **70**, 70–79.

- Burman, J. P. (1980), Seasonal Adjustment by Signal Extraction. *Journal of the Royal Statistical Society, Series A* **143**, 321–337.
- Casals, J., Jerez, M., and Sotoca, S. (2002), An Exact Multivariate Model-Based Structural Decomposition. *Journal of the American Statistical Association* **97**, 553–564.
- Chen, C. and Liu, L. (1993), Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association* **88**, 284–297.
- Cleveland, R. B., Cleveland, W. S., and McRae, J. E. (1990), STL: A Seasonal-Trend Decomposition Procedure Based on LOESS. *Journal of Official Statistics* **6**, 3–73.
- Commandeur, J. J. F., Koopman, S. J., and Ooms, M. (2011), Statistical Software for State Space Methods. *Journal of Statistical Software* **41**, 1–18.
- Dagum, E. B. (1980), The X-11-ARIMA seasonal adjustment method. Statistics Canada.
- Dagum, E. B., Quenneville, B., and Sutradhar, B. (1992), Trading-Day Variations Multiple Regression Models with Random Parameters. *International Statistical Review* **60**, 57–73.
- Durbin, J. and Koopman, S. J. (2001), *Time Series Analysis by State Space Models*. Oxford University Press, Oxford, UK.
- Durbin, J. and Quenneville, B. (1997), Benchmarking by State Space Models. *International Statistical Review* **65**, 23–48.
- Engle, R. F. (1978), Estimating Structural Models of Seasonality. In: Zellner, A. (ed.), *Seasonal Analysis of Economic Time Series*, U.S. Dept. of Commerce - Bureau of the Census, Washington, D.C., 281–297.
- Eurostat (2009), *ESS guidelines on seasonal adjustment*. European Communities, Luxembourg.
- Findley, D. F. (2009), *Stock Series Holiday Regressors Generated By Flow Series Holiday Regressors*. Statistical Research Division Research Report Series (Statistics #2009-04), U.S. Census Bureau. <http://www.census.gov/srd/papers/pdf/rrs2009-04.pdf>
- Findley, D. F. and Monsell, B. C. (2009), Modelling Stock Trading Day Effects Under Flow Day-of-Week Effect Constraints. *Journal of Official Statistics* **25**, 415–430.
- Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., and Chen, B. C. (1998), New capabilities of the X-12-ARIMA seasonal adjustment program (with discussion). *Journal of Business and Economic Statistics* **16**, 127–177. <http://www.census.gov/ts/papers/jbes98.pdf>
- Fischer, B. (1995), *Decomposition of Time Series - Comparing Different Methods in Theory and Practice*. Eurostat, Luxembourg.
- Gomez, V. and Maravall, A. (2001a), Automatic modeling methods for univariate series. In: D. Peña, G. C. Tiao, and R. S. Tsay (eds.), *A Course in Time Series Analysis*. John Wiley and Sons, New York, NY.
- Gomez, V. and Maravall, A. (2001b), Seasonal adjustment and signal extraction in economic time series. In: D. Peña, G. C. Tiao, and R. S. Tsay (eds.), *A Course in Time Series Analysis*. John Wiley and Sons, New York, NY.

- Harvey, A. C. and Cheung, C-H. (2000), Estimating the Underlying Change in Unemployment in the UK (with discussion). *Journal of the Royal Statistical Society, Series A* **163**, 303–339.
- Harvey, A. C. (1990), *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.
- Harvey, A. C. and Shephard, N. (1993), Structural Time Series Models. In: G. S. Maddala, C. R. Rao, and H. D. Vinod (eds.), *Handbook of Statistics*, Elsevier Science Publishers, 261–302.
- Harvey, A. C. and Koopman, S. J. (1992), Diagnostic Checking of Unobserved Components Time Series Models. *Journal of Business & Economic Statistics* **10**, 377–389.
- Harvey, A. C., Koopman, S. J., and Riani, M. (1997), The Modeling and Seasonal Adjustment of Weekly Observations. *Journal of Business and Economic Statistics* **15**, 354–368.
- Harvey, A. C. and Koopman, S. J. (1993), Forecasting Hourly Electricity Demand Using Time-Varying Splines. *Journal of American Statistical Association* **88**, 1228–1236.
- Harvey, A. C. and Todd, P. H. J. (1983), Forecasting Economic Time Series with Structural and Box-Jenkins Models: a Case Study. *Journal of Business and Economic Statistics* **1**, 299–306.
- Henderson, R. (1916), Note on Graduation by Adjusted Average. *Transactions of the American Society of Actuaries* **17**, 43–48.
- Hendry, D. F. (1995), *Dynamic econometrics*. Oxford University Press, Oxford.
- Hillmer, S. C. and Tiao, G. C. (1982), An ARIMA Model Based Approach to Seasonal Adjustment. *Journal of the American Statistical Association* **77**, 63–70.
- Hylleberg, S. (ed.) (1992), *Modelling Seasonality*. Oxford University Press, Oxford.
- Kaiser, R. and Maravall, A. (2001), *Measuring Business Cycles in Economic Time Series*. Springer, Berlin.
- Kaiser, R. and Maravall, A. (2003), Seasonal outliers in time series. Special issue on Time Series, Estadística, *Journal of the Inter-American Statistical Institute* **15**, 101–142.
- Koopman, S. J. and Durbin, J. (2000), Fast filtering and smoothing for multivariate state space models. *Journal of Time Series Analysis* **21**, 281–296.
- Koopman, S. J. and Franses, P. H. (2001), *Constructing seasonally adjusted data with time-varying confidence intervals*. Econometric Institute, Erasmus University, Rotterdam, Netherlands.
- Ladiray, D., and Quenneville, B. (2001), *Seasonal Adjustment With the X-11 Method*. Lecture Notes on Statistics, Springer-Verlag, New York.
- Maravall, A. (2012), Statistical and Econometrics Software. Banco de España.
http://www.bde.es/bde/en/secciones/servicios/Profesionales/Programas_estadi/Programas_estad_d9fa7f3710fd821.html
- Maravall, A. (2008), Notes on Programs TRAMO and SEATS. Bank of Spain.
http://www.bde.es/webbde/en/secciones/servicios/Profesionales/Programas_estadi/Notas_introduccion_3638497004e2e21.html

- Maravall, A. (1987), On Minimum Mean Squared Error Estimation of the Noise in Unobserved Component Models. *Journal of Business and Economic Statistics* **5**, 115–120.
- Maravall, A. and Perez, D. (2012), Applying and Interpreting Model-Based Seasonal Adjustment. The Euro-Area Industrial Production Series. In: W. R. Bell, S. H. Holan, and T. S. McElroy (eds.), *Economic Time Series: Modeling and Seasonality*, CRC Press, New York.
- Pankratz, A. (1991), *Forecasting with dynamic regression models*. John Wiley and Sons, New York.
- Pfeffermann, D. (1991), Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys. *Journal of Business and Economic Statistics* **9**, 163–176.
- Pierce, D. A. (1979), Signal Extraction Error in Nonstationary Time Series. *Annals of Statistics* **7**, 1303–1320.
- R Development Core Team (2012a), signalextraction: Real-Time Signal Extraction (Direct Filter Approach). *The Comprehensive R Archive Network*.
<http://cran.r-project.org/web/packages/signalextraction/signalextraction.pdf>
- R Development Core Team (2012b), *STL- Seasonal Decomposition of Time Series by Loess*. Vienna, Austria.
- Roberts, C. G., Holan, S. H., and Monsell, B. (2009), *Comparison of X-12-ARIMA Trading Day and Holiday Regressors With Country Specific Regressors*. U.S. Bureau of the Census, Statistical Research Division. <http://www.census.gov/ts/papers/rrs2009-07.pdf>
- SAS Institute (2009), *SAS/IML 9.2 User's Guide*.
- United States Census Bureau (2012), X-12-Arima Seasonal Adjustment Program. U.S. Census Bureau. <http://www.census.gov/srd/www/x12a/>
- Wildy, M. (2005), *Signal Extraction*. Springer, New York.
- Zellner, A. (ed.) (1978), *Seasonal Analysis of Economic Time Series*. Proceedings of a Bureau of the Census, NBER and ASA Conference. U.S. Department of Commerce, Bureau of the Census, Washington, D.C.

Specific section

8. Purpose of the method

The method discusses the theoretical background of seasonal adjustment with a comprehensive summary of methodological principles and the steps of the adjustment process: the main focus of this module is put on description of the decomposition based on ARIMA models, on moving averages and on STS-models, while the other classes of models are not treated.

9. Recommended use of the method

1. ARIMA: A particularly important part of seasonal adjustment is the identification of ARIMA models. This tool, as discussed by Box and Jenkins (1976), represents a practical way of dealing with moving features of seasonal time series.
2. STS: Any ARIMA-model may be expressed in terms of a STS-model but the STS-models are much more than extensions of an ARIMA-modelling framework. They may involve State-Space approach, Bayesian approach, multivariate seasonal adjustment, non-linear models, etc. The STS-framework can handle data with irregular structure, the data with missing values and they are likely to be robust to different misspecifications.

10. Possible disadvantages of the method

1. ARIMA: An ARIMA-based decomposition requires a preparatory step including reg-ARIMA- or TRAMO procedures to clean the data from irregularities. In this step differencing of time series to achieve stationarity is almost always imposed resulting in the loss of degrees of freedom. However, for some very noisy series the stationarity can not be achieved in this way, not even if differencing is performed several times. Hence, applying this approach would result in a relatively bad estimates of the so called de-noised series (the error from reg-ARIMA procedure). As this de-noised series is the one to be decomposed into the seasonal effect, the trend-cycle and the irregular component, such an approach which would in turn lead to a large uncertainty in the estimated components.
2. STS: This method is really flexible and robust, but these models have not been extensively used in official statistics as a result of the complexity of different modelling alternatives within this framework. Furthermore, the main-stream methods are likely to perform well for a large number of time series which quite naturally motivate for their use. And finally, complexity is not always easy to handle – not even for a specialist.

11. Variants of the method

1. ARIMA
2. STS

12. Input data

The original time series before seasonal adjustment.

13. Logical preconditions

In this module, this point is not relevant.

14. Tuning parameters

Not relevant.

15. Recommended use of the individual variants of the method

It is discussed in point 9.

16. Output data

The output data contains the results of seasonal adjustment: the components of time series after decomposition and the elimination of irregularities, and the adjusted time series.

17. Properties of the output data

Not relevant

18. Unit of input data suitable for the method

Not relevant

19. User interaction - not tool specific

Not relevant

20. Logging indicators

Not relevant

21. Quality indicators of the output data

The quality indicators represent the adequacy of the seasonal adjustment process. A primary purpose is identifying the available best model. The statistical tests such as Ljung-Box and Box-Pierce tests offer the opportunity to examine the adequacy of the chosen model. The robustness is also essentially important, which may be studied via sliding spans.

22. Actual use of the method

Discussed in point 9 and 10

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Seasonal Adjustment – Introduction and General Description
2. Seasonal Adjustment – Issues on Seasonal Adjustment

24. Related methods described in other modules

- 1.

25. Mathematical techniques used by the method described in this module

1.

26. GSBPM phases where the method described in this module is used

1. GSBPM Phase 6.1, 6.2, 6.3

27. Tools that implement the method described in this module

1.

28. Process step performed by the method

Administrative section

29. Module code

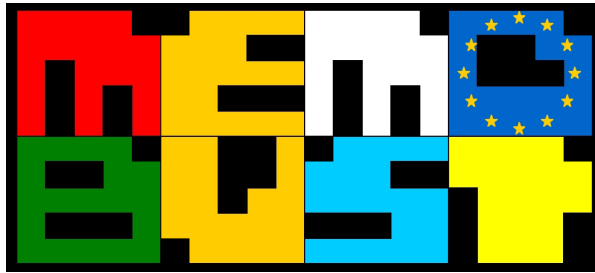
Seasonal Adjustment-M-SA of Economic Time Series

30. Version history

Version	Date	Description of changes	Author	Institute
0.0.1	01-02-2013	STS models	Suad Elezović	SCB Sweden
0.0.2	01-02-2013	step by step	Orsolya Kocsis	HCSO
0.0.3	02-04-2013	decomposition	Anna Ciammola	ISTAT
0.0.4	27-05-2013	ad hoc filters	Oyvind Langsrud	Statistics Norway
0.0.5	08-07-2013	reg-ARIMA	Anna Ciammola	ISTAT
0.0.6	02-09-2013	glossary	Anna Ciammola	ISTAT
0.0.7	18-09-2013	step by step	Orsolya Kocsis, László Sajtos	HCSO
0.0.8	20-09-2013	STS models	Suad Elezović	SCB Sweden
0.0.9	04-10-2013	new subsection 2.4 and other changes	Suad Elezović, Yingfu Xie	SCB Sweden
0.2	22-11-2013	specific section	László Sajtos	HCSO
0.2.1	11-12-2013	preliminary release		
0.3	20-12-2013	minor improvements based on EB-review	László Sajtos	HCSO
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:28



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Issues on Seasonal Adjustment

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Consistency issues	3
2.2 Special Issues	9
2.3 Treatment of the crisis	11
2.4 Data presentation, communication with users, and documentation	13
3. Design issues	14
4. Available software tools.....	14
5. Decision tree of methods	14
6. Glossary.....	15
7. References	15
Interconnections with other modules.....	17
Administrative section.....	18

General section

1. Summary

Seasonal adjustment, which is a routine activity in statistical offices nowadays, and the connected mathematical background have been a subject of theoretical investigations for several decades. However, methods and tools of seasonal adjustment are still under development and perpetual debates focus on them. Furthermore, there is significant flexibility regarding applied adjustment settings and model selection, which may lead to subjective and ambiguous results. As the number of the series to be adjusted is rapidly increasing and the quality of official seasonally adjusted data is increasingly important, the need of recommendations and guidelines is indisputable. ESS Guidelines on Seasonal Adjustment (2009) can be regarded as benchmark working material in this topic.

The goal of this module is to discuss important issues on seasonal adjustment, providing a guide on how to deal with them describing practices and giving some references to achieve further information. One of the most essential issue is providing temporal and cross-sectional consistency of time series. Although forcing consistency may hold disadvantages, it may be required to fulfil accounting constraints (as in the quarterly national accounts). Owing to statistical investigations and the available significant computer resources, nowadays, this task is much less demanding than it was a few years ago.

The module also focuses on the choice between indirect or direct approach to seasonally adjust time series derived as aggregation of other component time series. Choosing between these approaches is not obvious, comprehensive analyses have been performed in order to eliminate uncertainty. Practices and guidelines are described in the related subsection. Revision is also a crucial element of the seasonal adjustment: the updating of unadjusted data and the use of bilateral filters lead to revise the seasonally adjusted data previously released undermining the credibility of the producer agencies.

The financial crisis seriously undermined the reliability of the results of seasonal adjustment. The seasonal pattern and the behaviour of time series may change significantly. Therefore, it is necessary to take the impacts of the crisis into consideration. Although, this aim is available through outliers and ramp effect, monitoring time series is also essential part of treatment.

Seasonal adjustment and the comparison with raw data are often subject of confusion among users. Consequently, the details of publication and communication policy is essential to prevent misunderstanding.

2. General description

2.1 Consistency issues

2.1.1 General introduction

Most of time series belong to a system of series classified by attributes. For example, labour force series are classified by province, age, sex, part-time and full-time employment; short-term statistics on industry (production, turnover, etc.) are aggregated according to the classification of economic activities. Therefore, economic statistics are often linked by a system of relationships (for example accounting), thus constraints should be satisfied (for instance, GDP as balance of the uses and

resources account). Due to the different sample surveys, the ways of collecting data or the measuring equipment, it is challenging to ensure the consistency between the constraints and the observed variables. However, the discrepancies are usually the basis of confusion among users and criticism, according to Quenneville and Rancourt (2005). The adjustment of a set of data in order to satisfy a number of restrictions and remove any discrepancy is generally known as *reconciliation*. This method is entitled *balancing* in the aspect of national accounts (Di Fonzo and Marini, 2009; Dagum and Cholette, 2006). The statistical institutions are required to face this problem because they are often obliged to publish consistent sets of time series to fulfil legal regulations or common practices on statistics set by international institutions (UN, IMF, etc.).

The related restrictions can be of two types:

- *temporal* constraints: these require the consistency between the low-frequency aggregates and the high-frequency adjusted series.
- *contemporaneous(cross-sectional)* constraints, which assume the form of linear combinations of the variables which should be fulfilled in every observed period. In other words, this constraint requires that the values of the component elementary series add up to the marginal totals for each period of time (Dagum and Cholette, 2006). For example, if the system is classified by M industries (or sectors) and W provinces, the system must satisfy M sets of industrial cross-sectional constraints over the industries in each province.

The elimination of discrepancies between and within variables are handled by methods based on similar principles. The process of adjustment in time dimension is called *benchmarking* (or temporal disaggregation), while the former type of reconciliation is known as the balancing problem (Di Fonzo and Marini, 2009). These methods and the background are discussed in the following subsections.

2.1.2 Time consistency, benchmarking, and related techniques

Problem description

It is essential to provide the consistency of, for instance, sub-annual and annual industrial time series, or quarterly national accounts and the annual accounts in order to present clear view about the economy. This is entitled *time consistency*. The absence of it may confuse users.

Collecting large volume of comprehensive data with high accuracy is really expensive. As a result of this fact, annual or ten-yearly enterprise census provides them. These data are referred to benchmark and may be the basis of annual data. More frequent data, such as quarterly national accounts, also play important role in the economical and statistical system. However, they are less accurate compared with the comprehensive data as a result of the different sources of quarterly and annual data, sampling error, etc.

In general, *benchmarking* refers to techniques used to ensure coherence between time series data of the same target variable measured at different frequencies, for example, sub-annually and annually (Publications of Statistics Canada, 2009). The benchmarking problem arises because the annual sums of the sub-annual series are not equal to the corresponding annual values (due to the factors described above). In other words, there are annual discrepancies (d_m) between the annual benchmarks and the sub-annual values (Dagum and Cholette, 2006):

$$d_m = a_m - \sum_{t=t_{1m}}^{t_{Lm}} j_{mt} s_t, \quad m=1, \dots, M$$

where t_{1m} and t_{Lm} are respectively first and last, sub-annual periods, t covered by the m -th benchmark, e.g., quarters 1 to 4, 5 to 8, and the j_{mt} s are the coverage fractions here assumed to be equal to 1, a_m , $m=1, \dots, M$ refers to the annual series and the sub-annual series are denoted by s_t , $t=1, 2, \dots, T$. $\{1, 2, \dots, T\}$ refers to a set of contiguous months, quarters, days, etc., and $\{1, 2, \dots, M\}$ refers to a set of not necessarily contiguous periods, e.g., there may not be a benchmark every “year” (as it is described above). In some cases, benchmarks are available every second year, or even irregularly.

It is also important to note that the discrepancies are often expressed in terms of proportional discrepancies:

$$d_m = \frac{a_m}{\left(\sum_{t=t_{1m}}^{t_{Lm}} j_{mt} s_t \right)}, \quad m=1, \dots, M$$

Benchmarking also plays role in case of seasonal adjustment. In fact, seasonally adjusting monthly or quarterly time series causes discrepancies between the yearly sum of the unadjusted data series and the corresponding yearly sums of the seasonally adjusted series (Dagum and Cholette, 2006). There are several disadvantages of constrained equality in the annual sum, such as bias in the seasonally adjusted data or the non-optimality of the final seasonally adjusted data, see ESS Guideline (2009). As a consequence the application of any constraint should be avoided.

When users insist on temporal consistency or accounting constraints have to be fulfilled, seasonally adjusted series are then benchmarked to the yearly sums of the unadjusted series or to the yearly sum of the calendar-adjusted series, if significant calendar effects are present.

2.1.3 Methods to achieve time consistency: benchmarking methods

The method of benchmarking operates with the sum of modified sub-annual series in order to be equal to the corresponding benchmark. The formulation is the following:

$$a_m - \sum_{t=t_{1m}}^{t_{Lm}} j_{mt} \hat{\theta}_t = 0, \quad m=1, \dots, M$$

where $\hat{\theta}_t$ is the benchmarked series. Several benchmarking methods are available. The simplest ones are the *prorating* and the *Denton method*, which are widely known.

Prorating method (Dagum and Cholette, 2006)

Prorating consists of multiplying the sub-annual values by the corresponding annual proportional discrepancies. If the benchmark is not available, the closest proportional discrepancies are used. As a consequence, the proportional corrections are $\hat{\theta}_t/s_t$. The prorating method preserves the proportional movement within each year: $\hat{\theta}_t/s_t - \hat{\theta}_{t-1}/s_{t-1} = 0$. However, large discontinuities can emerge between the last quarter of a year and the first quarter of the following year, if the discrepancies are not uniform from year to year. For further details, see the module “Micro-Fusion – Prorating”.

Example

It is worth considering a simple example (Quenneville and Rancourt, 2005). Suppose there are three observations y_0, y_1, y_2 such that y_1 and y_2 must add up to y_0 . One way of reconciling the observed values of y_1 and y_2 with y_0 is prorating where the value of y_1 is set equal to $b_1 = \frac{y_1 y_0}{(y_1 + y_2)}$, the corrected value of y_2 is set equal to $b_2 = \frac{y_2 y_0}{(y_1 + y_2)}$, and so $b_1 + b_2 = y_0$.

Denton method

This is a quadratic programming method. The aim of this method is to make the quarterly data coherent with annual totals, while preserving all quarter-to-quarter changes as much as possible. According to the general solution (Denton, 1971), the adjusted values should be equal to the original values plus linear combinations of the discrepancies between the two sets of annual data. In the module “Macro-Integration – Denton’s Method”, a general overview about this method with examples is available.

There are related methods such as nonbinding benchmarking (these are not benchmarks in a strict sense, but simply low frequency measurements of the target variable) or interpolation, which are based on similar principles. These are also well-discussed in Dagum and Cholette (2006).

2.1.4 *Indirect vs. direct adjustment*

In practice, we usually examine the joint impact of more time series rather than a single given one. If a time series can be constructed as the sum of several time series it is called an aggregate series. An aggregate time series can be seasonally adjusted in two natural alternative ways:

- *Direct approach*: we produce the aggregated time series then we adjust it seasonally.
- *Indirect approach*: we apply the seasonal adjustment for components of time series (with the same method and software) then we sum the adjusted time series.

Apart from the methods above, there are further possibilities:

- *Spurious indirect approach*: This approach is applied only when it is unavoidable to calculate the aggregate series based on adjusted components which are generated in different ways (different approaches and software). *Example*: the described method realises when each European or Euro-zone state seasonally adjusts its series with its own method and strategy, and the European seasonally adjusted series is then derived as the aggregation of the adjusted national series (Astolfi, Ladiray, and Mazzi, 2001).
- *Mixed approach*: The methods described are not the only possible ones. A mixed approach is available. In this case, the method is based on subsets of the basic series, which are aggregated in one new component, this component and the remaining sub-series can then be adjusted and the adjusted aggregate derived by implication (Astolfi, Ladiray, and Mazzi, 2001).
- *Multivariate seasonal adjustment*. The multivariate seasonal adjustment consists of adjusting the series simultaneously, taking their covariance structure into account. Detailed mathematical background can be found in Birrell, Steel and Lin (2010).

Direct and indirect strategies could produce quite different results. While the aggregate is a linear combination of the components and the seasonal adjustment is a non-linear process, direct and indirect approaches *do not generally coincide*, except under special conditions (for instance, when the decomposition model is purely additive, or there are no outliers in the series).

Possible methods of the choice

In lack of general decision-making process concerning the between the methods, there are proposals. Examine the characteristics of the seasonal pattern in the component time series. If they show similar pattern then the direct approach is suggested. In spite of this, if the seasonal patterns of the different components show significant differences then one can suggest to use the indirect approach. However, the presence of residual seasonality is always to be checked in all of indirectly seasonally adjusted aggregates since the inadequately adjusted components can result in presence of residual seasonality.

Another way of model choice can be the following. It is possible to analyse the quality figures of the indirect and direct seasonally adjusted estimates (Astolfi, Ladiray and Mazzi, 2001). In this case one can examine, among other things, the smoothness of the components, revision rates, and analyses of the residuals.

The third way is to satisfy the user's requirements. On the one hand, users are interested in getting consistent and coherent outputs and therefore the indirect approach seems to be a good choice to avoid inconsistencies in data. On the other hand, the direct approach is favoured for transparency and accuracy.

Practice at statistical agencies

The choice between these methods is crucial, and has been the subject of articles and discussions for years. Theoretically, there are no guidelines to which of the methods is the best. The choice of method should depend on the system of series that is considered, according to Linde (2005) and Eurostat ESS Guideline (2009). According to ESS Guideline, the direct approach is preferred for transparency and accuracy, especially when component series show similar seasonal patterns. The indirect approach is preferred when components show seasonal patterns differing in a significant way.

Many national statistical institutes, such as Statistics Sweden or Statistics Netherlands (Bikker, Daalmans and Mushkudiani, 2010) prefer the direct approach to the indirect. The direct method was applied at Central Bureau of Statistics in Israel for composite series, but the indirect method has been adopted based on comprehensive studies linked to composite series and their components. However, aggregate series, such as national accounts, composite price index or manufacturing are still adjusted directly. As a consequence, the choice between direct and indirect approaches is a very complex issue and therefore, it is advisable to make a decision based on scrutiny.

2.1.5 Cross-sectional (aggregation) consistency, reconciliation

Problem description

While many economic data, for instance, national accounts are calculated based on an accounting system, the equality of the aggregate series and the sum of their components (along the whole length of time series) is desirable. The problems are the following:

- The additivity is not fulfilled as a result of the *non-linearity* (Xie and Elezovic, 2012) of the seasonal adjustment procedure. While time series resulting from aggregation of several sub-series can be seasonally adjusted directly or indirectly (Xie and Elezovic, 2012), the problem is the following in other words: there is discrepancy between the direct and indirect seasonally adjusted aggregates.
- The most important features of the dependence structure between the non-adjusted series should be preserved. However, there is *inconsistency* in the growth rates of related series after the seasonal adjustment (Xie and Elezovic, 2012).
- Lack of *coherence* in a system of time series because different accounting relationships are not preserved.

Methods to achieve aggregation consistency

- *Only indirect or only direct seasonal adjustment*: In this case, we apply one of the well-known seasonal adjustment methods discussed in the previous sub-section. The exclusive application of one of these methods is not preferable because it is difficult to maintain the quality of seasonal adjustment of the aggregate series.
- *Multivariate approaches*: This method operates either with structural time series models (further details are in Xie and Elezovic (2012) and Tsay (2005)) or coordinated seasonal adjustment. The approach based on structural time series models permits to derive simultaneously the seasonally adjusted series for the aggregate and the components. Coordinated seasonal adjustment entails an additive model and exactly the same filter applied to all series including in the contemporaneous constraints. This is unrealistic in practice, according to Xie and Elezovic (2012).
- *Reconciliation*: In case of reconciliation, all series are first seasonally adjusted (direct approach) and then the discrepancies are distributed according to some criteria. According to Statistics Canada, the contemporaneous constraint is satisfied while the distortion of reconciliation is minimised. Possible reconciliation methods:
 - *prorating*;
 - Denton method and methods derived from Denton's principle (Xie and Elezovic, 2012). Since Denton (1971), several extensions have been proposed (e.g., Di Fonzo and Marini, 2009). The extended approaches are also post-adjustment methods which can be computed based on the minimisation of special distance functions (distance means the closeness of the reconciled data to the original one).
 - *Regression model based on alterability coefficients* Quenneville and Rancourt (2005): The prorating method can be performed via regression model as well. Let $y_1 = b_1 + e_1$, $y_2 = b_2 + e_2$, $y_0 = b_1 + b_2$, $e_1 \sim (0, y_1)$, $e_2 \sim (0, y_2)$, where $e_i \sim (0, y_i)$ means that the error has mean 0 and variance y_i . It is possible to obtain a simplified model by eliminating b_2 : $y_1 = b_1 + e_1$, $y_0 - y_2 = b_1 + e_2$. Assuming e_1 and e_2 are uncorrelated, the best linear unbiased estimate of b_1 is a weighted average of y_1 and $y_0 - y_2$ where the weights are inversely proportional to the variances:

$$b_1 = \left(\frac{1}{y_1} + \frac{1}{y_2} \right)^{-1} \left(\frac{y_0}{y_1} + \frac{y_0 - y_2}{y_2} \right) = y_1 \frac{y_0}{y_1 + y_2}$$

The simple method is able to be applied to perform the results of the prorating method. However, the variance of the error associated with an observation can be artificially modified. In this case, let

$$e_1 \sim (0, a_1 y_1), e_2 \sim (0, a_2 y_2), y_0 = b_1 + b_2 + e_0, e_0 \sim (0, a_0 y_0),$$

where $a = (a_0, a_1, a_2)$ is a known vector of alterability coefficients. These coefficients must take non-negative values. The general practice consists of setting the coefficients of variation equal to 1 for all series and 0 for unalterable series. The alterability coefficients could also reflect the relative reliability of the various series. In the seasonal adjustment, these coefficients may depend on the importance of the indicators and the quality of the seasonal adjustment. Seasonally adjusted data of better quality (often they are the result of manual interventions) should not be modified and discrepancies should be distributed on less important series or series automatically decomposed. Consequently, the alterability coefficients increase or reduce the covariance matrices of some of the series in the system, thus these series are more or less affected by reconciliation (Dagum and Cholette, 2006). In this case, it can be proved that

$$b_1 = y_1 + \frac{a_1 y_1}{a_0 y_0 + a_1 y_1 + a_2 y_2} [y_0 - (y_1 + y_2)]$$

$$b_2 = y_2 + \frac{a_2 y_2}{a_0 y_0 + a_1 y_1 + a_2 y_2} [y_0 - (y_1 + y_2)]$$

Prorating is useful in case of one-way classification, but higher dimensional tables of time series require regression based model in order to simplify the treatment.

2.2 *Special Issues*

2.2.1 *Aggregation of seasonally adjusted chained indices*

Chain-linking

In quarterly (in case of, for example, national accounts) or monthly (industry or commercial) estimations the chain-linking method has been applied for constant price calculations. The introduction of chain-linking was necessary because the previous year weights reflect the economic structural changes better than the fix base year weight structure. In case of quarterly time series, first constant price data are calculated at average prices of the previous year from current price data, and then the whole time series is chain-linked back to the beginning of the series with the help of indices. The time series thus produced is built on reference year prices (for example year 2005 prices), and the base year determining the structure is the previous year for all data of the time series, i.e., the base year annually differs. As a result, data of the time series at average prices of the reference year are **not additive** within the given quarter, i.e., the sum of sub-aggregates are not necessarily equal to an aggregate, therefore chain-linking has to be carried out in case of every time series (separately for sub-aggregates and aggregates).

Linking techniques for annually chain-linked quarterly data

According to the literature, three linking techniques are known:

- **Annual overlap:** the average annual prices of the previous year are used as weights for each of the quarters in the current year, with the linking factors being derived from the annual data.
- **One-quarter overlap:** one quarter of the year (e.g., the fourth quarter) is compiled at both the average prices of the current year and the average prices of the previous year. The ratio between the estimates for the linking quarter provides the linking factor.
- **Over-the-year:** all quarters are compiled at the average prices of both the current year and the previous year. The year-on-year growth rates are calculated and then linked together. This technique is not supported by Eurostat.

Seasonal adjustment must be carried out after chain-linking. Further details about this method are available in Task Force report (2008) and on the webpage of Statistics Estonia.

2.2.2 Revision

The revision of the seasonally adjusted data can be derived from two main sources: the revision of the unadjusted data or the seasonally adjusted data.

The revision of the unadjusted data is important because of the deficiencies in the system of data collection. For example, the data providers send the information after the deadline, or they send erroneous data. These data must be revised and this revision influences the adjusted data.

The revision of seasonally adjusted data is important for sake of a better estimation. All new incoming data conveys new information, by that we get more accurate estimation for the seasonal pattern.

Strategies to revise seasonally adjusted data

There are two extreme types of approach to handle with revisions: current and concurrent adjustment.

- **Current (or forward factor):** According to this adjustment, seasonal and calendar factors are revised only once in a time span (generally one year), when the last month or quarter becomes available. This implies that models, outliers and filters are periodically revised. Forecasted factors are used to derive the calendar and/or the seasonally adjusted data before the review.
- **Concurrent:** According to this adjustment seasonal and calendar factors are revised whenever a new or revised data is received. This implies that models, outliers and filters are always revised.

In practice, a compromise should be found between the current and concurrent adjustment: the former may provide a misleading signals at the end of time series, the latter may cause a significant instability in seasonally adjusted data. Two alternative approaches are suggested in the ESS Guidelines (2009):

- **Partial concurrent adjustment:** this method contains forecasting the seasonal factors and identifying the model for the next period. If a new or revised observation becomes available, we re-estimate the parameters of the model but the model is the same. This process takes the new information derived from the received data into consideration and intends to avoid significant revisions.

- *Controlled current adjustment*: The current adjustment is considered, but its results are internally compared to those derived from the partial concurrent approach, which is preferred when discrepancies between the two approaches are considered important (see ESS Guidelines (2009)). Since each series needs to be seasonally adjusted twice this adjustment is practicable only for a limited number of series.

Horizon of the revision

The revision policy has to contain the **horizon of the revision**. The entire seasonally adjusted time series may change by re-estimating the seasonal factors. The publication of this change is not obligatory but it is worth doing because of the transparency and accuracy. Hence, we have to find the optimum length of the revision period which is short enough not to confuse the users, but it is not too short in order to assure the reliability of the seasonally adjusted data. The issue arises what the explanatory power of a new observation is. If a new data affects only the last some years of the observation then it can be useful to limit the revision period. Therefore, according to ESS Guidelines (2009), the best alternative is to revise the seasonally adjusted data from 3-4 years before the beginning of the revision period of unadjusted data. Another acceptable practice is the revision of the entire time series irrespective of the revision on the unadjusted data. The general revision strategy applied at *Statistics Denmark* is that all series should at least be revised 13 months/5 quarters back in time. At the most, they should be revised 4 years back in time (Linde, 2005).

2.3 *Treatment of the crisis*

In 2008, when the economic downturn burst out, the reliability and stability of the official seasonally adjusted and trend-cycle data became seriously undermined. Consequently, seasonal adjustment became not only more relevant, but much more difficult: the negative effect of the crisis significantly influenced the examined time series and it caused rapid changes in the earlier structure with the consequence of increasing the uncertainty of the data (in other words, it is required to handle more volatile data) (Ouwehand and Krieg, 2012). In similar circumstances, it is not obvious to decide whether the seasonal pattern changes based on the end of the time series. Therefore, the results of the seasonal adjustment are not only more uncertain than usual but they are accompanied by much larger revisions as well. An agreement about the treatment strategy of the crisis among statistical offices is still lacking because crisis may influence each time series in a different way. Although the European countries applied heterogeneous approaches to deal with the crisis in 2008-2009, it is possible to classify them according to some criterion. In particular, in accordance with the timing of the intervention on the seasonal adjustment specifications, it is possible to distinguish real time and ex-post treatments (Ciammola, Cicconi, and Marini, 2010).

- *Real time treatment* stands for methods which are applied in estimating the abrupt movements of the processes during the crisis. The most appropriate tool would be handling of outliers at the end of the series.
- *Ex-post treatment* requiring the inclusion of special intervention variables to model the effect of the crisis such as ramp effects.

The crisis became the basis of many studies and lectures about the available strategies and their effects.

The strategy applied at Statistics Netherlands incorporated several steps at Ouwehand and Krieg (2012). First of all, usual approaches were continued such as *concurrent with annual review* and *automatic outlier detection*. Moreover, issues as part of the pre-treatment process, such as setting outliers are also important ones. However, increased monitoring is also unavoidable. At the Hungarian Central Statistical Office, the standardised seasonal adjustment policy was supplemented during the global financial crisis. Beyond the conventional disquisitions, detailed analysis has been performed in each period paying special attention to the models and outliers. After all, the Office decided not to use level shifts in consecutive time periods in the concerned time series. In a study of Ciammola, Cicconi and Marini (2010) carried out with Italian series, the results of a real-time treatment method (based on some variants of the partial concurrent treatment) were compared with an ex-post treatment (based on the ramp effects). Ramp effect is a special intervention variable: it has a start and an end date allowing for a linear increase or decrease in the level of the series (see Figure 1).

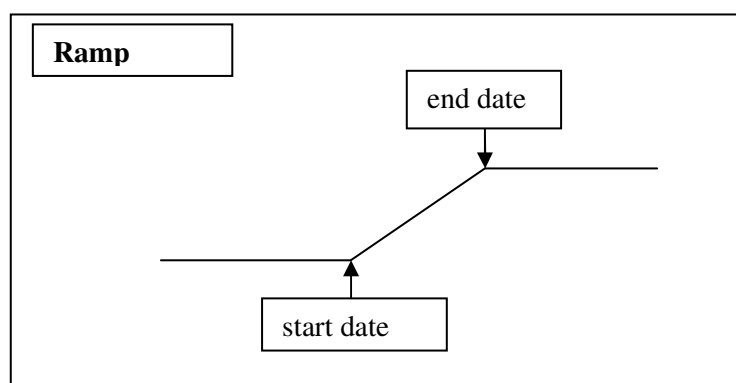


Figure 1: Ramp outlier

In Bell and Lytras (2013), ramp effects are considered to deal with the crisis, but a different procedure (based on the AICC information criterion) is used to set the beginning and the length of the ramps.

The common conclusion is that a carefully monitoring of seasonal adjustment is required in times of strong economic changes. Models should be closely reviewed by data producers as soon as new extraordinary data are detected in the raw series and special intervention variables could be used when regular (often automatically detected) outliers do not fit well the changes in time series.

Seasonal adjustment tools used by NSIs and other international organisations are capable to carry out outlier detection automatically.

In case of automatic outlier detection, level shifts and temporary changes may appear at the end of the time series as a result of the crisis. However, when several outliers are detected in a short time span, several disadvantages arise: firstly, it can be difficult to give an economic interpretation; secondly, these outliers can generate period-on-period growth rates with a very irregular pattern due to the fact that they are very close; finally, the application of level shift outliers implies permanent shocks that are not compatible with business cycle movements. The latter effect is mitigated through the use of temporary change, but the model forecasting ability could worsen. Another approach exclude any intervention as the trend is able to adapt to the long-time impact of the crisis.

Although, both level shifts and ramp outliers yield satisfactory results in terms of stability in parameter estimates, ramp outliers have many advantages, especially when the direct approach is used to seasonally adjust the aggregates. In fact it is difficult to explain to unprofessional users why the fallback derived from the use of level shift is concentrated in the periods where the outlier is identified. Furthermore, level shifts may fail in the “linearisation” procedure of time series preceding the decomposition. On the other hand, the main drawback in the use of the ramp effects is that they cannot be automatically detected and their features (the starting point and the ramp length) have to be set manually.

2.4 *Data presentation, communication with users, and documentation*

The estimations applied in seasonal adjustment are very complex issues. Hence, the responsible institutes such as Statistical Offices and National Banks have to undertake the task within seasonal adjustment, publish, and interpret the results in press releases.

According to ESS Guidelines (2009), data can typically be presented either in unadjusted, calendar adjusted, seasonally adjusted, or trend-cycle form. Apart from data, users can also be classified, according to OECD Handbook (2007):

- *general public*: they are usually not interested in technical details, thus they only need “basic” metadata;
- *informed users*: they need detailed information how the statistical program performing the seasonal adjustment was carried out, as well as statistics on the validity of the adjustment for specific series;
- *analytical users*: they need some of the results of the statistical program to reprocess them for their own use(s).

Discussions may focus on whether the unadjusted data should be published together with the seasonally adjusted data. The problem emerges due to the fact that two different time series linked to the same ‘phenomenon’ may confuse users. This uncertainty can be reduced by appropriate suggestions about which time series are recommended to be applied in different cases. However, if seasonally adjusted and unadjusted data are published apart from other data (such as series only adjusted for trading-day), then the risk of confusing the general public is significant.

Another question is whether it is advisable to publish the seasonally adjusted time series or the trend-cycle component. If the focus is on the underlying medium term movements, then trend-cycle estimates is the preferred form. According to the general recommendation, the focus of press release concerning the main sub-annual indicators should be on their appropriately seasonally adjusted version, but the original data should be sent to the users in any forms. In case of user’s request, the offices can publish the trend-cycle or other components of the seasonal adjustment process but it has to be clear that the seasonally adjusted data are the most important figure for the short-term variation.

Month-on-previous month and quarter-on-previous-quarter growth rates for original series are not very informative unless seasonal effects are negligible. Consequently, statistical agencies seldom use them in their releases of indicators affected by seasonal fluctuations. The users and the media often focus on the year-on-year changes (YoY) which are the rates of change with respect to the same period of previous year. This should be applied to the original data and also to the calendar adjusted data if the

latter are available. If necessary, special effects, e.g., the so-called base effect¹ contained in the base period should be highlighted when presenting YoY. Period-on-period growth rates and changes in level should be computed on seasonally adjusted time series, but in case of high volatility, it should be used with caution. When the seasonal component is not deterministic, the rate of change on original data and seasonally adjusted data can show conflicting signals, leading the general public and even some informed users to question the validity of the results. However, YoY change calculated on seasonally adjusted series is a common practice.

Moreover, the presentation of annualised level changes $-(1+\Delta_t)^{12}$ or $(1+\Delta_t)^4$, where Δ_t is the growth rate of one month or quarter (compared with the previous one) – is not recommended, because it can result misleading signals, especially for series displaying high volatility. Hence, where annualised changes are used, users should be provided with information regarding the possibility of misleading signals due to series volatility. Also, the annualised period-to-period growth rates are not recommended for the presentation of monthly or quarterly growth rates.

Beside the results, other important information can help the users to understand what the seasonal adjustment is all about. While the seasonal adjustment procedure is complex, it is recommended to explain it without presuming detailed mathematical and statistical background. This explanation should contain its benefits, its aims, and the steps of the process. “For the benefit of users requiring information about appropriateness of the seasonal adjustment method applied, statistical agencies should provide a minimum amount of information that would enable an assessment of the reliability of each seasonally adjusted time series.”

The situation is different concerning with the analytical users. They especially need the availability of metadata. Hence, one should publish more detailed information about the applied methods, the most useful figures, the main specifics of the adjustment, outliers, expected problems of the adjustment and recommendations of data publications. For example, if the time series contain neither seasonal effect nor calendar effect then the data provider should publish the original time series as seasonal adjusted time series.

3. Design issues

4. Available software tools

5. Decision tree of methods

¹ A base effect occurs when the evolution of a variable’s annual rate from month t to month t+1 varies because of the evolution of the variable’s level 12 months before and not because of the variation of the variable’s level between month t and t+1 (Banque centrale du Luxembourg, 2004).

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Astolfi, R., Ladiray, D., and Mazzi, G. L. (2001), Seasonal Adjustment of European Aggregates: Direct versus Indirect Approach. Office for Official Publications of the European Communities, Luxembourg.
http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/38.pdf
- Bikker, R., Daalman, J., and Mushkudiani, N. (2010), A multivariate Denton method for benchmarking large data sets. Report, Statistics Netherlands.
<http://www.cbs.nl/nr/rdonlyres/7b2387f2-5773-42cf-8c50-5f02b451a2e4/0/201002x10pub.pdf>
- Birrell, C., Steel, D. G., and Lin, Y. X. (2010), Seasonal Adjustment of an Aggregate Series using Univariate and Multivariate Basic Structural Models. Centre for Statistical & Survey Methodology Working Paper Series.
- Central Bureau of Statistics Israel;
<http://www.cbs.gov.il/publications/tseries/seasonal07/introduction.pdf>
- Ciammola, A., Cicconi, C., and Marini, M. (2010), Seasonal adjustment and the statistical treatment of the economic crisis: an application to some Italian time series. 6th Colloquium on Modern Tools for Business Cycle Analysis, 26-29 September 2010, Eurostat, Luxembourg.
- Dagum, E. B. and Cholette, P. A. (2006), *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*. Springer.
- Denton, F. T. (1971), Adjustment of Monthly or Quarterly Series to Annual Totals; An Approach Based on Quadratic Minimization. *Journal of the American Statistical Association* **66**, 99–102.
- Di Fonzo, T. and Marini, M. (2009), Simultaneous and Two-step Reconciliation of Systems of Time Series. Working Paper Series, N.9.
- ESS Guidelines on seasonal adjustment (2009);
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-09-006/EN/KS-RA-09-006-EN.PDF
- FAQ about chain-linking method, Statistics Estonia: www.stat.ee/dokumendid/29861.
<http://www.catherinechhood.net/safaqdiagnostics.html>
- Linde, P. (2005), *Seasonal Adjustment*, Statistics Denmark.
- OECD (2007), *Data and Metadata Reporting and Presentation Handbook*.
<http://www.oecd.org/std/37671574.pdf>
- Ouwehand, P. and Krieg, S. (2012), Seasonal adjustment at Statistics Netherlands in times of strong economic changes.
- Öhlén, S. (2006), *Benchmarking and seasonal adjustment – A Study of Swedish GDP*. http://epp.eurostat.ec.europa.eu/portal/page/portal/euroindicators_conferences/documents_seasons/OHLEN%20AB.pdf

Publications of Statistics Canada (2009), *Benchmarking and related techniques*.

<http://www.statcan.gc.ca/pub/12-539-x/2009001/benchmarking-etalonage-eng.htm>

Quenneville, B. and Rancourt, E. (2005), Simple methods to restore the additivity of a system of time series. Statistics Canada, Time Series Research and Analysis Centre.

Stuckey, A., Zhang, X. M., and McLaren, C. H. (2004), *Aggregation of Seasonally Adjusted Estimates by a Post-Adjustment*. Methodological Advisory Committee, November 2004, Australian Bureau of Statistics. http://www.uow.edu.au/~craigmc/abs_agg_2004.pdf

Task Force on Seasonal Adjustment of Quarterly National Accounts Final report (2008), Committee on Monetary, Financial and Balance of Payments Statistics.

<http://www.cmfb.org/pdf/TF-SA%20QNA%20-%20Final%20Report.pdf>

Tsay, R. S. (2005), *Analysis of Financial Time Series*, 2nd edition. John Wiley & Sons.

Xie, Y. and Elezovic, S. (2012), Reconciliation of seasonally adjusted data with application to the Swedish quarterly national accounts. European Conference on Quality in Official Statistics.

Interconnections with other modules

8. Related themes described in other modules

- 1.

9. Methods explicitly referred to in this module

1. Micro-Fusion – Prorating
2. Macro-Integration – Denton’s Method
3. Seasonal Adjustment – Seasonal Adjustment of Economic Time Series

10. Mathematical techniques explicitly referred to in this module

1. Interpolation
2. Extrapolation
3. Regression

11. GSBPM phases explicitly referred to in this module

1. GSBPM Phase 6.1, 6.2, 6.3

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Data reconciliation
2. Benchmarking

These processes are also incorporated in the topics “Macro-Integration” and “Micro-Fusion”.

Administrative section

14. Module code

Seasonal Adjustment-T-Issues on SA

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	21-11-2012	first draft	Attila Lukács	Hungarian Central Statistical Office
0.2	31-08-2013	second draft: changes based on reviews	Laszlo Sajtos	Hungarian Central Statistical Office
0.3	14-10-2013	third draft: changes based on reviews	Laszlo Sajtos	Hungarian Central Statistical Office
0.3.1	25-11-2013	version submitted to editorial board	Laszlo Sajtos	Hungarian Central Statistical Office
0.3.2	11-12-2013	preliminary release		
0.4	13-12-2013	minor improvements	Laszlo Sajtos	Hungarian Central Statistical Office
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:29



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Disclosure Control – Main Module

Contents

General section.....	3
1. Summary	3
2. General description.....	4
2.1 Tables versus microdata	4
2.2 Tabular data.....	5
2.3 Trade-off: Probability of disclosure versus information loss	5
2.4 User needs and SDC.....	5
2.5 Data access	6
3. Design issues	6
4. Available software tools.....	7
5. Decision tree of methods	7
6. Glossary.....	7
7. References	7
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

Statistical disclosure control (SDC), or Statistical Disclosure Limitation (SDL) as it is also called, is an activity aimed at the protection of data that are to be released by an NSI. Protection means that individual entities (such as businesses) are not (readily) identified, and more particularly, confidential or sensitive information about such entities is not released to third parties. This to prevent misuse of data intended for statistical purposes. Instead of focusing on aggregates, the attention is directed at individual entities and their response, or the information that is available from them. This shift of attention may even be inadvertent, because certain aggregates happen to consist of one, or a few, entities of which one dominates the contribution.

The aim of SDC is twofold: to identify the risks involved in releasing data, and secondly to modify 'risky data' in such a way that for the resulting data the disclosure risk is negligible. The challenge in modifying the data is to do it in such a way that no (possibly sensitive) information about individual entities is disclosed, directly or indirectly, whereas the protected data are still of interest for statistical research and policy studies. The aim of SDC is not to hamper statistics, but to hamper non-statistical use of the data, such as 'unearthing' information on certain individuals. Statistics is not about individuals but about groups of individuals. So there is room to protect the privacy of individuals whilst serving the interests of society to provide it with statistical information, for research, policy making or general interest. Note that, in the context of this handbook, individuals usually mean individual businesses.

In case of business statistics, tables are the usual pieces of information that are released to users outside statistical institutes. Business populations are usually too skewed so that safe release of business data in microdata form is usually not possible for public use: large units cannot be protected, without rendering the microdata useless. In some countries it may be possible to allow researchers from bona fide institutes to have access to microdata, under strict conditions, and/or in safe settings. But the final results of this research are also in the form of aggregates, such as tables. So in practice, disclosure control of tables is more of an issue for business data than is the protection of microdata. For that reason the focus of attention in the present module is on the SDC of tables.

For tabular data the first task in protecting them is to define rules that separate safe from unsafe data. Once these rules have been specified they can be applied to the tables at hand. In case cells (in tables) have been found that are considered unsafe according to the rules applied, the next thing to do is to try to eliminate them by modifying the tables. For this a range of techniques is available. The problem is to apply them to the tables, in such a way that the resulting tables are safe (according to the rules that have to be considered) and the modification of the tables is minimal. For microdata a similar problem exists, but that will not be highlighted in the present module, for the aforementioned reasons.

For more detailed information about Statistical Disclosure Control issues, we refer to Hundepool et al. (2012), Hundepool and De Wolf (2011), Willenborg and De Waal (2001) and Willenborg and De Waal (1996).

2. General description

Microdata are data about individual entities, such as persons, households, companies, municipalities, etc. At NSIs business data are usually stored in microdata files. These files are used at NSIs as sources to produce aggregate data that can be released to external users. Public release of microdata for business data is typically not an option.

Tabular data are aggregate data, about groups of individual entities. It is convenient to divide the tabular data into two kinds: quantitative tables and frequency tables. Quantitative tables contain data for continuous variables, such as income, turnover, weight shipped, etc. Frequency tables contain numbers of units that have the properties of the respective cells in such a table, such as the number of business involved in a certain business activity in a specific part of a country (province, district, municipality, etc.). Frequency tables are not as often used in business statistics as magnitude tables. Like microdata, frequency tables are more favoured in social statistics than in business statistics. For this reason we focus on the protection of quantitative tables in the present handbook.

So, when publishing business data in the form of quantitative tables, the question is what to be aware of. How to prevent that information on certain businesses is revealed, maybe not exactly but with sufficient precision, by deduction and using certain prior knowledge. Moreover, how to appropriately modify such tables, such that the resulting tables will still be useful but will satisfy the safety rules as well.

2.1 *Tables versus microdata*

Microdata contain information on individual entities such as business or enterprises in the business statistics area. The individual entities in this area are more complex than those in the social statistics area (persons, households). They are usually also different in terms of size, measured in a variety of ways (number of employees, turnover, profit, etc.). Because of the skewness of certain identifying characteristics of such entities in the business world, it is impossible to release microdata to external users, as the extremest individuals can be recognised immediately. Protecting business microdata using SDC techniques usually does not work, or would produce data that are safe but useless for statistical analyses. So whereas in social research protected microdata sometimes can be released, in business statistics this is not an option.

The publication of business data is therefore typically as aggregate data. This means that data not about individual businesses or enterprises are published, but about groups of such entities. For instance, one might want to publish about businesses providing financial services in the various regions (provinces, districts) of a country. This kind of information is usually published in the form of tabular data, or tables.

However one should not be fooled by the fact that these data are about aggregates, and therefore would need no protection. There is still the possibility that information about individual companies can be inferred from aggregate data, maybe not exactly, but with sufficiently high precision. This happens, for instance, if there is an entity that stands out in a group of entities, in the sense that it dominates this group's total (say total turnover). But in certain cases it is possible to publish this kind of tabular information, but after having modified the original table somewhat. How such tables can be modified in order to make them suitable for publication is the subject matter of statistical disclosure control of tabular data.

2.2 *Tabular data*

Whereas microdata contain information on individual entities, tables contain information of groups of entities. In other words they are aggregate data. Naively one would perhaps expect that aggregate data do not present any disclosure risks as they are about groups of individuals. But this is generally not true, if only because the size of a group (represented in a table by a cell) may correspond to one individual in the population. Or it may be the case that a group of entities is very heterogeneous with respect to a particular variable, such as turnover. It may very well be that a single individual dominates the individuals corresponding to a particular cell in a table. Publishing the contents of this cell (as part of a bigger table) would disclose the turnover of a particular company, say, in a particular year, with an error that is related to the contributions of the fellow companies represented in this cell. If, for instance, a company attributes 99% of a cell value (where the remaining 1% is attributed by, say, 5 other companies) that cell value is a very good estimate of the contribution of that largest company in that cell. Tabular data are very important for releasing business data. In particular this is true for tables of magnitude data. In the handbook the module “Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables” is devoted to the disclosure limitation of tabular data.

2.3 *Trade-off: Probability of disclosure versus information loss*

When selecting the method or methods to be used, two competing aspects must be taken into account:

- Probability of disclosure (also called ‘disclosure risk’). This is the probability, assuming some kind of disclosure scenario, that there will be an identification of an individual entity.
- Information loss. This is used to express the loss of data utility when applying SDC techniques to a data set or a set of tables. Often this concept is used informally, but in some cases it is formalised in the form of a target function. The SDC problem is then formulated as an optimisation problem. For example the number of suppressed cells in a table may be seen as an example of an information loss measure.

In general, there is a trade-off between disclosure risk and information loss: reducing the disclosure risk will lead to increased information loss, and vice versa. The choice of an acceptable risk level has to be made after careful deliberation by an NSI. This usually depends on the disclosure scenario that is assumed. Because it is easier in practice, the choices will crystallise into a set of rules that can be applied easily in practice, by various groups in a statistical office. Without such a set of rules, protecting data prior to release would be tailor-made, difficult to check, and arbitrary (each department uses its own rules). It is preferable to have a common set of rules, to be used across an NSI.

2.4 *User needs and SDC*

In practice there may be a conflict between user demands and SDC. The users want certain variables with certain detail in the data, but this is not possible due to the SDC applied by an NSI. There also may be different user groups, with different demands. Policy makers, academic researchers, journalists and the general public may all have their specific requests. The task of the NSI is to manage these requests as well as possible, keeping a firm eye on the protection of the data.

2.5 *Data access*

Access to business microdata may (in some countries), for instance, only be granted at the premises of the NSI (on site facility), under strict conditions (contractual arrangement, safe settings, controlled access, output checking, etc.). A more recent trend (especially with social statistics microdata) is to allow researchers to have remote access to such data. This access mode has advantages for both researchers and the NSI: the researchers can work with the data at their institute, whereas the NSI can keep the microdata within its own walls, and it can control and log the access to the data. See also the theme module “Dissemination – Dissemination of Business Statistics” in the current handbook.

Access to business microdata is only to allow researchers to use the detailed information they need for their analyses. In the end, however, only aggregate information can be published. The microdata are used as an intermediate data source. Or, in an alternative interpretation, the external users are given similar access rights to these data as the employees at the statistical office working in the area involved.

For restricted access the microdata are lightly protected. Direct identifiers are removed, as well as information that is irrelevant for the research purpose. Some regional variables may be recoded into broader categories if this is possible given the research goals. But this is not a modification comparable to the production of safe microdata for external release (in the area of social statistics). Restricted access is only available to a select group of researchers. A major part of confidentiality issues is dealt with using legal protection, i.e., is agreed upon in contracts. Moreover, each research proposal is evaluated beforehand and only that information is made available that is necessary to conduct this specific research.

3. **Design issues**

The following aspects need to be designed and organised for tables:

1. The formulation of criteria as to what are safe and unsafe tables. For tabular data there are certain rules that are applied to identify cells in a table that are considered unsafe (due to the dominance of a small group of contributors for such cells). For more information on this see the module “Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables”.
2. The measures to be taken to modify unsafe tables into safe data. For instance tabular data can be protected by a combination of table restructuring and cell suppression. Or by rounding, or adding noise. The choice of a method may depend on the user group for which the data are prepared. For instance, to the general public tables with suppressed cells are acceptable, whereas academics would perhaps prefer tables where noise is added for protection. The goal of the measures taken is to produce data that are safe, and with minimum information loss compared to the original data. This aspect, however, we do not consider as a design issue, but as an algorithmic problem. It may involve solving a formal optimisation problem (and sometimes a big one). See the module “Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables”.
3. Mode of access to the data, depending on the intended user group (researchers, policy makers, journalists, the general public, etc.). For each group it should be decided what data should be released to them, or what kind of access they should have to the data. There are several

possibilities. Data can be released on a website or in a publication. Certain researchers may be granted access to the microdata, under strict conditions, and via safe settings or via remote access or remote execution. See also the theme module “Dissemination – Dissemination of Business Statistics”.

4. Available software tools

τ -ARGUS is a package intended to protect tabular data by various techniques, such as table redesign, various versions of cell suppression, rounding and controlled tabular adjustment. For more information see Hundepool et al. (2011). This package requires a commercial LP-solver (either Xpress or Cplex) for certain techniques (like cell suppression and rounding). The τ -ARGUS package itself, however, is free of charge. See also <http://neon.vb.cbs.nl/casc/index.htm>

There are other packages for the protection of tabular data, such as sdcTable (R package, no user interface available) and G-Confid (see, e.g., Statistics Canada, 2011). For a general discussion of different software tools, see Giessing (2013).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Giessing, S. (2013), Software tools for assessing disclosure risk and producing lower risk tabular data. Data Without Boundaries Deliverable 11.1 – Part B, February 2013.
(http://www.dwbproject.org/export/sites/default/about/public_deliverables/dwb_d11-1b_software-tools-disclosure-risk-assessment.pdf).
- Hundepool, A. and De Wolf, P. P. (2011), *Statistical disclosure control*. Methods Series, Statistics Netherlands, The Hague. See: <http://www.cbs.nl/en-GB/menu/methoden/gevalideerde-methoden/publicatie-analyse/statistical-disclosure-control.htm>.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and De Wolf, P. P. (2012), *Statistical disclosure control*. Wiley-Blackwell.
- Hundepool, A., Van de Wetering, A., Ramaswamy, R., De Wolf, P. P., Giessing, S., Fischetti, M., Salazar, J. J., Castro, J., and Lowthian, P. (2011), *τ -ARGUS user manual 3.5*. Statistics Netherlands, Voorburg.
- Statistics Canada (2011), *G-Confid User Manual*. Internal report.
- Willenborg, L. and De Waal, T. (1996), *Statistical disclosure control in practice*. Lecture Notes in Statistics, vol. 111, Springer.
- Willenborg, L. and De Waal, T. (2001), *Elements of statistical disclosure control*. Lecture Notes in Statistics, vol. 155, Springer Verlag.

Interconnections with other modules

8. Related themes described in other modules

1. User Needs – Specification of User Needs for Business Statistics
2. Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables
3. Dissemination – Dissemination of Business Statistics

9. Methods explicitly referred to in this module

1. Cell suppression
2. Table redesign
3. Rounding

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 6.4 Apply disclosure control

12. Tools explicitly referred to in this module

1. τ -ARGUS
2. sdcTable
3. G-Confid

13. Process steps explicitly referred to in this module

1. Statistical disclosure control

Administrative section

14. Module code

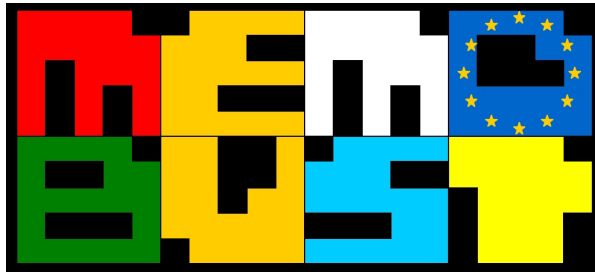
Statistical Disclosure Control-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	04-09-2013	first version	Leon Willenborg, Peter-Paul de Wolf	CBS (The Netherlands)
0.2	24-01-2014	revised version after review	Leon Willenborg, Peter-Paul de Wolf	CBS (The Netherlands)
0.3	04-02-2014	minor revision after EB review	Peter-Paul de Wolf	CBS (The Netherlands)
0.4	05-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:30



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Disclosure Control Methods for Quantitative Tables

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Tables of magnitude data.....	3
2.2 Tables of frequency count data.....	4
2.3 Sensitive cells	4
2.4 Table protection measures.....	5
3. Design issues	13
3.1 Sensitivity rules	13
3.2 Choice of table protection methods.....	13
3.3 Longitudinal aspects.....	13
4. Available software tools.....	13
5. Decision tree of methods.....	14
6. Glossary.....	14
7. References	14
Interconnections with other modules.....	15
Administrative section.....	16

General section

1. Summary

This module is about the protection of quantitative tables. Such tables are typically used to release data on business statistics. There are other forms that are sometimes used (such as microdata, frequency tables), but they are not dealt with here. More in particular we shall focus on a single quantitative table together with its marginals. The general case of linked tables (of which the one with hierarchical tables is a special case) is not treated here. A discussion of this case can be found in the literature. References will be provided.

The main issues with protecting quantitative tables are the identification of the unsafe cells in such tables, and how to protect them. Both issues will be addressed here. The one about actually protecting tables is ultimately rather technical, amounting to the solution of often complicated optimisation problems. How this is done is described in the literature, and references will be provided. We concentrate in this module on two techniques: table restructuring and cell suppression.

2. General description

2.1 Tables of magnitude data

Quantitative tables are tables in which the cell values are composed by summation of a continuous variable over all the contributors to a cell. This is in contrast to frequency tables in which only the *number* of contributors per cell is given. Other rules apply to frequency tables, and other protection methods may be more suitable than those for quantitative tables. In Section 2.2 there is more on frequency tables.

If exactly one or two contributors produce a cell total, it is clear that this cell cannot be published. In the case of a single contributor, individual information is released directly, and in the case of two contributors, one contributor can exactly calculate the other contribution by subtracting his or her own contribution from the cell total.

However, undesirable situations can arise also if there are more than two contributors in a cell. In principle, in the statistical disclosure control of quantitative tables, we must prevent (or at least make it more difficult) that any contribution can be estimated too accurately. This may occur, for example, also in the case that a very large contributor is present in a single cell along with several relatively small contributors. In this case, the second-largest contributor can calculate that the largest contribution does not contribute more than the cell total minus the second-largest contribution to the cell. A relatively good estimation of the contribution of the largest contributor can be obtained as a result, in conflict with the disclosure control rules of any NSI.

The presence of empty cells also requires extra attention. In some cases, an empty cell will be a so-called *structural zero cell*. This means that it is generally known that, logically, it is *impossible* for this cell to have a contribution. Such cells can therefore also not be used in the disclosure control: whatever you do, everyone knows that they must be empty cells.

At the same time, reliable information can sometimes be disclosed using *non-structural zero cells*. If there are contributors in such a cell, there is actually a sort of group disclosure: it is immediately clear that all the contributors to that cell have provided a contribution of zero (assuming that the

contributions are non-negative). If there are no contributors in the cell, but it is not logically impossible for a contributor to be in this cell, this in itself also reveals direct information.

2.2 *Tables of frequency count data*

Frequency tables are tables in which the number of contributors per cell is given. This is in contrast to quantitative tables in which the cell values are created by summation of a continuous variable over all the contributors to a cell. Other rules apply to quantitative tables, and other protection methods may be more suitable than those for frequency tables.

Frequency tables require the protection of recognisable data about statistical units. A violation of statistical confidentiality (a disclosure) may be two-fold: *identity disclosure*, i.e., disclosing the presence of an individual respondent in the table, and *attribute disclosure*, i.e., disclosing additional information about a single respondent. Some statistical laws do not allow identity disclosure on its own, while other statistical laws only care about attribute disclosure (for which identity disclosure is a necessary precondition).

For frequency tables, attribute disclosure can be formulated as follows. The user must first recognise a contributor or group of contributors in the table. This is followed by a statement about these contributor(s) due to the frequency distribution over the cells. The statement that the table makes possible about this group must provide more information about the members of the group than just the group size. In this sense, knowledge that is needed to recognise the members of the group can be considered not to be disclosive information about the members of the group. However, some statistical laws do not allow for disclosing this kind of information nonetheless.

The requirement is satisfied if the table does not provide any information about an individual statistical unit as such. However, the table should not provide information about groups of statistical units that can be identified (*group disclosure*). In particular, that is the case if the table contains variables that could provide harmful or potentially damaging information about these groups, like whether or not an environmental crime has been committed. Such data will be referred to as sensitive data.

2.3 *Sensitive cells*

The usual approach in SDC for tabular data is to identify the sensitive, or risky cells in a table. These are the ones that need to be protected. Various sensitivity measures are available that can be used for this task. All these measures need to be parameterised.

In Table 1 an overview of some well-known sensitivity rules is given. For a more detailed description, see Hundepool et al. (2012).

The first three sensitivity rules are so-called ‘concentration rules’. For concentration rules it should be borne in mind that in order to apply them, one needs to have information about individual contributions to the various cell values. In particular, one needs to know the n largest contributions to each cell.

In case of magnitude tables, often a combination of a concentration rule and a threshold rule is used to determine the sensitive cells. However, the concentration rules imply a certain threshold by definition.

Table 1. Various sensitivity rules

Sensitivity rule	Type of table	Cell is unsafe if
(n,k) rule / dominance rule	Magnitude	The n largest contributions to that cell make up for more than k% of the cell total.
(p,q) rule / ambiguity rule / prior posterior rule	Magnitude	Some contributor to that cell is able to derive an estimate of some other contributor to the same cell within p% of the true value, a-priori knowing all the other contributions within q% of their true values.
p % rule	Magnitude	Some contributor to that cell is able to derive an estimate of some other contributor to the same cell within p% of its true value.
Threshold rule	Frequency and Magnitude	The number of contributors is less than a prespecified threshold.

From a methodological point of view, the p% rule is preferred. Moreover, note that a concentration rule implies a certain threshold rule. E.g., under the p% rule a cell will always be unsafe when there are less than 3 contributors to that cell.

2.4 Table protection measures

To protect tabular data several methods are being employed in practice. In Table 2 some of the more important techniques for protecting tables (both magnitude and frequency tables) have been assembled. For detailed descriptions of these methods, we refer to Hundepool et al. (2012).

Table 2. Various SDC methods for tabular data

SDC Method	Type of table	Type of method	Short description
Barnardisation	Frequency	Perturbative	Randomly add/subtract 1 from some cell values.
Table redesign / table restructuring	Magnitude or frequency	Nonperturbative	Collapsing rows and/or columns.
Cell suppression	Magnitude or frequency	Nonperturbative	Completely suppress the value of some cells (put a “cross”).
Rounding <ul style="list-style-type: none"> • Controlled • Conventional / deterministic • Random 	Magnitude or frequency	Perturbative	Round each cell value to a prespecified rounding base.
Controlled Tabular Adjustment (CTA)	Magnitude	Perturbative	Selectively adjust cell values: unsafe cells are replaced by either of their closest safe values. Other cell values are adjusted to restore additivity.
Perturbation / adding noise	Magnitude	Perturbative	Add random noise to cell values.

As “Type of method” a two-fold classification is used: Perturbative or Nonperturbative. Whenever a method is of type Perturbative, this means that certain cell values will be replaced by adjusted cell values, i.e., they will be *perturbed*.

The most commonly used methods are table redesign and cell suppression. CTA is a promising recent technique, but is not used that often in Europe yet.

2.4.1 Table restructuring

In general, cells with a limited number of contributors or a cell with one or two large contributors are the obvious candidates to be characterised as risky. All risky cells must be protected. Before performing suppression on a large scale, restructuring the table can also be considered. By combining rows and/or columns, cells are pooled and the content per cell is increased. The result of this is that fewer cells are identified as risky by a sensitivity rule (such as the p % rule).

This method will generally lead to fewer risky cells in the table. Combining an unsafe cell with one or more safe cells may result in a cell that is safe.

There are no methodological conditions for using this method. However, externally imposed obligations sometimes specify what level of detail a table must have when published. This may be a Eurostat obligation. Or an NSI may have a publication policy requesting a certain level of detail for a table when published. So although restructuring could be applied successfully, publication policy might prevent this.

Furthermore, an assessment must be made between the information loss resulting from the larger number of suppressed cells that are needed to protect the table, and the information loss resulting from combining columns/rows, for which fewer crosses are needed.

The software package τ -ARGUS has provisions for recoding rows and/or columns in tables. Two situations are distinguished:

- In the case of a hierarchical spanning variable, the recoding implies that certain splits are omitted at the lowest level.
- In the case of an unstructured spanning variable, users are free to combine the columns or rows of a table as they choose.

Example. Figure 1 presents a fictitious table of turnover according to Region (hierarchical) and SizeClass. The crosses in Figure 1 are cells that are unsafe (or risky) according to some sensitivity rule. Figure 2 and Figure 3 provide two restructuring possibilities for this table.

In Figure 2 the variable SizeClass is recoded such that the categories 2 to 6 are combined into the category MediumSmall, and that the categories 7, 8 and 9 are combined into the category Large. Note that, in this way, all the risky cells are combined to create safe cells. In Figure 3 the recoding of the variable Region is such that the smallest detail level has been removed. This restructuring does not resolve all the problems: the risky cells at region level (for North and East) are still present in the table. This is not necessarily a problem. If the protector is satisfied with the structure of this table, he may decide to eliminate the remaining sensitive cells by, e.g., cell suppression. ■

	tot	2	4	5	6	7	8	9	99
tot	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
- North	4,373,664.00	×	×	719,049.00	659,680.00	688,962.00	756,529.00	1,549,049.00	385.00
.. 1	1,986,129.00	×	×	398,062.00	348,039.00	354,711.00	418,778.00	466,529.00	-
.. 2	1,809,246.00	0.00	-	223,990.00	221,332.00	241,913.00	258,233.00	863,393.00	385.00
.. 3	578,289.00	-	-	96,997.00	90,309.00	92,338.00	79,518.00	219,127.00	-
- East	3,703,896.00	15.00	×	642,238.00	515,003.00	534,147.00	620,392.00	1,392,096.00	-
.. 4	124,336.00	×	-	36,311.00	32,132.00	25,770.00	18,150.00	-	×
.. 5	526,279.00	-	-	93,589.00	94,957.00	110,930.00	81,799.00	145,004.00	-
.. 6	2,234,995.00	×	×	345,803.00	251,358.00	251,188.00	303,377.00	1,083,254.00	-
.. 7	818,286.00	-	-	166,535.00	136,556.00	146,259.00	217,066.00	151,870.00	-
- West	4,576,115.84	-	-	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	-
.. 8	485,326.00	-	-	63,767.00	75,442.00	87,305.00	59,953.00	198,859.00	-
.. 9	3,664,559.84	-	-	537,911.00	430,851.00	515,019.58	643,762.26	1,537,016.00	-
..10	426,230.00	-	-	47,294.00	37,277.00	61,572.00	71,417.00	208,670.00	-
- South	4,193,971.00	-	15.00	701,549.00	602,281.00	618,037.00	647,021.00	1,625,068.00	-
..11	2,752,743.00	-	15.00	488,613.00	392,395.00	363,490.00	402,925.00	1,105,305.00	-
..12	1,441,228.00	-	-	212,936.00	209,886.00	254,547.00	244,096.00	519,763.00	-
.99	-	-	-	-	-	-	-	-	-

Figure 1. Quantitative table for turnover according to region and size class

	tot	Large	SmallMedium	99
tot	16,847,646.84	11,814,874.84	5,032,387.00	385.00
- North	4,373,664.00	2,994,540.00	1,378,739.00	385.00
.. 1	1,986,129.00	1,240,018.00	746,111.00	-
.. 2	1,809,246.00	1,363,539.00	445,322.00	385.00
.. 3	578,289.00	390,983.00	187,306.00	-
- East	3,703,896.00	2,546,635.00	1,157,261.00	-
.. 4	124,336.00	55,888.00	68,448.00	-
.. 5	526,279.00	337,733.00	188,546.00	-
.. 6	2,234,995.00	1,637,819.00	597,176.00	-
.. 7	818,286.00	515,195.00	303,091.00	-
- West	4,576,115.84	3,383,573.84	1,192,542.00	-
.. 8	485,326.00	346,117.00	139,209.00	-
.. 9	3,664,559.84	2,695,797.84	968,762.00	-
..10	426,230.00	341,659.00	84,571.00	-
- South	4,193,971.00	2,890,126.00	1,303,845.00	-
..11	2,752,743.00	1,871,720.00	881,023.00	-
..12	1,441,228.00	1,018,406.00	422,822.00	-
.99	-	-	-	-

Figure 2. Recoding of SizeClass (all risky cells have disappeared)

	tot	2	4	5	6	7	8	9	99
tot	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
.North	4,373,664.00	×	×	719,049.00	659,680.00	688,962.00	756,529.00	1,549,049.00	385.00
.East	3,703,896.00	15.00	×	642,238.00	515,003.00	534,147.00	620,392.00	1,392,096.00	-
.West	4,576,115.84	-	-	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	-
.South	4,193,971.00	-	15.00	701,549.00	602,281.00	618,037.00	647,021.00	1,625,068.00	-
.99	-	-	-	-	-	-	-	-	-

Figure 3. Recoding of Region (not all risky cells have disappeared)

2.4.2 Cell suppression

2.4.2.1 Short description

A frequently used method to protect risky cells is to suppress (not publish) certain cells. The cell value is then simply replaced by a certain symbol, e.g., a cross (×).

In a quantitative table when the marginals are also provided, however, it is often not sufficient to suppress only the risky cells (i.e., only use so-called primary suppressions). If a suppressed cell is the only suppressed cell in a row, the suppressed value can, after all, simply be calculated by subtracting the other cell values in that row from the corresponding marginal.

To sufficiently protect risky cells, it is therefore also necessary to suppress other cells which, in themselves, are safe. This is called *secondary suppression*. It is not easy to perform this in such a way such that the risky cells are protected sufficiently, while also ensuring that not too much information is

removed from the table. Furthermore, account must also be taken of the fact that structural zero cells cannot be used as secondary suppressions: everyone knows that, by definition, these cells are empty.

To prevent a situation where suppressed, risky cells can be (re)calculated exactly, secondary suppressions are therefore necessary. However, also a “too accurate” estimation for a suppressed cell is not desirable. Indeed, what is the difference between the following statements: “This suppressed cell actually has a value of 10000” and “This suppressed cell actually has a value of between 9998 and 10002”. Given a suppression pattern, it is always¹ possible to calculate an interval in which a suppressed cell must lie. The method of “Cell Suppression” must then also produce a suppression pattern, for which the intervals that can be calculated are sufficiently large. The size of these intervals is determined by the rule that is used to determine the risky cells.

Fischetti and Salazar (2000) have developed a method to solve the above problem in an optimal manner. Their method is, in theory, applicable to arbitrary, additive tables with non-negative contributions. In practice, however, their solution involves too much computing time if the tables become too large, either in size or complexity. This is why a number of sub-optimal methods have been developed to find suitable suppression patterns for larger and/or more complex tables.

For example, the “modular approach” (also known as HiTaS) splits a hierarchical table into a large number of non-hierarchical sub-tables and applies the optimal method to each individual sub-table. By correctly combining the results, a sub-optimal solution can be obtained for the entire table, with a significantly shorter computing time.

The “hypercube approach” can also protect large tables by protecting the sub-tables in a certain iterative way. The protection of each sub-table also takes place sub-optimally. Consequently, the approach is relatively fast, but, in general, more cells are suppressed than strictly necessary to obtain a protected table.

2.4.2.2 Applicability

This method can be used to adequately protect quantitative tables with cells that do not satisfy the requirements of the NSI’s statistical disclosure control policy. In particular, if the table cannot be restructured further or at all, the cell suppression method can be used effectively.

The contributions to the table to be protected must not be negative² and the table must be additive. If no marginals are provided, secondary cell suppression is not needed. When marginals are provided, secondary cell suppression is usually needed to properly protect the sensitive cells.

In the modular approach, the table may be at most three-dimensional. Each dimension may be hierarchical. The limit on the dimensionality of the table is due to the fact that for higher dimensional tables, the calculation time would grow exponentially and effectively become too large.

¹ In case the table is composed of non-negative contributions and the marginals are provided as well.

² The requirement of non-negativity can be relaxed to the requirement that the values should be uniformly bounded from below. However, this requires an adaptation of the concentration rules. See, e.g., Hundepool et al. (2012).

Linked tables can be protected by copying the suppressions from one table to the other, and then protecting the tables. This should then possibly be performed in an iterative manner. The current version of τ -ARGUS is able to solve certain classes of linked tables problems automatically.

In the hypercube approach as implemented in τ -ARGUS, the table may be at most seven-dimensional. The table may be hierarchical in every dimension. Linked tables are also *possible*.

In theory, neither the modular approach nor the hypercube approach are limited in dimensionality of the tables. It is purely for performance issues, that the dimensionality is limited in the way these approaches are implemented in τ -ARGUS.

Moreover, it should be mentioned that for both approaches, from a performance perspective, the recommendation is to avoid using long, unstructured (non-hierarchical) code lists.

2.4.2.3 Detailed description

To apply statistical disclosure control techniques to tabular data, specialised software is available. In Europe, the most commonly used “generally available” software is τ -ARGUS. For that reason, the following paragraphs are dedicated to explaining methods as implemented in τ -ARGUS.

Other software packages that are available are: *sdcTable* (R package, no user interface available) and *G-Confid* (see, e.g., Statistics Canada, 2011). For a general discussion of different software tools, see Giessing (2013).

The software package τ -ARGUS has a provision to apply cell suppression to quantitative tables. If the original microdata is used as input, τ -ARGUS will determine the risky cells with the associated safety intervals.

After this, τ -ARGUS will have to determine a suppression pattern that guarantees the necessary safety intervals. There are various options for this. We will discuss the two approaches that are the most interesting for Statistics Netherlands.

2.4.2.4 Modular approach

Generally, the modular approach can be described as follows:

1. Split the hierarchical table into all logical non-hierarchical sub-tables.
2. Group the sub-tables in classes in such a way that all tables in a single class can be protected independently of each other. For a suitable classification, see De Wolf (2002).
3. Protect all tables in class K .
4. If no secondary suppressions are placed in the marginals of the sub-tables of class K , continue with class $K + 1$, including any secondary suppressions in the inside of a table as primary suppressions for class $K + 1$.
5. If secondary suppressions do have to be placed in a marginal of at least one sub-table, go back to class $K - 1$, including only the secondary suppressions in the marginals as primary suppressions.
6. Repeat steps 4 and/or 5 until all sub-tables have been protected at the lowest (most detailed) hierarchical level.

All non-hierarchical sub-tables will be protected using the mixed integer approach from Fischetti and Salazar (2000). In this approach, the required safety intervals are guaranteed, while a certain cost function is minimised. This cost function can be selected in different ways, as a result of which various forms of information loss can be minimised. This minimisation takes place *locally*, so that the ultimate solution for the entire (hierarchical) table does not necessarily also have to be optimal.

Note that in this way, the required safety intervals are guaranteed when using the subset of table relations that define the sub-table. In certain specific situations it might be possible that a required safety interval is not attained when using the complete set of table relations that defines the hierarchical table.

In selecting the cost function in τ -ARGUS, several options can be selected, including:

- A variable from the dataset (such as the quantitative value on which tabulation takes place);
- A constant (so that the number of suppressions is minimised);
- The number of contributors per cell (so that the total number of suppressed contributions is minimised).

In the disclosure control of a sub-table, also the so-called singletons problem must be taken into account: cells with only one contribution. If such cells are in a suppression pattern, the contributors involved can reverse part or all of the suppression pattern. After all, they know what their own contribution is and can therefore fill in that suppressed value, as a result of which it may also be possible to calculate other suppressed cells. In the current implementation of the mixed integer approach in τ -ARGUS, it is not possible to keep each conceivable combination of a singleton with another suppressed cell under control while searching for a suppression pattern. However, it is possible to take account of the combinations within a single row, column or layer³ in the table. The combinations which must be taken into account consist of exactly two risky cells in a single row, column or layer, of which at least one cell is a singleton. By requiring a small safety interval for the combination of these two cells, it will be made sure that even with knowledge of one of these cells, it is not possible to exactly disclose the other risky cell.

In a similar way, it is ensured that, within a single row, column or layer, all the suppressed cells together contain more than the minimum required number of contributors for a safe cell.

For a detailed description and an elaborated example of the modular approach, see De Wolf (2002). For a detailed description of the adjustments to be able to deal with linked tables, see De Wolf and Giessing (2009).

2.4.2.5 Hypercube approach

In this approach too, a hierarchical table is split into non-hierarchical sub-tables. The non-hierarchical sub tables are then protected in a certain order, where the sub-tables at the highest level are dealt with first.

³ A row consists of the cells with coordinates (r, k, l) where k and l are fixed. A column consists of the cells with the coordinates (r, k, l) where r and l are fixed. A layer consists of the cells with coordinates (r, k, l) where r and k are fixed.

For each sub-table, all possible hyper cubes are constructed for each risky cell in which that risky cell is one of the corner points. For each hypercube, the interval is calculated around the risky cell if all other corner points of the hypercube are also suppressed. If that interval is large enough (depending on the protection rule used), the associated hypercube is designated as “feasible”. The information loss is then calculated for each feasible hypercube. Finally, the feasible hypercube with the smallest information loss is selected to protect the risky cell concerned.

No linear programming problem needs to be solved in order to calculate the safety intervals resulting from a hypercube. This significantly accelerates the procedure. The hypercube approach is therefore, in general, faster than the modular approach, for which a mixed integer programming problem needs to be solved.

After all sub-tables are protected in this way, the entire procedure is repeated. Secondary suppressed cells from a certain sub-table that also occur in other sub-tables are considered as sensitive cells in those other sub-tables, and dealt with as such. This process is repeated until no more changes take place.

Note that the use of hyper cubes to protect risky cells is a sufficient but not necessary condition for a safe suppression pattern. In other words, in some cases, the combination of the different hyper cubes will not lead to an optimal suppression pattern, but it will always produce a safe suppression pattern. Consequently, this approach tends to suppress more cells than necessary for a safe suppression pattern.

This approach also takes account of the so-called singletons. A cell with only one contributor would indeed allow all suppressed corner points of a hypercube to be calculated. Therefore the extra requirement in the case of singletons is that this type of cell must be a corner point of at least two different hypercubes.

As said, the hypercube method for hierarchical tables also splits a hierarchical table into non-hierarchical sub-tables. Therefore, the protection that is provided is of a similar level as with the modular approach. I.e., the required safety intervals are guaranteed when using the subset of table relations that defines the sub-table. In certain specific situations it might be possible that a required safety interval is not attained when using the complete set of table relations that defines the hierarchical table.

2.4.2.6 Example

Using τ -ARGUS, it is easy to apply cell suppression to a quantitative table. Both the modular approach and the hypercube approach are implemented in τ -ARGUS. It is also possible to select multiple information loss measures for the cost function that must be minimised. See Section 4 for more information on τ -ARGUS.

Figure 4 shows an example of a table with some sensitive cells suppressed.

It is clear that this is not sufficient: both the cell (East, 4) and the cell (4, 9) can be directly calculated: (East, 4) = $3\,703\,896 - 15 - 642\,238 - 515\,003 - 534\,147 - 620\,392 - 1\,392\,096 = 5$ and (4, 9) = $1\,392\,096 - 145\,004 - 1\,083\,254 - 151\,870 = 11\,968$.

	Total	2	4	5	6	7	8	9	99
Total	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
- ..Nr	4,373,664.00	×	×	719,049.00	659,680.00	688,962.00	756,529.00	1,549,049.00	385.00
.. 1	1,986,129.00	×	×	398,062.00	348,039.00	354,711.00	418,778.00	466,529.00	-
.. 2	1,809,246.00	0.00	-	223,990.00	221,332.00	241,913.00	258,233.00	863,393.00	385.00
.. 3	578,289.00	-	-	96,997.00	90,309.00	92,338.00	79,518.00	219,127.00	-
- ..Os	3,703,896.00	15.00	×	642,238.00	515,003.00	534,147.00	620,392.00	1,392,096.00	-
.. 4	124,336.00	×	-	36,311.00	32,132.00	25,770.00	18,150.00	-	×
.. 5	526,279.00	-	-	93,589.00	94,957.00	110,930.00	81,799.00	145,004.00	-
.. 6	2,234,995.00	×	×	345,803.00	251,358.00	251,188.00	303,377.00	1,083,254.00	-
.. 7	818,286.00	-	-	166,535.00	136,556.00	146,259.00	217,066.00	151,870.00	-
- ..Ws	4,576,115.84	-	-	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	-
.. 8	485,326.00	-	-	63,767.00	75,442.00	87,305.00	59,953.00	198,859.00	-
.. 9	3,664,559.84	-	-	537,911.00	430,851.00	515,019.58	643,762.26	1,537,016.00	-
..10	426,230.00	-	-	47,294.00	37,277.00	61,572.00	71,417.00	208,670.00	-
- ..Zd	4,193,971.00	-	15.00	701,549.00	602,281.00	618,037.00	647,021.00	1,625,068.00	-
..11	2,752,743.00	-	15.00	488,613.00	392,395.00	363,490.00	402,925.00	1,105,305.00	-
..12	1,441,228.00	-	-	212,936.00	209,886.00	254,547.00	244,096.00	519,763.00	-
..99	-	-	-	-	-	-	-	-	-

Figure 4. Quantitative table for turnover according to region and size class

Figure 5 shows the suppression pattern that was determined with τ -ARGUS using the hypercube approach. Figure 6 shows the same based on the modular approach. Of course, in a publication, it should be impossible to make a distinction between primary and secondary suppressions.

	Total	2	4	5	6	7	8	9	99
Total	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
- ..Nr	4,373,664.00	×	×	719,049.00	-	×	688,962.00	756,529.00	1,549,049.00
.. 1	1,986,129.00	×	×	398,062.00	-	×	354,711.00	418,778.00	466,529.00
.. 2	1,809,246.00	×	-	223,990.00	-	×	241,913.00	258,233.00	863,393.00
.. 3	578,289.00	-	-	96,997.00	90,309.00	-	92,338.00	79,518.00	219,127.00
- ..Os	3,703,896.00	×	×	642,238.00	-	×	534,147.00	620,392.00	1,392,096.00
.. 4	124,336.00	×	-	36,311.00	-	×	25,770.00	-	×
.. 5	526,279.00	-	-	93,589.00	94,957.00	-	110,930.00	×	×
.. 6	2,234,995.00	×	×	345,803.00	-	×	251,188.00	303,377.00	1,083,254.00
.. 7	818,286.00	-	-	166,535.00	136,556.00	-	146,259.00	217,066.00	151,870.00
- ..Ws	4,576,115.84	-	-	648,972.00	543,570.00	-	663,896.58	775,132.26	1,944,545.00
.. 8	485,326.00	-	-	63,767.00	75,442.00	-	87,305.00	59,953.00	198,859.00
.. 9	3,664,559.84	-	-	537,911.00	430,851.00	-	515,019.58	643,762.26	1,537,016.00
..10	426,230.00	-	-	47,294.00	37,277.00	-	61,572.00	71,417.00	208,670.00
- ..Zd	4,193,971.00	-	×	701,549.00	-	×	618,037.00	647,021.00	1,625,068.00
..11	2,752,743.00	-	×	488,613.00	-	×	363,490.00	402,925.00	1,105,305.00
..12	1,441,228.00	-	-	212,936.00	209,886.00	-	254,547.00	244,096.00	519,763.00
..99	-	-	-	-	-	-	-	-	-

Figure 5. Suppression pattern for the table from Figure 4, using the hypercube approach

	Total	2	4	5	6	7	8	9	99
Total	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
- ..Nr	4,373,664.00	×	×	719,049.00	-	×	688,962.00	756,529.00	1,549,049.00
.. 1	1,986,129.00	×	×	398,062.00	-	×	354,711.00	418,778.00	466,529.00
.. 2	1,809,246.00	0.00	-	223,990.00	221,332.00	-	241,913.00	258,233.00	863,393.00
.. 3	578,289.00	-	-	96,997.00	90,309.00	-	92,338.00	79,518.00	219,127.00
- ..Os	3,703,896.00	×	×	642,238.00	515,003.00	-	534,147.00	620,392.00	1,392,096.00
.. 4	124,336.00	×	-	36,311.00	32,132.00	-	×	-	×
.. 5	526,279.00	-	-	93,589.00	94,957.00	-	110,930.00	×	×
.. 6	2,234,995.00	×	×	345,803.00	251,358.00	-	×	303,377.00	1,083,254.00
.. 7	818,286.00	-	-	166,535.00	136,556.00	-	146,259.00	217,066.00	151,870.00
- ..Ws	4,576,115.84	-	-	648,972.00	543,570.00	-	663,896.58	775,132.26	1,944,545.00
.. 8	485,326.00	-	-	63,767.00	75,442.00	-	87,305.00	59,953.00	198,859.00
.. 9	3,664,559.84	-	-	537,911.00	430,851.00	-	515,019.58	643,762.26	1,537,016.00
..10	426,230.00	-	-	47,294.00	37,277.00	-	61,572.00	71,417.00	208,670.00
- ..Zd	4,193,971.00	-	×	701,549.00	-	×	618,037.00	647,021.00	1,625,068.00
..11	2,752,743.00	-	×	488,613.00	-	×	363,490.00	402,925.00	1,105,305.00
..12	1,441,228.00	-	-	212,936.00	209,886.00	-	254,547.00	244,096.00	519,763.00
..99	-	-	-	-	-	-	-	-	-

Figure 6. Suppression pattern for the table from Figure 4, using the modular approach

For a more detailed description of the hypercube approach, see Hundepool et al. (2011, Section 2.8). References to the original literature on this method can also be found there.

2.4.3 *Waivers*

Sometimes, the need to maintain the confidentiality of the contribution of a particular respondent may result in disastrous results. E.g., just to protect that single respondent, it may be that many additional (secondary) suppressions are needed. In those cases, if the local law permits, it may be good practice to ask the respondent in question for a so-called *waiver*. That is, permission is asked to publish a table cell that contains the contribution of that respondent, even though it may not pass the primary confidentiality rule. According to the Dutch Statistical Law, waivers are permissible in economic surveys, provided a formal agreement of the respondent is present.

When waivers are used, the sensitivity rule that is used to identify the risky cells needs to be adjusted. This follows from the fact that some but not all respondents to a particular cell may have given a waiver. To adjust sensitivity rules in the presence of waivers, see Hundepool et al. (2012, Chapter 4).

3. **Design issues**

The issue here is to make the necessary preparations for protecting tables to be issued by an NSI. In order to facilitate the production of safe (enough) tables relatively quickly, it is mandatory that standardised procedures are available for the staff responsible to protecting tables. These persons will be typically scattered over an NSI, working in different departments. In this section we discuss what elements are important for such rules. However, we will not go into this matter exhaustively nor discuss the choice of the various parameters. This is impossible and depends on local circumstances in a country, and the statistical laws and practices that have to be taken into account.

3.1 *Sensitivity rules*

The criteria to test the safety of tables typically operate at the cell level. So they can be used to test which cells are considered safe and which not. These criteria have parameters that have to be specified by the NSI responsible for the disclosure control of its tables. They should be specified along with the criteria, and should be part of the disclosure control policy of the NSI. The specification, apart from the choice of the kind of sensitivity measure, is a choice for the parameters to use

3.2 *Choice of table protection methods*

To protect the sensitive cells in tabular data, the NSI has to specify what SDC methods will be used to protect the tables they want to release. There may be a choice of techniques available, but which one(s) are to be applied in a particular case depends also on the user demand.

3.3 *Longitudinal aspects*

Special attention needs to be paid to longitudinal data or panel data, in which the same entities (say businesses) yield data at several points in time. It is then not sufficient to protect the data at each point in time as if they are cross-sectional data.

4. **Available software tools**

τ -ARGUS is a package intended to protect tabular data by various techniques, such as table redesign, various versions of cell suppression, rounding and controlled tabular adjustment. For more information see Hundepool et al. (2011). This package requires a commercial LP-solver (either Xpress or Cplex)

for certain techniques (like cell suppression and rounding). The τ -ARGUS package itself, however, is free of charge. See also <http://neon.vb.cbs.nl/casc/index.htm>. Currently, non-commercial Open Source LP Solvers are investigated to be included in future versions of τ -ARGUS. In τ -ARGUS one can apply cell-suppression to unstructured tables and hierarchical tables, to single tables and sets of lined tables.

There are other packages for the protection of tabular data, such as sdcTable (R package, no user interface available) and G-Confid (see, e.g., Statistics Canada, 2011). For a general discussion of different software tools, see Giessing (2013).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- De Wolf, P. P. (2002), HiTaS: a heuristic approach to cell suppression in hierarchical tables. *Proceedings of the AMRADS meeting in Luxembourg 2002*.
- De Wolf, P. P. and Giessing, S. (2009), How to make the τ -ARGUS modular approach to deal with linked tables. *Data & Knowledge Engineering* **68**, 1160–1174.
- Fischetti, M. and Salazar Gonzales, J. J. (2000), Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints. *Journal of the American Statistical Association* **95**, 916–928.
- Giessing, S. (2013), Software tools for assessing disclosure risk and producing lower risk tabular data. Data Without Boundaries Deliverable 11.1 – Part B, February 2013.
(http://www.dwbproject.org/export/sites/default/about/public_deliverables/dwb_d11-1b_software-tools-disclosure-risk-assessment.pdf)
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and De Wolf, P. P. (2012), *Statistical disclosure control*. Wiley-Blackwell.
- Hundepool, A., Van de Wetering, A., Ramaswamy, R., De Wolf, P. P., Giessing, S., Fischetti, M., Salazar, J. J., Castro, J., and Lowthian, P. (2011), *τ -ARGUS user manual 3.5*. Statistics Netherlands, Voorburg.
- Statistics Canada (2011), *G-Confid User Manual*. Internal report.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Disclosure Control – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

1. Linear programming
2. Mixed integer programming

11. GSBPM phases explicitly referred to in this module

1. 6.4 Apply disclosure control

12. Tools explicitly referred to in this module

1. τ -ARGUS
2. sdcTable
3. G-Confid

13. Process steps explicitly referred to in this module

1. Statistical disclosure control

Administrative section

14. Module code

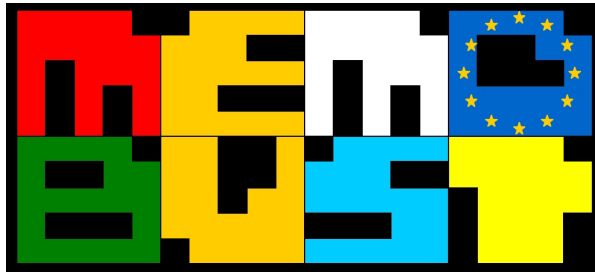
Statistical Disclosure Control-T-Methods for Quantitative Tables

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	04-09-2013	first version	Leon Willenborg, Peter-Paul de Wolf	Statistics Netherlands (CBS)
0.2	24-01-2014	revised version after review	Leon Willenborg, Peter-Paul de Wolf	Statistics Netherlands (CBS)
0.3	04-02-2014	minor revision after review by EB	Peter-Paul de Wolf	Statistics Netherlands (CBS)
0.4	05-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:30



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Specification of User Needs for Business Statistics

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Sub-processes for specification of user needs in Business Statistics	3
2.2 Examples of the specification of user needs in Business Statistics	6
3. Design issues	6
4. Available software tools.....	6
5. Decision tree of methods	6
6. Glossary.....	6
7. References	6
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

As stated in principle 11 of the Code of Practice (Eurostat, 2011) European statistics must meet the needs of users. The Quality Assurance Framework of the European Statistical System (Eurostat, 2012) describes activities, methods and tools that facilitate the implementation of the Code of Practice (CoP). One indicator related to the needs of users is to start the design of a new statistical production process for business statistics with the identification of user needs. In the GSBPM the identification of the user needs forms a preparation before the actual design for a new business statistics can start. The activity of identifying user needs starts when a need for new statistics is identified or current statistics appear to be inappropriate. Then there is a demand that is not satisfied, externally or internally, for the identified statistics. In this preparatory phase a statistical organisation has to

- 1 Determine the needs for information: what statistics, methods, sources are needed
- 2 Confirm, in more detail, the statistical needs
- 3 Establish the high level objectives of the statistical outputs
- 4 Identify the relevant concepts and variables for which data are required

Already in this stage the feasibility and necessity of the new statistics must be evaluated. For that a statistical organisation has to

- 5 Check if current data collections and methodologies can meet these needs
- 6 Prepare the business case to get approval to produce the statistics.

In these sub-processes the user needs are put in a broader scope of existing statistics and common methodologies and available data sources. These are generally six sequential sub-processes, but in a less formal context they can also occur in parallel, and can be iterative if prior assumptions have to be adapted taking into account additional information collected from users or stakeholders. So in the GSBPM before the actual design of a production process starts, the user needs are compared to the possibilities to produce them and common practice in related statistics. In this module we will follow the GSBPM approach that immediately turns the user needs into a business case, with optionally an iteration to the user needs, taking into consideration their costs for actual production.

2. General description

2.1 *Sub-processes for specification of user needs in Business Statistics*

In this section we will focus on each of the sub-processes in the determination of user needs.

2.1.1 *Determine the needs for information*

This sub-process includes the initial investigation and identification of what statistics are needed. It also includes consideration of existing practice amongst other national and international statistical organisations producing similar data, and in particular the methods used by those organisations.

2.1.2 Consult and confirm needs

Subsequently, the user needs have to be translated into specific statistical output that would meet those user needs. This includes consulting with the stakeholders and confirming in detail the needs for the statistics. A good understanding of user needs is required so that the statistical organisation knows not only what is expected to deliver, but also when, how, and, perhaps most importantly, why. For second and subsequent iterations of this phase, the main focus will be on determining whether previously identified needs have changed. This detailed understanding of user needs is the critical part of this sub-process.

2.1.3 Establish output objectives

This sub-process identifies the statistical outputs that are required to meet the user needs identified in the previous sub-process. It includes agreeing with users the suitability of the proposed outputs and their quality measures. In a new field of information, user needs can be determined by tracing and analysing existing publications in related fields. It is necessary to specify the user needs, because most users have a broad interest. This can be done by letting the potential user specify the frequency (daily, monthly), the quantity (one figure, large number), the depth (microdata) and the purpose (to report, to teach) of data use. Once a first indication of user needs being available, the statistician has to find out to what extent these needs can be satisfied by statistical information already available, either from resources within his own NSI, or from other data providers.

2.1.4 Identify concepts

This sub-process clarifies the required concepts to be measured by the business process from the point of view of the user. At this stage the concepts identified may not align with existing statistical standards. Defining the statistical output for a particular process is not an isolated activity exclusively based on an interpretation of user needs in the field of interest. We mentioned already that constraints with respect to data availability as well as considerations regarding response burden should be kept in mind. When the user needs are translated into a specification of the intended statistical output, contents are more essential than names. Therefore the choice of definitions should, in principle, precede the choice of vocabulary (e.g., variable names).

The choice of the variables (definitions and terminology) for which data are to be published is in the first place a matter of user needs. However, there are more aspects to take into account, such as coherence of concepts with existing publications as a general quality component. The choices made should as much as possible comply with international lists of concepts such as standard classifications.

The step from user needs to statistical output comprises the delineation of the target population (including the desired statistical unit type, e.g., businesses, enterprises or establishments, e.g., as defined by a legal authority or a statistical institute, and the desired coverage of the population), and comprises the identification of the variables for which data are to be produced.

In most cases several important user groups will exist with related but deviating needs on certain issues, such as individual users, users in the public domain, governmental users, local and national authorities, and commercial users. All of these different user groups should be managed properly in their needs, which may require creative solutions concerning the identified concepts and statistical disclosure control (see also the modules “Statistical Disclosure Control – Main Module” and

“Dissemination – Dissemination of Business Statistics”). In this sub-process it cannot be guaranteed that every user group is fully satisfied for cost reasons and for confidentiality reasons. The user needs should also be harmonised to accepted standards to ensure that outcomes can be compared with existing statistics.

The systematic overview of user needs leads to a detailed specification of the intended statistical output. However, after issuing the outcomes of a new survey, users will be inclined to reconsider their needs and priorities. Therefore, measuring user satisfaction as well as market research should be recurrent operations. Such monitoring is worthwhile not only to keep ahead with expanding user needs, but also to reconsider the usefulness of existing data output.

2.1.5 Check data availability

This sub-process checks whether current data sources can meet user requirements and the conditions under which they would be available, including any restrictions on their use. An assessment of possible alternatives normally includes research into potential administrative data sources and their methodologies, to determine whether they would be suitable for use for statistical purposes. When existing sources have been assessed, a strategy for filling any remaining gaps in the data requirement is prepared. This also includes a more general assessment of the legal framework in which data would be collected and used, and may therefore identify proposals for changes to existing legislation or the introduction of a new legal framework.

2.1.6 Preparation of a business case

The previous specification of user needs and available data sources and methodologies are input for a business case to get approval to implement the new statistical business process. Such a business case would typically include:

- a description of the “as-is” business process, if it already exists, with information on how the current statistics are produced, highlighting any inefficiencies and issues to be addressed;
- the proposed “to-be” solution, detailing how the statistical business process will be developed to produce the new or revised statistics;
- an assessment of costs and benefits, as more detailed data will be more expensive, and an assessment of any external constraints, such as statistical disclosure control considerations, as more detailed data are likely to create confidentiality problems. In this way both costs and statistical disclosure control counterbalance the identified user needs.

Eventually products are produced to meet the user needs. The products can take many forms, including printed publications, press releases and web sites. In this situation the user needs are monitored and reviewed regularly as they can be rather dynamic. In particular after issuing the outcomes of new statistics, users will be inclined to reconsider their needs and priorities. Therefore, measuring user satisfaction should be a recurrent operation. Such monitoring is not only worthwhile to keep ahead with expanding user needs, but also to reconsider the usefulness of existing data collection. Balancing effort and returns is then an ongoing concern. The fact that a certain data item is published is not a justification towards users that become ever more critical. See also the module “Evaluation – Evaluation of Business Statistics”.

2.2 Examples of the specification of user needs in Business Statistics

Two examples of the identification of user needs can be found in the following literature.

1. Danish experiences are described in *How to fulfil user needs* (Thygesen and Nielsen, 2012). The paper recognises uncovered needs and gives end-users assistance when they use statistics or wish to find relevant statistics. The quality and metadata are implemented with the aim to support processes at statistical organisations that handle administrative data as generic fulfilment of user needs, since user needs can to a certain extent be fulfilled dynamically by combining data from different sources. The paper contains in-depth studies and schemes based on the Danish experience.
2. Another example is the Nordic project of recognition of user needs concerning business statistics, described in *Mapping user needs* (Jørgensen, 2010). A characteristic feature of this paper is a perception of a user as a stakeholder or co-producer (co-developer). An important question in this context is the measurement and evaluation of user needs. One measure is the user satisfaction index. The Eurostat Handbook on Data Quality Assessment Methods and Tools (Eurostat, 2007) shows how it can be used in practice. A formal definition of the user satisfaction index can be found in the Guidelines for the implementation of a data quality framework for the UNCCD process (Committee for the Review of the Implementation, 2013).

3. Design issues

A specification of the user needs in terms of output tables, and a listing of the available data sources and required methods, provides essential input for the design for business statistics. If the user needs or the available data sources are not clear, the requirements for the process are also unclear. So the specification of user needs is an indispensable preparatory phase before design can start. The six sub-processes described in the previous section explain this activity in more detail.

4. Available software tools

5. Decision tree of methods

Descriptions of methods to monitor and review user needs can be provided in method modules, but at the moment they are not yet planned.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Committee for the Review of the Implementation of the Convention (2013), Guidelines for the implementation of a data quality framework for the UNCCD process. Bonn.

Eurostat (2007), *Handbook on Data Quality Assessment Methods and Tools*.

- Eurostat (2011), *European Statistics Code of Practice for the National and Community Statistical Authorities*. Luxembourg.
- Eurostat (2012), Quality Assurance Framework of the European Statistical System (ESS QAF), version 1.1. Deliverable of the Eurostat working group on Quality in Statistics.
- Jørgensen, L. L. (2010), *Mapping user needs, Nordic project 'Measuring innovation in the public sector in the Nordic countries: Toward a common statistical approach' ("Copenhagen Manual")*, DAMVAD.
- Thygesen, L. and Nielsen, M. G. (2012), How to fulfil user needs – metadata, administrative data and processes. European Conference on Quality in Official Statistics (Q2012), Athens, Greece.
- UNECE (2009), Generic Statistical Business Process Model. Version 4.0 – April 2009 (prepared by the UNECE Secretariat). Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS).

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – GSBPM: Generic Statistical Business Process Model
2. Overall Design – Overall Design
3. Statistical Disclosure Control – Main Module
4. Dissemination – Dissemination of Business Statistics
5. Evaluation – Evaluation of Business Statistics

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 1: Specify Needs

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

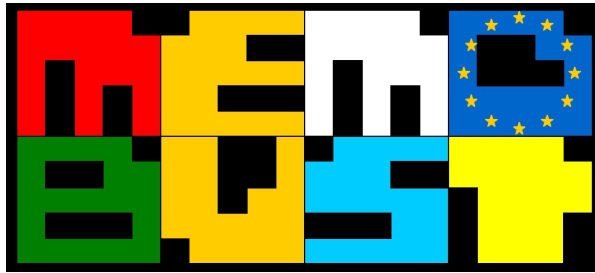
User Needs-T-Specification of User Needs

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-02-2013	first version	Rob van de Laar	CBS (Netherlands)
0.2	27-03-2013	updated version after review NL	Rob van de Laar	CBS (Netherlands)
0.3	12-11-2013	updated version after review PL	Rob van de Laar	CBS (Netherlands)
0.4	14-03-2014	updated version after review EB	Rob van de Laar	CBS (Netherlands)
0.4.1	14-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:28



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Dissemination of Business Statistics

Contents

General section	3
1. Summary	3
2. General description.....	3
2.1 Updating the output systems for dissemination products	3
2.2 Production of dissemination products	3
2.3 Managing the release of dissemination products.....	4
2.4 Promotion of dissemination products.....	5
2.5 Managing user support	5
3. Design issues	5
4. Available software tools	5
5. Decision tree of methods	5
6. Glossary	5
7. References	5
Interconnections with other modules.....	6
Administrative section.....	7

General section

1. Summary

Before the dissemination starts, the business statistics are produced, examined in detail and made ready for dissemination. For statistical outputs produced regularly, the analysis of outputs, their validation and application of disclosure control occur in every iteration, before the dissemination of the statistical products can start. In this module the dissemination of statistical products is described as in phase 7 “Disseminate” of the GSBPM (Eurostat, 2009), as related to other phases in the GSBPM, for instance phase 1 “Specify needs” (cf. the module “User Needs – Specification of User Needs for Business Statistics”), phase 2 “Design” (cf. the module “Overall Design – Overall Design”), and phase 9 “Evaluate” (cf. the module “Evaluation – Evaluation of Business Statistics”). These phases are applicable fully to the production and dissemination of business statistics. The value of statistical products does not depend only on the amount and quality of data produced but also on the use that is made of them. We refer to the Code of Practice (Eurostat, 2011) principles 11 on Relevance: European Statistics meet the user needs, and principle 15 on Accessibility and clarity: European Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance. An adequate dissemination policy applies in order to maximise user satisfaction.

2. General description

According to the GSBPM the dissemination phase of a statistical production process manages the release of the statistical products to customers. This phase occurs in each iteration for statistical outputs produced regularly. In GSBPM phase 7 “Disseminate” five sub-processes are distinguished, which are generally sequential, but can also occur in parallel and can be iterative. The sub-processes are described in the following sections.

2.1 *Updating the output systems for dissemination products*

This sub-process manages the update of output systems (databases) where data, and metadata, will be stored for dissemination purposes. It includes:

- Formatting data and metadata, ready to be put into output databases.
- Loading data and metadata into output databases.
- Ensuring data are linked to the relevant metadata.

The formatting, loading and linking of the metadata should preferably mostly take place in earlier phases, but this sub-process includes a check that all metadata are in place, ready for dissemination.

2.2 *Production of dissemination products*

This sub-process produces the products (from the output systems), as previously designed to meet user needs. The products can take many forms, tailored to specific demands and needs of different user groups, such as users in the public domain, governmental users, local and national authorities, and commercial users. These products include printed publications, press releases and web sites. Typical steps include:

- Preparation of the product components: explanatory text, tables, charts etc.
- Assemblage of components into products.
- Editing the products and checking that they meet publication standards.

In the preparation of quantitative tables an important issue is statistical disclosure control. This is an activity aimed at the protection of data that are to be released by an NSI. Protection means that individual entities (such as businesses) are not (readily) identified, and more particularly, confidential or sensitive information about such entities is not released to third parties. This to prevent misuse of data intended for statistical purposes. See also the modules “Statistical Disclosure Control – Main Module” and “Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables”. Disclosure Control is part of phase 6 “Analyse” of the GSBPM, as is also the preparation, validation (quality assessments) and finalisation of the statistical output products, immediately before the dissemination phase, but here it must be checked that disclosure control was carried out for all different user groups and dissemination channels. In many cases the production of statistics is iterative, so at first a preliminary value is derived with a lower data quality, and it is followed later on with a higher quality final value of the statistic. When planned these are called ‘revisions’, as opposed to corrections of individual values as part of the regular production process or during evaluation, or rare unplanned corrections after publication.

Publication standards are used to prevent ambiguous or unclear tables, hardly readable charts and figures, unexplained or incorrect metadata, missing or incorrect explanations in text or tables, etc. The dissemination products can take a more flexible form that allows additional user interaction. In that way the user needs can be met at the moment that a user searches for information. For instance users who want to choose which dimensions of an output table are visible, and who want to change the layout of an output table for tailor-made information. See also the module “Repeated Surveys – Repeated Surveys” for examples of iterative steps for production and validation of business statistics. Also in all cases of more flexible output products the necessary disclosure control should prevent publication of data on individual enterprises and businesses. For different user groups, different output channels can be used. For business microdata, remote access or access to data in a safe environment are sometimes possible options. In all cases statistical disclosure control may directly interfere with the user needs, and may result in dissemination solutions that do not fully satisfy user needs (see the module “User Needs – Specification of User Needs for Business Statistics”).

2.3 Managing the release of dissemination products

This sub-process ensures that all elements for the release are in place and the managing of the timing of the release. It includes briefings for specific groups, such as the press or ministers, and the provision of products to subscribers. Relevant in this context of the timing and availability of release products is Code Of Practice (Eurostat, 2011) principle 6: Statistical authorities develop, produce and disseminate European Statistics respecting scientific independence and in an objective, professional and transparent manner in which all users are treated equitably. In this principle indicators 6.3 through 6.8 are particularly relevant for the dissemination of products:

- Indicator 6.3: Errors discovered in published statistics are corrected at the earliest possible date and publicised.
- Indicator 6.4: Information on the methods and procedures used is publicly available.

- Indicator 6.5: Statistical release dates and times are pre-announced.
- Indicator 6.6: Advance notice is given on major revisions or changes in methodologies.
- Indicator 6.7: All users have equal access to statistical releases at the same time. Any privileged pre-release access to any outside user is limited, controlled and publicised. In the event that leaks occur, pre-release arrangements are revised so as to ensure impartiality.
- Indicator 6.8: Statistical releases and statements made in press conferences are objective and non-partisan.

2.4 Promotion of dissemination products

Whilst marketing in general can be considered to be an over-arching process, this sub-process concerns the active promotion of the statistical products produced in a specific statistical business process to help them reach the widest possible audience. It includes the use of customer relationship management tools to better target potential users of the products, as well as the use of tools including web sites, wikis and blogs to facilitate the process of communicating statistical information to users.

2.5 Managing user support

This sub-process ensures that customer queries are recorded and that responses are provided within agreed deadlines. The queries should be regularly reviewed to provide an input to the over-arching quality management process, as they can indicate new or changing user needs (see the module “User Needs – Specification of User Needs for Business Statistics”).

3. Design issues

4. Available software tools

In this version of the module no standard tools for the dissemination of output product are mentioned. There is only a general incentive to link data to metadata, and an indication of the many different forms the output products can take, including printed publications, press releases and web sites.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Eurostat (2011), *European Statistics Code of Practice for the National and Community Statistical Authorities*. Luxembourg.

UNECE (2009), Generic Statistical Business Process Model. Version 4.0 – April 2009 (prepared by the UNECE Secretariat). Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – GSBPM: Generic Statistical Business Process Model
2. User Needs – Specification of User Needs for Business Statistics
3. Overall Design – Overall Design
4. Repeated Surveys – Repeated Surveys
5. Statistical Disclosure Control – Main Module
6. Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables
7. Evaluation – Evaluation of Business Statistics

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 7: Disseminate

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Dissemination

Administrative section

14. Module code

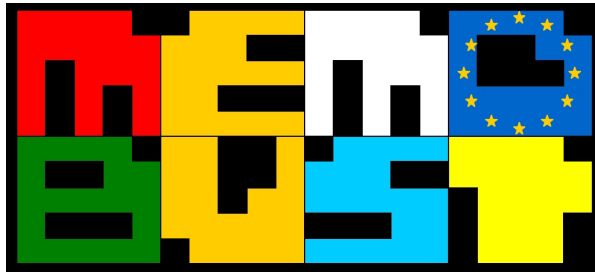
Dissemination-T-Dissemination of Business Statistics

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-02-2013	first version	Rob van de Laar	CBS (Netherlands)
0.2	27-03-2013	updated version after review NL	Rob van de Laar	CBS (Netherlands)
0.3	14-11-2013	updated version after review PL	Rob van de Laar	CBS (Netherlands)
0.4	16-01-2014	updated version after second review PL	Rob van de Laar	CBS (Netherlands)
0.5	10-03-2014	updated version after second review EB	Rob van de Laar	CBS (Netherlands)
0.5.1	14-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:23



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Evaluation of Business Statistics

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Evaluation in overall quality management vs. evaluation of specific instances of a process 3	
2.2 Evaluation as part of the overall quality management	3
2.3 The evaluation of iterations of a statistical business process	3
2.4 Quality frameworks	5
3. Design issues	6
4. Available software tools.....	6
5. Decision tree of methods	6
6. Glossary.....	6
7. References	6
Interconnections with other modules.....	8
Administrative section.....	9

General section

1. Summary

In the GSBPM the evaluation of business statistics manages both the more general over-arching process of statistical quality management, and the evaluation of specific instances of a statistical business process. In the GSBPM the latter type of evaluation is divided into three separate sub-processes. We will first describe the first type of evaluation in more detail, and next the evaluation of specific iterations of a production process for business statistics. Where appropriate a further specification to Business Statistics will be made.

2. General description

2.1 *Evaluation in overall quality management vs. evaluation of specific instances of a process*

Two levels of evaluation can be distinguished. On the one hand is the evaluation as part of the over-arching quality management, on the other the evaluation of individual instances of statistical business processes. Compared to the latter the over-arching part has both a deeper and broader scope. Both types of evaluation are complementing activities. We focus first on evaluation as part of the overall quality management.

2.2 *Evaluation as part of the overall quality management*

The evaluation as part of the over-arching quality management is dealing with several kinds of quality management: quality training, measuring quality, audit system, and quality awareness. It can also be called institutional quality.

Quality management involves the evaluation of groups of statistical business processes, and can therefore identify potential duplicates or gaps. All evaluations should result in feedback, which should be used to improve the relevant process, phase or sub-process, creating a quality loop. On the process or sub-process level plan-do-check-act quality circles are expected.

Metadata generated by the different sub-processes are used as input for quality management. These evaluations can apply within a specific process, or across several processes that use common components.

The current multiplicity of quality frameworks enhances the importance of the benchmarking and peer review approaches to evaluation. Whilst these approaches are unlikely to be feasible for every iteration of every part of every statistical business process, they should be used in a systematic way according to a pre-determined schedule that allows for the review of all main parts of the process within a specified time period.

2.3 *The evaluation of iterations of a statistical business process*

The evaluation of specific instances of a statistical business process logically takes place at the end of the instance of the process, but relies on inputs gathered throughout the different phases. As such it is the ninth and last phase in the GSBPM, after phase 8 'Archive', and part of the Quality management, over-arching all phases 1 to 9 (UNECE, 2009; UNECE, 2013). For statistical outputs produced regularly, evaluation should, at least in theory, occur for each iteration, determining whether future

iterations should take place, and if so, whether any improvements should be implemented. However, in some cases, particularly for regular and well established statistical business processes, evaluation may not be formally carried out for each iteration. In such cases, the evaluation only provides the decision as to whether the next iteration should start from a re-specification of the user needs (phase 1) (see also the module “User Needs – Specification of User Needs for Business Statistics”) or from some later phase, often the data collection phase (phase 4), or phase 2 with regard to adjustment or re-allocation of resources..

According to the GSBPM the evaluation process is made up of three sub-processes, which are generally sequential, but which can in practice overlap to some extent. The sub-processes as defined in the GSBPM are described in the following sub-sections.

2.3.1 Gathering of the evaluation inputs

The sub-process 9.1, the gathering of the evaluation inputs for the evaluation of separate iterations of a statistical business process, can use material produced in any other phase or sub-process. It may take many forms, including feedback from users (changing user needs), process metadata (for logging indicators and logging related to the efficiency of the process see the module “General Observations – Logging”), system metrics and staff suggestions. Reports of progress against an action plan agreed during a previous iteration may also form an input to evaluations of subsequent iterations. This sub-process gathers all of these inputs, and makes them available for the person or team producing the evaluation.

2.3.2 Conduct evaluation

The sub-process 9.2 analyses the evaluation inputs and synthesises them into an evaluation report. The resulting report should note any quality issues specific to this iteration of the statistical business process, and should make recommendations for changes if appropriate. These recommendations can cover changes to any phase or sub-process for future iterations of the process, or can suggest that the process should not to be repeated. Major goals of this evaluation are

1. To compare the outcomes with the targets, regarding
 - a. accuracy and other output quality components
 - b. production targets, such as resources and also Quality and Performance Indicators.
2. To improve efficiency in future production. This partly related goal is an issue in both sub-processes 9.2 and 9.3.

Some recommendations found during evaluation may be easy to implement, whereas others may need investments and studies with regard to possible side-effects. Examples of types of evaluations in business processes are given in the module “Repeated Surveys – Repeated Surveys”.

2.3.3 Agree an action plan

The sub-process 9.3 brings together the necessary decision-making power to form and agree on an action plan based on the evaluation report. It should also include consideration of a mechanism for monitoring the impact of those actions, which may, in turn, provide an input to evaluations of future iterations of the process.

2.4 *Quality frameworks*

For evaluation activities, quality assurance frameworks and institutional frameworks have the objective to establish in a specific statistical organisation a system of coordinated methods and tools guaranteeing the adherence to existing requirements concerning the statistical processes, products, and the quality of statistical systems as a whole. Recent quality frameworks from Eurostat are:

- The Eurostat European Statistics Code of Practice (CoP) for the national and community statistical authorities (Eurostat, 2011).
- The Eurostat Quality Assurance Framework of the European Statistical System (QAF) (Eurostat, 2012). The QAF identifies activities, methods and tools that provide guidance for the operationalisation of the indicators that are required to adhere to the principles of the Code of Practice. In this way it facilitates the implementation of the European Code of Practice, to which national and community statistical authorities will be judged through peer reviews and other forms of assessment at both the process level and at the institutional level, as an important instrument of the ESS.
- The ESS Handbook for Quality Reports (Eurostat, 2009) provides detailed guidelines and examples of quality reporting practices, and assists National Statistical Institutes and Eurostat in meeting the Code of Practice standard by providing recommendations for preparing comprehensive quality reports for the full range of statistical processes and their outputs.
- The Handbook on Data Quality Assessment Methods and Tools (Eurostat, 2007) aims at facilitating a systematic implementation of data quality assessment in the ESS. It presents the most important assessment methods: quality reports, quality indicators, measurement of process variables, user surveys, self-assessment and auditing, as well as labeling and certification. The handbook provides a concise description of the data quality assessment methods in use.

Some recent national quality frameworks are:

- Statistics Canada's Quality Guidelines (Statistics Canada, 2009) provides guidance with experiences and conclusions about best practices in survey design and survey methodology. With care and judgment it can also be used for other data acquisition processes.
- The Statistics Finland Quality Guidelines for Official Statistics (Statistics Finland, 2007) aims to support the development of statistics production and interaction with stakeholders for statistical surveys in the broad sense: census surveys, sample surveys, administrative registers, and derived statistical data from existing data pools.
- The ISTAT Quality Guidelines for Statistical Processes, December (ISTAT, 2012).
- Statistics Netherlands' Quality Assurance Framework at Process Level (Statistics Netherlands, 2014) integrates the CoP and the QAF at the process level (not the institutional level), relevant Dutch laws and regulations and guidelines established by Statistics Netherlands. A feature of this document is the detailed breakdown in objects, characteristics of these objects and requirements, according to the Object-oriented Quality and Risk Management model (see the module "General Observations – Quality and Risk Management Models"). This enables a

structured assessment of statistical business processes and self-assessments. Examples of these reports can be found in audit reports (for internal use only).

3. Design issues

In order to be a fixed part of a statistical business process, the evaluation itself must be designed: whether the separate phases and sub-processes are evaluated each time they are applied or according to an agreed schedule. Then the evaluation can result in timely reconsideration of the design of the process. In that case the evaluation of a business process yields indispensable input for improvement or redesign. See also the module “Overall Design – Overall Design” on the use of earlier evaluations of a business process as inputs for the redesign of a statistical process. See the module “User Needs – Specification of User Needs for Business Statistics” for possible changes in user needs over time.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Eurostat (2007), *Handbook on Data Quality Assessment Methods and Tools*.

Eurostat (2009), *ESS Handbook for Quality Reports (EHQR)*. This handbook (planned to be revised soon) is accessible on the webpage of Eurostat, currently:

http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/EHQR_FINAL.pdf

Eurostat (2011), *European Statistics Code of Practice for the National and Community Statistical Authorities*. Luxembourg.

Eurostat (2012), *Quality Assurance Framework of the European Statistical System (ESS QAF)*, version 1.1. Deliverable of the Eurostat working group on Quality in Statistics.

ISTAT (2012), *Quality Guidelines for Statistical Processes*, December.

Statistics Canada (2009), *Statistics Canada Quality Guidelines*, Fifth edition.

Statistics Finland (2007), *Quality Guidelines for Official Statistics*, 2nd Revised Edition.

Statistics Netherlands (2014), *Quality Assurance Framework at Process Level*.

UNECE (2009), *Generic Statistical Business Process Model*. Version 4.0 – April 2009 (prepared by the UNECE Secretariat). Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata.

UNECE (2013), *Generic Statistical Business Process Model*. Version 5.0 – December 2013. The United Nations Economic Commission for Europe (UNECE). See:

<http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0>

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Quality and Risk Management Models
2. General Observations – Logging
3. General Observations – GSBPM: Generic Statistical Business Process Model
4. User Needs – Specification of User Needs for Business Statistics
5. Overall Design – Overall Design
6. Repeated Surveys – Repeated Surveys

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 9: Evaluate

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Evaluation

Administrative section

14. Module code

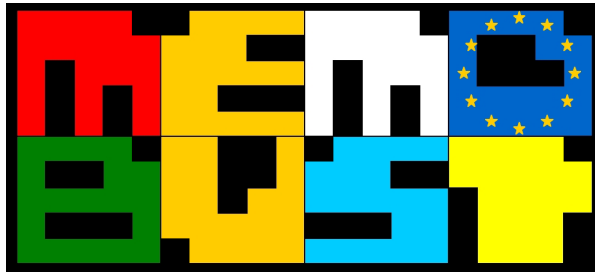
Evaluation-T-Evaluation of Business Statistics

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-02-2013	first version	Rob van de Laar	CBS (Netherlands)
0.2	27-03-2013	updated version after review NL	Rob van de Laar	CBS (Netherlands)
0.3	28-01-2014	updated version after review SE	Rob van de Laar	CBS (Netherlands)
0.4	10-03-2014	updated version after review EB	Rob van de Laar	CBS (Netherlands)
0.4.1	14-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:24



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Overall Design

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction	3
2.2 Theories and principles.....	4
2.3 Design work	13
3. Design issues	19
4. Available software tools.....	19
5. Decision tree of methods	19
6. Glossary.....	19
7. References	19
Interconnections with other modules.....	21
Administrative section.....	23

General section

1. Summary

Design refers to the design of a new survey, to re-design of a survey, and to continuous improvements in a repeated survey. Two core activities in design is to choose methods – e.g. for sampling and estimation, data collection mode(s), contact strategies, and editing – and to allocate resources to the sub-processes in the statistics production. Adjustments of allocations may dominate the work with improvements, whereas the choices frequently are more prominent for new and renewed survey designs. The aim of the design is, in principle, to find some optimum, e.g. maximum quality for a given cost. However, quality is multifaceted and depends on both uses and users, so the task to find an optimal solution has to be further developed and specified. In practice the “optimisation” rather means striving for something good and appropriate based on requests and costs rather than solving complex optimisation problems.

Mostly much of the practical design work is devoted to the accuracy of the statistics. There are further important quality components, e.g. timeliness and coherence. The “optimisation” may include one or more of these components in the search of a solution, often with trade-offs. An alternative – which may be more frequent – is to treat some components, such as timeliness, as constraints. There are further aspects, which may act both as restrictions and support. For instance, standards and common tools have a strong influence on the design. This means, for example, that the sampling and estimation methods may be chosen with regard to the practices and the IT-tools of the statistical office.

The present module gives an overall description of design and provides some general examples. There are a few handbook modules devoted to design, and there are sections within modules about specific design aspects in those topics, for instance editing and estimation. There is a topic on repeated surveys, for which more knowledge and possibilities are available when striving towards optimisation.

2. General description

This section consists of an introduction followed by two main sub-sections, one more theoretical-principal and the other more practical. There is no sharp line between the two; theory and practice should go hand-in-hand. Both have a further sub-division into a third level.

2.1 Introduction

In 2009 there was a communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade. Important ingredients in the vision are comprehensive production systems, horizontal and vertical integration in the system, and combinations of data sources, for instance directly collected survey data and administrative data. These ingredients should be considered in every survey design and be taken into at least some account. The GSBPM (Generic Statistical Business Process Model) is well in line with the vision, which is underlined by the G for Generic, see the handbook module “General Observations – GSBPM: Generic Statistical Business Process Model”. Design is the second phase, after Specify needs, out of nine phases in the GSBPM (version 2009).

In business statistics one well-known aspect of integration is a system in three layers with a Business Register (BR), primary statistics, and secondary statistics such as the National Accounts (NA).

Typically the NA sets requirements on the primary statistics, e.g. populations and variable definitions. The NA is a user with strong influence on the primary statistics. There are many EU regulations within this system, which is more encompassing than the system of (primary) business statistics. It is sometimes called the economic-statistical system.

The BR is an important basis and contributes to the co-ordination of the surveys and the statistics in the system. It holds information on units with classifications and some administrative data. It is essential when creating survey frames, delineating populations for the statistical unit(s) to be used etc. Time aspects are important to consider, e.g. the delay from an event until the information is registered in the BR. Such aspects influence the quality throughout the system. Obviously the BR should be updated regularly and frequently. The longer term “Statistical Business Register” is used to underline the role for statistical purposes, like a backbone. Concepts are fundamental, as further discussed below.

The word “design” has several aspects. The scope may be methodological or technical. Design may refer to the whole statistical survey, to a specific sub-process, to a tool, or to a system. Design is important for a new survey, when a survey is redesigned, and also in continuous improvements of repeated surveys. By first saving information – well-chosen data about the production process – during the production and then analysing these process data (often called paradata), possible improvements can be seen and hopefully also put in place: quality improvements or cost savings or both. Changing user requirements/needs, occurrence of mistakes in the production, high costs, and new laws and regulations are other examples of situations that may lead to decisions about redesign or improvement of individual sub-processes or tools. New methods can lead to redesign and improvements, too, and so can new research findings.

Before a change is implemented, the consequences must be analysed, as well as the investments that may be required. The benefits of change (increased quality, lower cost) are then compared to the cost of implementing the change and possible negative effects, for example costs of changes in IT-systems. Some changes are conveniently introduced immediately, while others should rather wait and be introduced simultaneously as a package. Risks to introduce unintentional breaks in time series must be considered, as well as opportunities to eliminate or overcome these time series breaks. Usability testing, pilot surveys, and experiments are different ways to examine the consequences, for example to test whether and how new technology influences measurements or systems.

2.2 *Theories and principles*

2.2.1 *Quality*

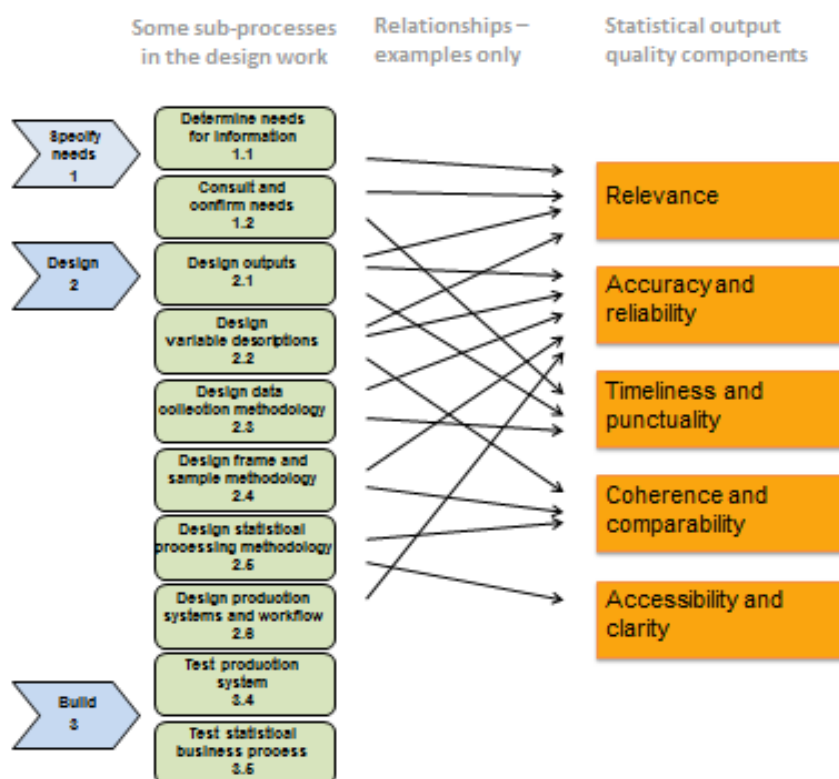
Quality of statistical output is in the European Statistical System described with five main components:

- Relevance;
- Accuracy and reliability;
- Timeliness and punctuality;
- Coherence and comparability;
- Accessibility and clarity.

See for instance Eurostat (2011) for the European Statistics Code of practice, Eurostat (2009) for a handbook (soon to be revised) on reporting quality of statistical data according to the European output quality components, and the handbook module “Quality Aspects – Quality of Statistics”.

There are several general definitions of quality, for instance fitness for use, fitness for purpose, and the degree to which a set of inherent characteristics fulfils requirements. Since quality of statistics depends on the use, the producer should work together with users (carefully selected) and specify important needs. These can be expressed as a purpose of the statistics and the quality needed for this purpose. This quality level is essential; it is sufficient for the purpose.

The figure below illustrates the complex relationship between the resulting quality of the statistical output and the three preparatory sub-processes, especially those in the phase 2 *Design*. The two phases 1 *Specify needs* and 3 *Build* also have an influence, of course. There are more sub-processes and many more relationships than those shown. The main message here is the complexity in both directions. Each quality component depends on many design sub-processes, and most design sub-processes influence several quality components.



Phase 1 *Specify needs* provides requests, which are designed and described in statistical output terms in sub-process 2.1 *Design outputs*. Sub-processes 2.2–2.5 design the production process from data collection to analyse. This is, however, not a simple mainstream route. It may be necessary to go back and change or modify some choice. Sub-process 2.6 *Design production systems and workflow* explicitly considers and ensures that the sub-processes fit together and that there are no gaps or overlaps. Phase 3 *Build* is more practical through building, enhancing, and testing the collection instrument, tools, and the production system as a whole. The survey may be tested on a small scale, for instance the data collection instrument with new variables. Tests may lead to design improvements.

There are dependencies between choices, for instance the choices of sampling and estimation procedures should be considered and chosen together, even if they are a bit apart in the GSBPM.

The user(s) may – depending on their interest(s) and their use(s) – put different priorities and constraints on each of the quality components. There may also be cost constraints. The producer has to find an appropriate balance between the interests, as further discussed in the next sections.

The user(s) may – depending on their interest(s) and their use(s) – put different priorities and different requests or constraints on each of the quality components. There may also be cost constraints, for instance budget cuts. The producer has to find an appropriate balance between the interests, as further discussed in the next sections.

2.2.2 Some basics of statistics – such as interest, target, and observation

Statistics can, perhaps most easily, be thought of in terms of statistical tables. Statistics are estimates of statistical parameters or characteristics, which can be described as follows.

- A statistical measure (e.g. sum, average or median) is used to summarise
- individual variable values (e.g. turnover) for
- the statistical units (e.g. enterprise) in a group.
- The totality of considered statistical units is called the population.
- There are sub-populations; also called domains of estimation.
- There are reference times for variables, units etc.

The producer has to find an appropriate balance between the possibly many interests of the users, for instance for the variables. They are here described first. These variables, which the users demand or express a need for, are called variables of interest. The producer has to consider trade-offs between

- different variables of **interest**;
- the possibilities to collect this information with regard to quality, costs and response burden.

This balancing and cognitive judgements result in:

- **target** variables, that is the variables of the statistical estimates/output;
- **observation** variables, that is the variables to be collected either from administrative data (or other accessible registers) or directly from respondents.

The observation variables may be the same as the target variables. Another possibility is that the target variables can be derived from the observation variables, for instance through summation. A third possibility is that a model with some assumptions is needed to arrive at the target variable. (For instance, invoice value is fairly simple for the respondent, in comparison with value of production and with statistical value in the Intrastat system for trade between EU countries.)

Similarly, statistical units and populations have to be considered with respect to interests, suitable target, and possibilities to observe.

2.2.3 Accuracy, errors, and total survey error

There is considerable potential and normally also need to work with the quality component accuracy in the design phase in order to influence its size. The accuracy – or, conversely, the inaccuracy or the

uncertainty – depends on errors of various types. The errors have different causes and characteristics, and they should be handled accordingly. Considering their effects on costs they should – depending on possibilities – be eliminated, minimised, reduced, or possibly ignored.

In the ESS quality concept there is a breakdown by sources of error:

- Sampling;
- Coverage;
- Measurement;
- Non-response;
- Processing;
- Model assumption.

Some errors can be avoided, but far from all. Some errors are unknown, or perhaps rather not yet known. New technology implies new error types and new error structures. Some errors will become known when the evaluation phase is run. The list of error sources and causes is not constant, nor are the interactions among errors. A debate taking place right now (especially for surveys to individuals and households) is that the non-response error does not have a strong relationship with the non-response rate. An intense search to get responses from non-response units may become expensive with little effect, and measurement errors may increase for units with intensive follow-up.

It is important for both the user and the producer to have knowledge of structures and sizes of errors. The most interesting for the user should be the quality of the statistical output (the product). That means, for example for the non-response, that it is not its size or rate that is important but its effect on the estimates. The producer needs a more thorough knowledge of the errors in order to direct resources towards reducing the errors that have a major impact on the final product.

Two important issues in the example of non-response are: if the non-response pattern leads to bias, and if there is auxiliary information that can reduce the bias. Methods and formulas are available to handle some, but not all, types of uncertainties and errors. The method of calibration weights can be used to compensate at least partly for the error sources sampling, coverage, and non-response (see the handbook module “Weighting and Estimation – Calibration”). For measurement and processing errors there is no correspondence yet. Administrative data can be more difficult to evaluate than directly collected data. The inaccuracy is often related to the requirements related to the administrative use. Some errors can be estimated only after some event and some time, such as coverage when the registers are updated. Studies in retrospect may be useful to match the data and to make estimates of various anomalies and errors. Models may be useful in the assessment of errors and resulting quality. There is much to be done methodologically in terms of quality in administrative data and statistical registers, also in statistical inference and design. See for instance Zhang (2012), Laitila (2012), the handbook module “Data Collection – Collection and Use of Secondary Data” and references there (and also the handbook module “Statistical Registers and Frames – Quality of Statistical Registers and Frames”).

The sensitivity to errors is different between estimators of levels and estimators of changes and flows. If only certain types of errors are included in calculations and estimates of size, this must be understood. An uncritical use of the estimate thus obtained as a measure of the total error must be avoided. A simply calculated mean squared error is often a too “optimistic” estimate which takes into

account some error sources but not all. Typically, the effect of a random sampling part may be included, while for example systematic measurement errors, non-response errors, and coverage errors are not.

Total survey error, which has been much discussed, is described for instance in two fairly recent overview papers: one by Biemer (2010) with the subtitle “Design, implementation, and evaluation”, and the other by Groves and Lyberg (2010) with the subtitle “Past, present, and future”.

2.2.4 *Quality Assurance and Quality Control*

Quality assurance has two main aspects:

- Approaches and methods to achieve the intended/stated quality.
- Providing confidence that the quality requirements will be met.

For the statistics to achieve the quality that has been stated the following is needed: a good and realistic planning, control of the production, as well as assessments and checks on the quality of processes and the final statistical product.

To use proven techniques, methodologies, checklists, etc. have several positive effects. It is for example easier to predict end product quality and to avoid situations where the desired quality is not achieved. Common methods, tools and practices thereby contribute to the quality assurance of individual statistical products.

While quality assurance is everything you do to get a good quality, quality control is verification that the quality achieved was as expected. Quality control is used to monitor that the planned methods, tools, routines etc. are used, operating as intended and result in the intended quality. Checks may be of various types. Checking that design specifications are followed may be necessary. It is important that quality control is used to control and also to improve each process that does not work as intended.

Hence, the survey design influences both quality assurance and quality control, and conversely. There is a close similarity to fitness for purpose, where the purpose includes a quality level.

Some references here are Eurostat (2007), Eurostat (2012b), and Lyberg (2012).

2.2.5 *Theory and criteria for some survey parts and sub-processes*

There exists no coherent theory for the design of statistical surveys. However, a variety of theories can be used, singly or in combination in various sub-processes. Such theories can be used to select appropriate methods or at least get assistance in the choice. Some examples of theories and principles follow below.

- For sampling and estimation there are theories with clear criteria for achievements (probability sample, minimising the mean squared error MSE, small variance, minor/small systematic error) and for many situations, even formulas that make it possible to calculate what is best or at least good. This area is more highly developed than many others in theory. See for instance the handbook modules “Sample Selection – Main Module” and “Weighting and Estimation – Main Module” and references provided there.
- Theories for the response process for different types of surveys, respondents, and data are developed in the behavioural sciences. Measuring techniques utilise theories and experiences

in order to avoid or reduce measurement errors (such as reducing response error as a result of difficult words and memory errors). Response processes for business surveys are less well known, but they are gradually developed. See the handbook module “Response – Response Process” and the recent book by Snijkers, Haraldsen, Jones, and Willimack (2013).

- For direct data collection there are theories and knowledge of advantages and disadvantages with different methods (questionnaire, telephone interview, visit interview, etc.) in different situations and circumstances (for example subject, cost, and time). In some situations, the choice of data collection method is evident, but in other situations discussions are needed. For example, telephone interview is a data collection method that can be implemented quickly but that is not suitable for all types of questions and question structures. See for instance the handbook main theme module “Data Collection – Main Module” and further modules on design of data collection and mixed mode. Snijkers et al. (2013) describe some practices.
- Editing is part of the quality control, specifically the quality control of data collection. The design of the editing is included in the survey design. Previously there was in many cases “over-editing” with too much time spent on units with little influence on the estimates. Nowadays statistical approaches have led to methods such as selective editing and macro editing, using resources in a cost-effective way. It is, of course, important to know how the statistics will be used – which estimates are needed with what accuracy. See the handbook module “Statistical Data Editing – Main Module” for an overview.

2.2.6 *What is optimal?*

An ideal in survey design is to achieve a “best” or “optimal” solution. The minimum sample size for a given accuracy requirement is an example, which refers to a specific part of the survey design and where the wording is in terms of an optimum. It may, in specific examples like this one, be possible to compute an optimal solution, when other factors and conditions for the survey have been defined, such as statistical units, population, variables, etc. Such situations and sub-processes – where the task can be formulated as an optimisation problem with a solution – are fairly rare.

Design work focuses on and around quality and costs. An optimal design is then the design providing the highest quality for a given cost, or conversely, the design that achieves a certain quality at the lowest cost. This is the classic efficiency criterion for the planning of a survey, expressed from two points of view. However, quality is a multifaceted concept, which needs to be separated into its components. Then the components are studied and balanced.

Considerations of and between quality components that are “pulling in different directions” are often included in the optimisation discussions and design work. Accuracy and timeliness provide an example of such a conflict. The accuracy can be increased by follow-up work on non-response and editing signals. This takes time, though. Many other conflicts exist, for instance between reduction of non-response and reduction of measurement error.

Since there are so many sub-processes, possible combinations, and conflicts to consider, it is hardly meaningful to talk about an optimisation problem with a global optimum. The optimisation is rather an overarching principle. The implication is that careful and often iterative work is used to strive towards a solution, which is good and without flaws. There is a need to formulate principles and restrictions for this “optimisation” at an early stage. The way is not to search for many optimal sub-processes. Rather,

the focus should be on factors and sub-processes that are deemed highly influential on quality and costs for the particular survey.

It is, of course, easier to make computations for a certain part of the survey, like the sample design, than for the whole survey. Yet, the whole survey must be taken into consideration, using earlier experiences as a starting point. It may be necessary to collect new information to enable calculations or to make reasonable assumptions and assessments.

2.2.7 What is included in the optimisation?

Below are some examples of factors and conditions that have to be considered in the “optimisation” procedure. This procedure searches best and good solutions for the design. In a few cases there may be a best choice or allocation, achieved in a formal optimisation procedure. Mostly the search for an “optimum” is largely non-formal – but still taking valuable knowledge into account in making choices and allocations. Many factors imply constraints for the design, and they are not included in the search as such.

- There are national regulations for the statistical office, e.g. so that a user-desired level of detail may not be achieved due to rules for disclosure control.
- There are survey-specific international requests or recommendations to take into account.
- There are general national and international standards, recommendations etc. to take into account.
- There are rules for data collection and requirements to reduce the response burden. It may, for instance, have effects on the level of detail and on the survey variables and the questionnaire.
- The users may have requirements, e.g. on timeliness, which limits the possibilities in the design work.
- Quality is related to the use. It is important that the user dialogue clarifies the constraints (if any) and what aspects should be included in the optimisation work.
- Lack of resources can influence and constrain the possibilities for a survey.

Although the optimisation seemingly can be expressed simply – to minimise the cost for a given quality or maximise the quality for a given cost – the optimisation itself is not a simple calculation. The situation is from the optimisation perspective described in terms of constraints and room for manoeuvre. Accuracy often dominates the design work. Other quality components are constraints in many cases.

Certain constraints are set out in the user dialogue, early or gradually. Perhaps the most common constraint is that the financial resources are limited. Timeliness is an obvious quality aspect, easy to require. The dialogue should be allowed to take time to explain also quality aspects that may be less obvious to the user but nonetheless important. As said before, the users should be carefully selected to have a broad view and an interest in discussing balances and trade-offs. A rough design and plan of the required production should be made during the user dialogue to find out if the goals are realistic.

When the user dialogue is completed, further work includes most of the following major issues: laws/regulations, quality requirements, quality wishes, response burden, staff, and costs. The work differs, of course, between a new survey, a re-design, and continuous improvements. In the latter cases there is already a starting point and experience, mostly both quantitative information and qualitative

information: what went well, what could be improved, possibly some idea of how. Knowledge of relationships between quality/errors and costs seems limited, at least generally available. Groves (1989) is a basic and early source: a book on survey errors and costs. Linacre and Trewin (1993) give an interesting practical example from a statistical office, a rare and classic example. Already a rough model can be of good guidance in the design work. Marella (2007) discusses errors with a perspective of total error and costs. Lyberg (2012) makes an extensive overview of survey quality; covering many aspects, for example quality management, process quality, product quality, and also total survey error.

Questions about design, such as those discussed, about effects on quality from different choices and allocations are difficult to answer. The product manager usually needs help from experts, including methodologists, cognitive experts, and IT-professionals.

2.2.8 Resources, intensities and their allocation

Design does not simply consist of choosing methods, but it is also about “intensities”, i.e., the extent to which each method is used. Examples of intensities are the sample size, amount and focus of reminders, validation levels, etc. For the sample size it is often relatively easy to calculate how an increase of the sample size reduces sampling error. Similar considerations can be made for other sources of error, although it is often very difficult. An increased intensity requires more resources. It can bring both positive and negative effects. For example, a further reminder may give an increased data inflow, but the response quality may be worse, and there will obviously be fewer resources available to reduce other sources of error.

The choice of the intensity involves both the specific sub-process – what advantages and disadvantages an increased intensity implies – and the full set of sub-processes. Where should the “last euro” be used – increased sample size, more tracking of non-respondents, more work on questionnaire and instructions, and so on? These important questions have no easy answers. Overall experience of process implementers and methodological expertise should be utilised.

2.2.9 Metadata and other information for different purposes

Users need documentation of various kinds in order to understand and use data – microdata and macrodata (statistics) – properly. A user need is information about data, so-called metadata, for example definitions of variables and quality information. From a user’s perspective, metadata have two main objectives: (i) to make it easier for the user to find relevant data, given an information problem, and (ii) to help the user to interpret and analyse data. Different users have different needs for metadata depending on usage, experience, and competence.

For the producer of statistics, metadata are also used in the production processes to control the processes. The producer needs detailed information about the processes behind the data. Such data are known as process data or paradata. Data on production and how it works should be produced and saved for several reasons. Paradata contribute to information on both process quality and product quality. Hence, the collected paradata is one of several sources for evaluation and feedback from the statistics production stage to future production. This applies especially to successive rounds in repeated surveys, but not only. Lessons can be learnt also for similar surveys.

Metadata and paradata can be used as “drivers” of the production system. This is called a metadata-driven production system, and there may similarly be a paradata-driven management/survey. The latter

terminology is quite recent. Paradata can be used dynamically to modify or change operations in the production. From the large number of possible paradata it is important to select and save those that are most useful to improve process quality and efficiency. One use is to adjust the intensity of the process, for example invest more or less on non-response follow-up depending on the results of the analysis of paradata. An alternative is to change the process. For example, do not intensify non-response follow-up but try to find ways to increase response with motivation and facilitation for the respondents.

Efforts to provide documentation and metadata for products and processes are often perceived as costly. It is therefore important to design the processes so that documentation, metadata, and paradata are as much as possible automatically generated as by-products from the processes. The GSBPM is about to be complemented with GSIM, the Generic Statistical Information Model, which will facilitate communication. For a short introduction see the paper by UNECE (2013) describing GSIM and a few other initiatives, for example the Common Statistical Production Architecture (CSPA), and including some links. Eurostat (2012a) describes an ESS strategy, which is more technical. It mentions for instance the CORE (Common Reference Environment) architecture and that bridging between CORE and GSIM is under development.

Documentation should thus be generated, written, and saved continuously, not postponed until a product is designed or a production process round is completed. It is important to document not only the production phases 4–7 (Collect, Process, Analyse, and Disseminate, respectively) of the GSBPM, but also the preparatory work, i.e., user requirements and preferences, the choices made and the reasons for these choices. A detailed documentation should be available for the internal users, and a less detailed one for the external users, with focus on the usages of the statistics.

There is a European standard for quality reporting; see the handbook by Eurostat (2009), which is now revised. There is since long international cooperation on metadata standards. There is a recent proposed integration of the two structures stated below, where ESQRS includes the revised handbook.

- ESS Standard for Quality Reports Structure (ESQRS)
- Euro SDMX Metadata Structure (ESMS)

The result of the integration is a framework for both quality reporting and reference metadata: the Single Integrated Metadata Structure (SIMS), see Eurostat (2013).

When no paradata are available, a pilot study may be made, see for instance the handbook module “Repeated Surveys – Repeated Surveys”.

2.2.10 Architecture and infrastructure

There are several types of resources that must be taken into account in the planning, in addition to human and financial resources. An example of such resources is the technological infrastructure that the statistical office has at its disposal. The standardised process and information system architecture provides another example. The architecture and infrastructure imply certain constraints in the planning. They also provide a springboard for new products, which do not need to be developed “from scratch”. These new statistical products can benefit from and build on standard solutions and standard components of the existing architecture.

Statistical production is in many statistical offices moving from tailor-made stove-pipes for single surveys/products towards architecture with re-use of data, common tools, data warehousing, statistical

systems with services etc. Standards simplify exchange of data and metadata, for example between sources, surveys, and countries. Standardisation is a key word – and a word with many meanings and many different interpretations. It is sometimes over-interpreted to mean one and only one method for sampling (estimation, editing and so on) irrespective of the preconditions. It is a challenge to find a balance between standardisation and flexibility, for instance to design and build a common tool which is functional and user-friendly for many. Obviously architecture needs to be well designed and planned for the future. Methodology is an important part, to foresee future needs of statistics and data, future sources and collection possibilities, methods of statistical inference etc.

Two examples, among many, of work in statistical offices on standardisation are provided by Merad and Brodie (2011) on UK sub-annual business surveys and by Godbout (2011) on post-collection processing in business surveys at Statistics Canada. These examples are standardisation in several ways, such as contents, methods, and tools. Hofman (2011) describes redesign at Statistics Netherlands with the aim to improve efficiency and quality of key statistics. Again, this involves statistical design, with special advice for methodology, and software architecture.

The Journal of Official Statistics (JOS) devotes its first issue in 2013 to “Systems and Architectures for High-Quality Statistics Production”, see JOS (2013). Several national statistical production systems and ongoing changes are described. Eltinge, Biemer, and Holmberg (2013) present a potential framework, including for instance (i) survey, quality, cost, and stakeholder utility, (ii) integration of system architecture with models for total survey quality and adaptive total design, (iii) possible use of concepts from the GSBPM and the GSIM, and (iv) the role of governance processes in the practical implementation.

The previous sub-section mentions GSIM and some more technical initiatives, like CSPA and CORE.

2.3 Design work

2.3.1 Teamwork

The design work is teamwork. Such a team, which is devoted to elaborate the design of a survey, should include at least the competences of a subject-matter statistician, a methodologist, an IT-expert, a dissemination specialist, and selected persons on behalf of the users: either external representative(s) or internal knowledge, for example National Accounts. Design work itself is an iterative process, which needs co-ordination in order to build effectively on the different kinds of expertise. It is important to be aware of the possibilities to influence the design: the first time, a redesign, and – the frequent option – continuous improvements. There are many choices and allocations to make, and there should be paradata and experience to summarise. The survey manager has an important role.

2.3.2 Some different situations

There are some differences between one-off surveys, repeated single surveys, and a system of surveys. Business statistics produced in a statistical office are largely a system with co-ordinated repeated surveys. The EU regulations put requests and restrictions on national surveys, motivated by comparability and the European perspective, and there may be additional national requests. Overall design summarises and balances the different parts: the sub-processes of a specific survey and also the different surveys in the statistical system.

A brief overview of some important steps to consider in the overall design follows; they are not necessarily all relevant in the individual case.

- Fulfil laws and regulations, nationally and internationally; state the influence on statistical units, populations, variables, level of detail with regard to disclosure control, timeliness, revisions etc.
- For repeated surveys: include further time aspects and possibilities to utilise process data.
- For surveys in a statistical system: include further possibilities and restrictions, e.g. data for editing and coherence of the statistical outputs.
- Specify optimisation and constraints in the design work for the survey, including the choices and the allocations when balancing the sub-processes of the survey.

2.3.3 *The economic-statistical system*

A survey in a system, for instance the economic-statistical system in European official statistics, normally is subject to a regulation, or a set of recommendations, or both. These may apply to the statistical output or more, for instance timeliness and naturally, domains of estimation, statistical units and variables. Some different issues follow.

- There is a business register – or rather a statistical business register – providing frames, which are subject to quality requirements.
- The system is a basis for coherence between statistics.
- The system is essential for national accounts and other secondary statistics.
- A system approach is likely to enable lower response burden by its joint perspective.
- The level of detail may be an issue, depending on the requirements from different stakeholders and users.
- There is international work on classifications, systems of output statistics etc. It is essential for comparability between countries and other geographical regions.
- There may be some conflicts or differences between national and international needs. There may be ways to resolve different needs, e.g. to fulfil both.
- There may be some common development of methods and tools, e.g. for seasonal adjustment.

Producing business statistics in an EU country means that much is already settled. There are different degrees of regulation and freedom when it comes to variables and other content parts, statistical output, quality achievements etc. There are surveys that are not in the system (yet), and there is some “freedom” for surveys in the system. Quality management and cost-effective production are important in both cases.

2.3.4 *Specify information needs*

The general starting point of design work is to clarify and specify the information needs. It is most important for the outcome of the specification that the assessment is made both for and with users. Pre-requisites for a good result are communication skills, and also skills and understanding of basic scientific methodology. In principle, each individual user has unique information needs. This makes it impossible to standardise the design work to a high degree or assume that the design is a quick fix. However, there are procedures which will increase the possibilities for the design to be good and

dedicated. Documentation is important, and not only decided actions but also underlying reasoning should be included. This makes it easier for those involved later to understand and communicate.

The design should take advantage of the flexibility and experience in different fields that the statistical office has. It is especially important to do a thorough job when it is a redesign or a completely new survey that will be repeated. The knowledge acquired on the information needs determines the appropriate type(s) of study or survey: a planned experiment, a statistical survey, an observational study, re-use of existing data, or any mix of these. Typically, for a survey, fundamental concepts such as target population, target parameters and major domains of estimation must be considered.

Depending on the type of survey it may be necessary or desirable to have contacts with users “along the way”. This may involve early warnings of possible problems, in spite of the planning, and decisions to be taken about adjustments.

There are mostly several uses and many users. This may – and often does – imply conflicting demands. These demands should be communicated with users, at least together with stakeholders and a selected set of important users. There are handbook modules related to this: “User Needs – Specification of User Needs for Business Statistics” and “Evaluation – Evaluation of Business Statistics”.

2.3.5 Concepts, level of detail, and accuracy requests

The work leading to the statistical output characteristics involves much communication with the user (one or usually more). Conceptualisation and conceptual modelling are central – conceptualisation is one of the most difficult and most important tasks of the user dialogue about the statistical output. It involves variables, statistical unit types etc. It involves iteration between user interests and response burden, including discussions where definitions are made operational and possible errors are assessed. Contents of business accounting systems should be taken into consideration as well as different user needs, including the economic-statistical system with the national accounts (as already mentioned in Sections 2.3.3–4 above). In the case of a redesign, the same questions may be raised in discussions with the user(s), but it is likely that the issues are more specific (reasons behind the re-design), more detailed, and also operational. The natural starting point is experience and earlier data. Forbes and Brown (2012) discuss conceptual thinking. There are specific examples in many different modules, for example “Statistical Registers and Frames – The Statistical Units and the Business Register”.

Coherence and comparability are important to consider early, as has been already indicated in terms of the economic-statistical system. They influence types of statistical units, populations, and variables. Such issues have been mentioned repeatedly above, both national and international perspectives. It is valuable to have classifications and other metadata (such as variable definitions and value domains).

The level of detail for the output needs to be discussed thoroughly and confirmed with the user. In general, the greater level of detail, the greater the costs. By providing a variety of design options with different degrees of detail or indeed different degrees of accuracy, the user gets a picture of marginal costs. It also gives the user choices and options, which may modify the preferences.

A further aspect to include in the communication is provided by the type of requests on accuracy. Are all domains of estimation equally important? How should the accuracy be expressed; in absolute or relative terms? The choices made have a considerable influence on the allocation of a sample. See for instance the handbook module “Sample Selection – Main Module”; the main theme with references.

When estimating the accuracy all sources of errors should be considered and included. Non-response and coverage deficiencies will occur, for instance. It is usually wise to consider such influencing factors already at the design stage. There are also possible deficiencies in measurement, for instance due to the statistical concepts used, possible difficulties with statistical units like Kind of Activity Unit (KAU), and difficulties to distinguish national and international activities.

Some further aspects are ethical and legal rules and also policies. They may restrict the survey design. For instance, disclosure control may affect the level of detail; see the handbook module “Statistical Disclosure Control – Main Module” and references provided there. This has to be considered early, already together with user needs and design. There is otherwise a risk of collecting data without being able to publish the planned detailed statistical tables – possibly with a larger sample and higher response burden than motivated.

The emphasis in the descriptions here is on statistics, but the output may alternatively be microdata or both micro- and macrodata.

2.3.6 Some specific parts of the design, which often are important

One of the more important decisions to make early in the design is about data collection. There are a few major issues related to sources and modes.

- Are there existing data (administrative data or other registers) which can be used? This means lower response burden and normally also lower costs and shorter production time. However, there may be a delay in administrative data in comparison with direct data collection. It may be motivated and cost-efficient to put some effort into statistics production based on such existing data, e.g. into editing these data and into building models to enhance the contents. See the handbook module “Data Collection – Collection and Use of Secondary Data”.

It may be motivated to mix direct data collection and use of administrative data. For instance, data could be collected directly from the large and often complex enterprises, whereas administrative data are used for medium-sized and small enterprises. There may be a delay in administrative data for small enterprises, though. See the handbook module “Weighting and Estimation – Estimation with Administrative Data”.

- In case of direct data collection the collection mode(s) should be chosen with regard to important factors, such as character of variables, timeliness, and cost. See the three handbook modules “Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method”, “Data Collection – Design of Data Collection Part 2: Contact Strategies”, and “Data Collection – Mixed Mode Data Collection”.

Information needs, concepts, levels of detail, and accuracy (at least first notions) are all aspects to include early and continuously in the design work. Comparability and coherence requests are included. Type(s) of statistical unit(s), population(s), and variables need to be considered. It may be wise to use further variables and also types of statistical units in the data collection than may first seem necessary to build the target statistics. This may lower the response burden and increase the quality of the data. Effort should put into variable definitions, formulating questions, and questionnaire design, see the handbook module “Questionnaire Design – Main Module” and further references there. It may be an option to combine direct data collection with existing data, thus reducing the amount of questions and

the sample size. See for instance the handbook module “Data Collection – Collection and Use of Secondary Data”.

Further parts are frame construction and sample design (if relevant), including co-ordination with other surveys/statistics and, again, considering response burden. The time of the collection depends on several factors, for instance availability of data, suitability for respondents, and timeliness of the statistics. Reminders should also be designed appropriately. The estimation should be considered together with the sampling, including the use of auxiliary information in either or both steps. Editing and imputation are examples of further aspects to take into account. Editing is done in several sub-processes, which should be balanced appropriately. This provides an important example of so-called “intensities” (discussed in Section 2.2.8). The cost of editing is often a considerable part of the total costs, so it is important to allocate the resources in an “optimal” way: both the editing share of the total and the allocation to sub-processes within editing. See the discussion in the handbook module “Statistical Data Editing – Main Module”.

Response burden should be considered as a special issue. Many countries have goals on reduction and on a low burden, especially for small business. Work should be done both for each survey and for the system of surveys with regard to response times, avoidance of double reporting, and possibilities to use administrative or other already accessible data. The sampling procedure could include co-ordination between surveys and over time. It is easier for a business to participate in a certain survey for a limited period than it is to jump in and out of several surveys. This is discussed in the handbook module “Sample Selection – Sample Co-ordination” and a few related modules.

The later phases of the production are perhaps designed in less detail at an early stage. What analyses to make, for instance, may be more suitable to consider later. However, accessibility to variables and enough time should be included from the beginning. Similarly, publication and other communication and deliveries should be planned in time but mostly not designed in detail early. Some further aspects to design and plan follow. Automatic procedures and little manual work to be done under time pressure is mostly a desirable target, especially towards the end of the production. It is important to study the output when it is first produced: Is it reasonable and is the quality as expected? Another check – related, but partly different – should be made just before publication or delivery: Have the intended tables been included with the correct contents, is the explanatory text as intended with the figures correct etc.

Design includes the organisation of the work with staff, team work, and responsibilities; more about this is described in the next section. There is also the production system (phase 3 in the GSBPM), hardly considered in this handbook. It is, of course, important with a system that works smoothly and is well tested in advance. This is an investment. It is time-consuming and often expensive to go back and re-start early processes due to failures discovered later on.

2.3.7 Responsive and adaptive design

The term responsive design is relatively new to survey methodology. In other statistical areas, adaptive design existed for quite some time as a way of working for instance with clinical trials, where trials are not optimised until sufficient information is accumulated. For surveys both terms responsive and adaptive are used, often with adaptive designs being somewhat broader, see for instance Schouten, Calinescu, and Luiten (2013). Business statistics may have used some of the ideas of responsive

design before the term came into use, but often in a less formal way. Schouten et al. (2013) provide some examples, and there is reasoning in the handbook module “Data Collection – Design of Data Collection Part 2: Contact Strategies”.

A simple example of planning for a responsive design with possible adjustments is to have milestones in the production, especially during data collection. At certain times or production situations there is a pause to see how the production works and to make appropriate adjustments. This applies, for example, to the data inflow or to the examination of questions: Is the data inflow sufficient in all strata (or correspondingly)? Does the editing run as expected or are there worrying error signals? If justified, take actions, for example for follow-up or re-contacts. The allocation between groups may be adjusted, for instance, or a more expensive data collection mode may be used if motivated according to the design.

Reasoning of this kind shows the importance of paradata. By measuring the production process and studying the paradata, the continuous process can be controlled. The responsive design means that the design is prepared for adjustments to be made, in a scientific way, so that the production process is safely improved. Adjustments of processes should, of course, be in line with design and randomisation principles used; not be too data-driven. For repeated surveys previous rounds of production may provide useful information for an adaptive design.

2.3.8 The plan and assessment of its sustainability

The time frame for a product must be clear and communicated to all involved. Internally at the statistical office, it must be much more detailed than it is to customers. The preliminary plan must – especially for a new but also for a redesigned survey – include:

- A plan for deliveries and publications;
- Conceptualisation of types of statistical units, population, and the main variables in data collection and production, and an outline of questions in direct data collection;
- A draft production-flow with methods and tools, a rough picture of the IT-solution, a list of tools and systems to be built or modified and tested;
- Plans for a pilot study and other quality assurance efforts;
- Resource requirements, broken down into key competencies and time when they are needed;
- A plan for organisation of the work, e.g. how to distribute workload and responsibilities.

The preliminary plan is successively refined and adjusted.

In statistics production with continuous improvements, the planning basically means evaluation of the previous production round(s) during a suitable period followed by appropriate adjustments of the earlier plan.

The first proposals for the survey design normally need to be revised, as part of the iteration which gradually approaches the final design. Reasons for revisions may include lack of resources or that a more careful analysis shows higher costs than expected.

When a tentative plan is developed it must be reconciled. Some important questions are:

- Does the plan fulfil the promises to the user?
- Has the response burden been sufficiently taken into account?

- Are all tools and a production system in place? Can remaining additions and modifications be ready in time?
- Are sufficient amounts of time and resources allocated to testing?
- Have paradata, metadata, and other documentation been prepared and scheduled?
- Are the necessary personnel resources available for the production, or must changes be made?
- Have quality controls been built?

The responses to these questions could lead to reassessments and revisions of the plan. The results of pilot studies and performed tests may also lead to such reviews.

2.3.9 The “optimisation”

As stated several times design work is not just stating and solving an optimisation problem. It rather involves finding influential and critical sub-processes, which then are studied and tuned. Re-use of well-known methods and tools has many advantages. Successive improvement work may improve quality or reduce costs considerably. Paradata are needed and personnel resources.

3. Design issues

/Already treated above/

4. Available software tools

5. Decision tree of methods

/Not on this high level, but for specific parts treated in other modules/

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Biemer, P. P. (2010), Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly* **74**, 817–848.
- Eltinge, J. L., Biemer, P. P., and Holmberg, A. (2013), A Potential Framework for Integration of Architecture and Methodology to Improve Statistical Production Systems. *Journal of Official Statistics* **29**, 125–145.
- Eurostat (2007), *Handbook on Data Quality Assessment Methods and Tools*. Output from a Eurostat-project with the editors in Wiesbaden. Eurostat webpage.
- Eurostat (2009), *ESS Handbook for Quality Reports*. Eurostat Methodologies and Working papers.
- Eurostat (2011), *European Statistics Code of Practice*. For the national and community statistical authorities. Adopted by the European Statistical System Committee 28th September 2011 (revised version).

- Eurostat (2012a), *Implementation of the ESS Joint Strategy: “the plug and play concept as an architectural principle”*. Supporting paper from Eurostat (prepared by J.-M. Museux, N. Hilbert, and G. Pongas) to the Meeting on the Management of Statistical Information Systems (MSIS 2012).
- Eurostat (2012b), *Quality Assurance Framework of the European Statistical System. Version 1.1*. Eurostat webpage.
- Forbes, S. and Brown, D. (2012), Conceptual thinking in national statistics offices. *Statistical Journal of the IAOS* **28**, 89–98.
- Godbout, S. (2011), *Standardization of post-collection processing in Business Surveys at Statistics Canada*. Proceedings of Statistics Canada Symposium 2011.
- Groves, R. M. (1989), *Survey Errors and Survey Costs*. Wiley, New York.
- Groves, R. M. and Lyberg, L. (2010), Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly* **74**, 849–879.
- Hofman, F. (2011), *Redesign at Statistics Netherlands*. Proceedings of Statistics Canada Symposium 2011.
- JOS (2013), Special Issue on Systems and Architectures for High-Quality Statistics Production. *Journal of Official Statistics* **29**, No. 1 (containing a set of articles and discussions).
- Laitila, T. (2012), *Quality of registers and accuracy of register statistics*. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.
- Linacre, S. J. and Trewin, D. J. (1993), Total Survey Design – Application to a Collection of the Construction Industry. *Journal of Official Statistics* **9**, 611–621.
- Lyberg, L. (2012), Survey Quality. *Survey Methodology* **38**, 107–130.
- Marella, D. (2007), Errors Depending on Costs in Sample Surveys. *Survey Research Methods* **1**, 85–96.
- Merad, S. and Brodie, P. (2011), *Standardizing UK sub-annual Business Surveys*. Proceedings of Statistics Canada Symposium 2011.
- Schouten, B., Calinescu, M. and Luiten, A. (2013), Optimizing quality of response through adaptive survey designs. *Survey Methodology* **39**, 29–58.
- Snijders, G., Haraldsen, G., Jones, J., and Willimack, D. K. (2013), *Designing and Conducting Business Surveys*. John Wiley and Sons, Inc.
- UNECE (2013), *What’s New from the High-Level Group?* Working paper from the UNECE (prepared by S. Vale) to the meeting on the Management of Statistical Information Systems (MSIS 2013).
- Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistics Neerlandica* **66**, 41–63.

Interconnections with other modules

8. Related themes described in other modules

Choosing the most relevant ones

1. General Observations – Methods and Quality
2. General Observations – Different Types of Surveys
3. General Observations – The European Statistical System
4. General Observations – GSBPM: Generic Statistical Business Process Model
5. User Needs – Specification of User Needs for Business Statistics
6. Repeated Surveys – Repeated Surveys
7. Questionnaire Design – Main Module
8. Statistical Registers and Frames – Main Module
9. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
10. Statistical Registers and Frames – Survey Frames for Business Surveys
11. Statistical Registers and Frames – The Design of Statistical Registers and Survey Frames
12. Statistical Registers and Frames – Quality of Statistical Registers and Frames
13. Sample Selection – Main Module
14. Sample Selection – Sample Co-ordination
15. Data Collection – Main Module
16. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
17. Data Collection – Design of Data Collection Part 2: Contact Strategies
18. Data Collection – Mixed Mode Data Collection
19. Data Collection – Collection and Use of Secondary Data
20. Response – Response Process
21. Response – Response Burden
22. Micro-Fusion – Data Fusion at Micro Level
23. Statistical Data Editing – Main Module
24. Imputation – Main Module
25. Weighting and Estimation – Main Module
26. Weighting and Estimation – Design of Estimation – Some Practical Issues
27. Weighting and estimation – Estimation with Administrative Data

- 28. Quality Aspects – Quality of Statistics
- 29. Quality Aspects – Revisions of Economic Official Statistics
- 30. Macro-Integration – Main Module
- 31. Statistical Disclosure Control – Main Module
- 32. Dissemination – Dissemination of Business Statistics
- 33. Evaluation – Evaluation of Business Statistics

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1. Phase 1. Specify Needs
- 2. Phase 2. Design
- 3. Phase 3. Build
- 4. Phase 4. Collect
- 5. Phase 5. Process
- 6. Phase 6. Analyse
- 7. Phase 7. Disseminate
- 8. Phase 9. Evaluate

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

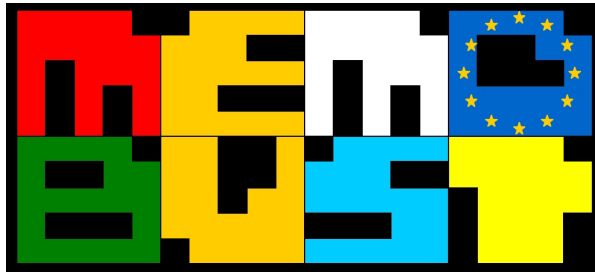
Overall Design-T-Overall Design

15. Version history

Version	Date	Description of changes	Author	Institute
0.0.1	27-03-2012	first overview	Eva Elvers	Statistics Sweden
0.0.2	23-04-2012	adj. after Rome meeting	Eva Elvers	Statistics Sweden
0.0.5	20-06-2012	adj. after reviews	Eva Elvers	Statistics Sweden
0.0.6	11-03-2013	template, glossary, add's	Eva Elvers	Statistics Sweden
0.1	15-05-2013	some expansions	Eva Elvers	Statistics Sweden
0.1.1	29-06-2013	glossary	Eva Elvers	Statistics Sweden
0.1.2	23-08-2013	adjustments	Eva Elvers	Statistics Sweden
0.1.5	16-11-2013	some clarifications, add's	Eva Elvers	Statistics Sweden
0.2	17-12-2013	EB review, addition	Eva Elvers	Statistics Sweden
0.2.1	18-12-2013	preliminary release		
0.2.2	14-01-2014	some expansions, L-m'g	Eva Elvers	Statistics Sweden
0.3	10-02-2014	other modules, glossary	Eva Elvers	Statistics Sweden
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:25



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Repeated Surveys

Contents

General section.....	3
1. Summary	3
2. General description.....	4
2.1 Introduction	4
2.2 Frames and sampling	5
2.3 Data collection and data processing	6
2.4 Estimation.....	8
2.5 Time series issues	9
2.6 Tests, experiments, and evaluation.....	11
3. Design issues	13
3.1 Introductory remarks	13
3.2 A new repeated survey – some issues	13
3.3 Frame, sampling, and estimation.....	14
3.4 Improvements of repeated surveys	16
3.5 Redesign and other considerable changes of a repeated survey	16
4. Available software tools.....	17
5. Decision tree of methods	17
6. Glossary.....	17
7. References	17
Interconnections with other modules.....	19
Administrative section.....	21

General section

1. Summary

A repeated survey is a survey carried out more than once, mostly with regular frequency, for example monthly, quarterly, or annually. Most surveys in a statistical office are repeated. Samples in a repeated survey may be independent over time, or the sample design may deliberately involve a unit at several occasions. The sample design should balance accuracy requests, which often imply considerable overlap between samples over time, and response burden. A sample design, where a business is selected each time during a period and then is not selected for a time period, has advantages for both accuracy and the respondent. Panel surveys and longitudinal surveys are particular cases of repeated surveys. There are other arrangements, for instance based on permanent random numbers. A repeated survey may use administrative data, either only or in combination with directly collected data.

Measures of change are normally an important part of the statistical output of a repeated survey, for example indices and in many cases also time series. Seasonal adjustment may be used for short-term statistics to make comparisons easier for the users. Usually there are time-related requests on the output, such as comparability over time and high accuracy in estimates of change. The changes over time are due both to population changes and changes in values of variables. The requests have implications for the survey design. Differences in definitions and methods between two points in time mostly have a negative effect on the comparability between the two sets of statistics. Considering comparability over time only, such differences should be avoided. It is in the nature of a repeated survey to use the same definitions, methods etc.

A break in the time series may become unavoidable for external or internal reasons, and it can be justified when the advantages outweigh the disadvantages. It is important to measure its size, if possible, and to inform the users in advance about the introduced changes and the break. When statistics from repeated surveys are published it is often the case that the statistics for one or more of the earlier time periods are revised. It is recommended to have a revision policy, preferably aligned with other statistics, both nationally and internationally.

The repetitive character of the survey gives possibilities to improve the statistical production process and the quality of the output by utilising both previously collected data and process data (paradata). These possibilities should be taken into account before the first production round and incorporated in the design to ensure that appropriate data, paradata, and metadata are collected and saved for future use. An imbedded experiment is an example of a method to study effects of a suggested change in advance and possibly avoid time series breaks or at least reduce the effects.

There are three major reasons for a separate description in the handbook of repeated surveys: the possibilities to make improvements over time, the possibilities to utilise previous data if a unit is included repeatedly in the survey, and issues related to time series breaks. This topic provides an overview of the specifics of repeated surveys. Most methods are already described in other parts of the handbook, for example methods for sampling and estimation. References are given to relevant modules and also to the general literature, mainly on specific issues. There are few books dedicated to repeated business surveys, perhaps a bit surprisingly, since business surveys often are regularly repeated surveys. The overview edited by Cox et al. (1995) has good coverage. Snijkers, Haraldsen, Jones, and Willimack (2013) describe how to design and conduct business surveys; a recent book.

2. General description

2.1 Introduction

Business statistics are largely based on regularly repeated surveys. One reason is that measures of change are important, often more important than measures of level. Comparability over time, which then is an essential property, depends highly on concepts being the same. Stability of methodology is also essential – methodology needs to be unchanged if it has systematic influence on the output. These two facts are important to consider and handle when working with a repeated survey. It may be desirable or necessary to change concepts or methodology in order to improve quality (mostly the contents or the accuracy) or to reduce costs. Some resulting time series issues are mentioned here, and they are further discussed below in Section 2.5. Effects of the considered changes on the statistics should be measured, if possible, to avoid or reduce a break in the time series. It may be possible to adjust or extrapolate a time series forwards or backwards, depending on the knowledge about the break and an assessed likely behaviour over time. It is also important to inform the users. See for instance OECD (2007), a handbook on presentation.

A repeated survey can have samples that are independent over time or samples that are deliberately overlapping. Here focus is on the latter type. There are several advantages of including the same statistical unit in several rounds of the survey. Accuracy in measures of change is usually improved – but the sampling design needs to balance between accuracy gains and response burden. When a statistical unit has provided survey data, these data can be used in later rounds of the repeated survey, for example in editing or by providing them to the respondent as a support in a later round. In the latter case it sometimes happens that a statistical unit corrects earlier data, for example when seeing the previous data or realising for other reasons that it has provided erroneous data to the survey. This may occur for instance when there is a new respondent in the business, looking with new eyes. These new data (correction on micro level) can be used to revise the earlier published statistics (macrodata) later on; or even to correct the statistics outside the regular publishing scheme, if needed. Units that are outliers – values are correct but extreme and influential in regular estimation – should be studied to understand reasons; such knowledge may be used to improve size measures and estimation procedures in forthcoming rounds of the survey. It is typical for business statistics that the population changes quickly due to births, deaths, splits, and other re-organisations. It is important to have a frame that is regularly updated and gives access to the survey population, directly or in several steps. Some care is needed, though, when using previous information from the sample in the sampling frame, in order not to introduce bias; see further Section 3.3 below.

Experiences and more formal evaluations can be used to draw conclusions from earlier to later production rounds, thus improving the cost-efficiency of the survey. If the possible consequences of a methodological change are not known it may be useful and relatively cheap to use an embedded experiment, where the sample is divided into two or more groups, to compare the different methodologies (similar to study and control groups in other experiments).

A repeated survey may be based on direct data collection or on other types of data sources, such as administrative data, or possibly a combination. Some parts of this module are relevant only for surveys with direct data collection; surveys which often are sample surveys. This is particularly the case for sample co-ordination, providing previous values to the respondents, estimation with weights, outliers, and variance estimation. Other parts are relevant for all types of surveys. This is the case for updates

of frames and other information on populations, comparability over time, time series breaks, evaluation, and successive improvements. Use of administrative data means work with administrative and statistical units, populations, variable definitions, possibly models to make “transformations”, estimation models etc. See, for instance, the handbook modules “Data Collection – Collection and Use of Secondary Data” and “Overall Design – Overall Design”.

Most of the methods that are used for repeated surveys are described fully or partly in other topics of the handbook. The sub-sections 2.2–2.6 below have the headings Frames and sampling, Data collection and data processing, Estimation, Time series issues, and finally Tests, experiments, and evaluation. Even if the phases of the Generic Statistical Business Process Model (GSBPM) are not mentioned as such, most of them are discussed here: Specify needs, Design, Build, Collect, Process, Analyse, Disseminate, and Evaluate.

2.2 Frames and sampling

In repeated surveys considerable attention should be given to frame construction, sample design, and estimation. Some general remarks are given in Sections 2.2 and 2.4, and Section 3 describes design issues in some more detail. With administrative data the main issue in this context is to be careful with updates and reference times for population information.

When estimating change over time a large overlap of samples between occasions is desirable. Large businesses will be selected with probability 1. Small businesses may, for instance, be in the sample for one, a few, or some years. It is easier both for the statistical office and the respondent if there are longer periods of being inside and outside the sample, rather than frequently jumping in and out. This is part of the sampling design, as described below in Section 3.3. Often some sample co-ordination is used to increase or decrease the overlap of samples in comparison with independent samples. Simply expressed, negative sample co-ordination between two surveys means that samples for these have as few businesses in common as possible or reasonable. Conversely positive sample co-ordination between two surveys means that these surveys have as many businesses in common as possible or reasonable. Similarly, positive co-ordination over time for a survey means a high overlap between the samples of the two periods. As stated previously, accuracy and response burden are the two main reasons for such positive or negative co-ordination of samples.

Since business populations usually change rapidly, frames and samples need to be updated with some frequency, rather than retaining the same sample for a long period of time. The frequencies of updating need to be decided. The frame and the sample could be updated at the same times. Alternatively, the frame could be updated more often, with just small changes of the sample. For instance a sample corresponding to new parts of the population could be drawn, as a complement. There is a balance between improved information and the work with new frames and samples (also for the affected respondents). See further Section 3.3 below. Another way of handling coverage problems due to using ‘old’ samples is in the estimation; see Section 2.4 below and the overview in the handbook module “Weighting and Estimation – Main Module”.

Maintaining the business register is essential to capture births, deaths, and other changes of the units before sampling to avoid problems in estimation. Updating registers and frames with data from samples in repeated surveys is, however, not straightforward. In repeated surveys in which there is a controlled overlap of samples between occasions, survey feedback can lead to bias in estimates, see

Section 3.2. The problem is discussed also in the handbook module “Sample Selection – Sample Coordination”. See Cox et al. (1995), too.

2.3 Data collection and data processing

One of the characteristics of a repeated survey is that many statistical units are included several times, often every month or quarter during a period of one or several years. Values for one or several previous periods are available when the unit has responded, and that information can be utilised in several ways. Such data, with measurements of the same variable at several times, are sometimes called longitudinal data.

Some particular comments for surveys with administrative data follow. There are often large amounts of data, and re-contacts are in general not possible. See the specific editing module “Statistical Data Editing – Editing Administrative Data”. The comments below on possible corrections are valid but probably not very frequent.

2.3.1 Providing previous data

A provided value may be printed or filled in when sending out the next questionnaire (paper or other mode). This gives support to the respondent, who fills in the questionnaire, and it reduces the response burden. Also, if a mistake was made, the respondent has an opportunity to correct the erroneous figure. On the other hand – if there was misinterpretation of which information to provide – showing these data may conserve this misinterpretation.

The layout of a printed or electronic questionnaire can take the form of two parallel columns beside the question. One column refers to a previous period and is pre-filled with the values from that period. The other column refers to the current period, where the information is to be filled in by the respondent. The respondent may be asked to correct the previous value in case it is in error. Such an error may be due to an earlier mistake of the respondent when providing the information, an update of that information, or a mistake when registering the information (for instance scanning).

The situation is similar when interviewing the respondent instead of using a questionnaire. When previous information is utilised in the questions of the interview, it is called “dependent interviewing”. The previous information can be used in different ways, such as using a value without re-asking, probing in case of an unexpected change, verifying in case of unlikely combinations, and checking the time of a rare event. The advantages and disadvantages of dependent interviewing are similar to those of making previous information visible in questionnaires.

Holmberg (2004) carried out an experimental study, one of the few in business statistics, for a survey with pre-printed self-administered questionnaires for an establishment population. He lists several reasons for the use of such questionnaires: respondent support (reducing the response burden, questionnaire guidance, memory support, anchoring, and feedback purposes), improved efficiency in the data collection, and reduction of measurement errors and improved data quality. He also gives reasons against pre-printing: risk of bias due to underreporting of changes and conservation of errors, loss of confidence and goodwill (if the pre-printed data are of poor quality or if they are not recognised by the respondent), and disclosure risk. He used three treatments in the experiment: no pre-printing, pre-printing for one period, and pre-printing for two time periods. The survey that was studied was complex and asked for data during fourteen months. He considered the effects of pre-printing on

various aspects: response variability, presence and size of outliers, effects on other months, and experiences in general. In this study there were advantages: fewer and smaller problems with outliers and less spurious variation. Care has to be taken before generalising from one survey to another, though.

Overall, from the relatively few reported studies, it seems as if the advantages for preprinting are stronger than the disadvantages. When planning or using previous data, different factors behind the advantages and disadvantages should be noted and taken care of, as far as possible. Morrison (2009) provides some guidelines and further references on questionnaire design. Snijkers et al. (2013) provide broad information on questionnaire design.

2.3.2 Corrections of previous microdata – and macrodata

Corrections from the respondent of previous data mean that the statistical office can improve the output of earlier periods. For statistics with regular revisions this is a natural action in the next publication round. If the statistics are already final when additional corrections are made on the micro level, the effects on the estimates have to be considered. It may happen that the effect is great enough to motivate an unplanned correction of the statistics published. Specifically, two aspects are the size of the effect(s) and the status of the statistics, preliminary or final.

The size of the effect should be considered together with several factors, such as the accuracy and the aggregation level of the statistics. This leads to the following cases:

- If the size is considerable, a correction may be necessary.
- If the statistics are preliminary and the next planned revision date is close, it may be better to wait until this regular revision, even if the correction size is not very small.
- If the size is small and the statistics are preliminary, the correction and its effects simply go into the next revision.
- If the size is small in comparison with the accuracy of the statistics and the statistics are final, it is probably better to refrain from correction.

2.3.3 Using previous values in editing and imputation

Previous values are useful for both editing and imputation. This – longitudinal data – is an inherent strength of repeated surveys with some overlap between survey rounds in comparison with surveys that are made just once.

In editing a comparison with a previous period may show a value to be spurious due to an unrealistic change – and thus lead to the detection of an error that had otherwise gone unnoticed. It may also happen that the previous value was wrong, as discussed above. A re-contact may be motivated in case of a surprising change. Perhaps, there has been a mistake or, as a possible alternative, a change in organisation of the enterprise. See further the handbook module “Statistical Data Editing – Editing for Longitudinal Data”.

A further use of previous values is in imputation. If the unit is late in responding this period it may be reasonable to utilise the value of one or more previous periods and bring them forward in time, with regard taken also to possible seasonality – unless there are signals about the unit undergoing some

change. It may be reasonable to assume that a group of similar units (for example in a stratum) have similar sizes of change since that period. This method, which uses previous information of the unit, may be better than using only information from a group of similar units in the current period. See further the handbook module “Imputation – Imputation for Longitudinal Data”.

2.4 *Estimation*

This section is mainly relevant for sample surveys with direct data collection. When administrative data are used, estimation issues will involve coverage deficiencies and possibly late data, especially for small enterprises. See for instance the handbook module “Weighting and Estimation – Estimation with Administrative data”, which has a description and provides references to the ESSnet project AdminData with many relevant deliverables.

Data from previous runs of the survey are available in repeated surveys, at least to some extent. This means that further and more advanced estimation methods are available. For instance, there are different imputation methods based on previous values, as mentioned above.

The changes over time for a variable, like turnover, have two basic causes: the target population changes due to births, deaths, splits etc. and the “stable” units change in size resulting in larger or smaller variable values. The estimation has to take into account these causes based on the sample design, the frame, sample information, and possibly auxiliary information. The frequency of new register and frame information has an influence on the estimation procedure, and so has the frequency of new or partly renewed samples.

In short-term statistics, using a frame with information that is already somewhat old due to reporting delays and a sample for a longer period leads to coverage problems and problems with businesses that merge or split. These problems are, of course, present in all business surveys, but in short-term surveys with sample selection perhaps once or a few times a year they grow and become more serious at the end of the period that the sample is used. Mergers and splits mean that the units are not the same as they were at the sampling occasion. Sometimes data on the sampled units can still be collected, in case the data providers can report values for the old units. If that is not possible the problem with unit changes has to be handled in the estimation. There are several principal and practical issues to consider in the estimation (not only for repeated surveys), see the handbook module “Weighting and Estimation – Main Module”.

Coverage problems may be handled if there is an updated register or frame, which can be used to “adjust” the estimates to correspond to the updated information. One or more auxiliary variables, which are known on both sample and population levels, are used for such a calibration; see the handbook methodological module “Weighting and Estimation – Calibration”. There may be some practical issues with merging, comparing and handling large businesses.

Problems with outliers are common in business surveys of economic variables, as indicated above. Usually there is stratification by some size measure or a sampling design using a probability proportional to size. However, since the size measure is not always up to date or the size measure is not perfectly correlated with the study variable(s), it is difficult to order the businesses by size. For some study variables this is more difficult than for others, for instance an investment variable with a skewed distribution. There is always a risk of observations with a high influence on the estimates, because the size measure in the sampling design is not up-to-date or the variable has a highly skewed

distribution. Some type of outlier treatment must be in place in the estimation and included already in the design. See Section 3.2 and the handbook module “Weighting and Estimation – Outlier Treatment (Robust Estimation)”.

When samples are co-ordinated over time variance estimation for measures of change is not as straightforward as variance estimation for parameter estimates based on one sample. Some methods are in place; see for instance Nordberg (2000), who suggests a variance estimator when samples are co-ordinated by using permanent random numbers, and Knottnerus and Van Delden (2012) for a more general setting. This is discussed also in the handbook module “Sample Selection – Sample Co-ordination”.

The population parameters to be estimated may be indices, both price indices and volume indices of different types. The choices should be made from the start in agreement with important users.

2.5 *Time series issues*

An important reason for repeated surveys is to measure population changes over time. Such interests from users have implications for both production and dissemination. The quality component comparability over time is important, but it has, of course, to be balanced with other quality components. A change may occur, for instance by introducing new methodology for accuracy reasons, even if such a change implies a break in the time series. Alternatively, a break may be caused by external changes, for instance new tax rules. Use of new concepts in the survey may be justified because they describe the current situation better than the old concepts. Hence contents and relevance may motivate a disturbance to comparability over time. This section is relevant for all types of surveys.

2.5.1 *Comparability over time*

There is always inaccuracy in statistics: random variation and systematic variation. The random variation makes comparisons less “sharp” or conclusive, but the comparisons are still meaningful with just random variation. However, if there is some systematic deviation, this normally disturbs or destroys the comparability. Examples where such possibly systematic influence is introduced are:

- A different way of updating the business register, for example a new source or a different time schedule. This influences for instance the frame coverage and the accuracy in classifications.
- A new data collection mode influences possibly, but not necessarily, the data that the respondent provides.
- Changes in the editing procedure influence possibly, but not necessarily, the output in a systematic way.

This is simply expressed as follows. With “everything unchanged” in production, response processes, and the society context, comparability over time is not affected. Otherwise possible systematic influence has to be investigated.

Some systematic errors may have different effects on estimators of levels and estimators of change measures. For example coverage deficiencies mainly affect the level of estimation. However, care has to be taken also for estimators of change measures with regard to over-coverage and under-coverage. Such coverage deficiencies have different sizes over time for instance between different parts of a

business cycle (up-and-down movements in economic activity). Hence, a simple assumption in the estimation procedure of over-coverage and over-coverage being equal may work when the economy is stable but be quite misleading in times of change.

Comparability over time means that a presentation of the time series in a graph or in a table is meaningful. There may still be calendar and seasonal effects. Removing such effects makes the study over time easier. Sometimes just simple comparisons with the corresponding period in the previous year are made.

Comparability defects, if any, should be clearly stated and explained in all presentation modes. There are two major situations where breaks occur. The causes indicate at least partly which methods to use in order to handle the problems that occur. Such different methods are discussed in the next two sections.

2.5.2 Methods to overcome breaks, for instance caused by redesign

There are typically two time series, the old and the new one. They cover different time periods, possibly with a “double” period. The two series show “the same thing”, but there is a difference, which for instance is caused by using another method. Hence, there is a “jump”.

One possibility to quantify the effect of a change without a full implementation is to make an experiment, as described in Section 2.6.2 below. At best, use of experiments can reduce, or even avoid, time series breaks. If the change is introduced without a controlled experiment, there should be a double period in order to have some possibility to estimate the effects of the change.

A further possible situation is to have two separate time series that describe the same or nearly identical phenomena but without a clear connection between the series. This may for instance be the case when the National Accounts replace one source (the old one) with another source (the new one). There may be one or a few time periods where both sources exist. There is often a “gap” between the two time series. De la Fuente (2009) describes a set of methods that is more flexible than a simple “vertical movement” of the old series. An error term is introduced to describe the difference between the two time series, and some appropriate assumption is made about this error; this assumption may not be obvious. Then the adjustment is derived.

Van den Brakel and Roels (2010) describe an approach where state-space models and intervention analysis are used to estimate the discontinuities, typically caused by survey redesign. Theory and illustrations are provided.

When several related time series are involved, they have to be considered together so that consistency is preserved, for example so that parts of a sum add up to the total.

2.5.3 Revision of a classification

Much methodological and practical work has been devoted to the situation where a classification is revised, for instance a new version of NACE. The revised classification is introduced because it is now a better description of society than the old classification. The users like long time series and often ask for “back-casting”. This may be reasonable for a high aggregation level and for a moderate time period. Care has to be taken in computations, presentations, and interpretations of the back-casted series. Perhaps some industries in the revised classification did not exist ten years back in time.

There is a user need to extend the time series, especially backwards, in spite of the break. The relationship between the old and the new versions of the classification is complex in many cases. The Business Register is normally double-coded, that is coded according to both classification versions, usually for a brief period of time.

There are two major approaches to extend time series: the micro and the macro approach. They are briefly described below.

The micro approach is based on business units. The new coding is extended backwards for units in the survey, and estimates are made for these new domains of estimation. These estimates will be somewhat inaccurate. There is uncertainty and difficulty with coding backwards. Some units no longer exist when the double-coding is made, and they have to be coded with limited knowledge. Moreover, the survey was designed for the old classification. There may be relatively few units in some domains of estimation.

The macro approach builds on relationships between the two versions of classification. The double-coded business register is an important source of knowledge about the relationship in that period. Choosing some appropriate level of detail, a cross-classification can be made. Imagine a matrix showing the industries in the old system row-wise and the industries in the new system column-wise. Each cell shows some quantity, like the number of employees or the turnover. Each row will show the “flow” from the old version to the new version. Each column will similarly show, for each new industry on that level, which old industries contributed and to what extent.

In the macro approach such a matrix is used for “conversion” of the time series in the old classification version to time series in the new classification version. The computations are fairly easy as such for many surveys. The difficulties lie in choosing appropriate information for the conversion matrix and finding appropriate levels of detail.

There are many papers, both from national statistical offices and from Eurostat as well as other international organisations. Only two references are given here.

Brunauer and Haitzmann (2010) describe a case at Statistics Austria for the Structural Business Statistics. They compare the micro and the macro approaches. The former may seem more accurate at a first glance, but it is difficult to code statistical units backwards in time and to code units no longer existing. It is also resource-consuming.

Van den Brakel (2010) describes several aspects, especially sampling and estimation techniques with regard to the two classifications and back-casting procedures. Both the micro and the macro approaches are included. Conversion factors are discussed in some detail. Time dependency and indices are included. This is a methodological paper.

2.6 Tests, experiments, and evaluation

Embedded experiments and pilot studies mainly refer to surveys with direct data collection. Studies and evaluations are relevant for all types of surveys.

2.6.1 Tests and experiments

Testing is mostly an investment that prevents future work with corrections and other problems. Some types of mistakes mean that one or more processes have to be run again. In some cases it is not

possible to re-run the process, though, in spite of the mistake. This is not specific for repeated surveys, but a general observation.

In a repeated survey much is the same from round to round. A new reference period may require new settings in the production system, by parameters (preferably) or manual settings. Changes may be needed for data set names etc. Every change means a potential error. Changes should be tested early.

A main goal of a repeated survey is to measure changes in population parameters, so comparability over time is mostly important. Changes in production may be motivated for a number of reasons, for example indications about desirable or possible improvements from earlier production rounds or other internal changes in the production environment. Cost-effectiveness is always essential, and small improvements may mean considerable savings in the long run for a survey that is conducted many times.

However, consequences of changes should be foreseen before embarking on them. A pilot study is always a possibility to make investigations on a small scale. It may be a good and cheap way to discover weak points and unexpected implications. Repeated surveys provide a setting that enables more than separate pilot studies, because experiments can be made within the survey setting: so-called embedded experiments.

In an embedded experiment the sample is randomly divided into two or more subsamples according to an experimental design. There is normally a control group and one or more treatment groups (subsamples). The treatment(s) may be, for example, new advance letters, new data collection modes, or new contact strategies. Care has to be taken in planning in order to avoid confounding with other influencing factors and also to make sure that the experiment is feasible in practice. Moreover, the hypotheses should be formulated in advance, and the powers of the tests should be estimated at the same time to understand which differences can be considered as significant. Otherwise the experiment may turn out to be an inconclusive disappointment.

A methodological description of how to make an embedded experiment is given by Van den Brakel and Renssen (2005), where also further references can be found. Another description with both an experiment and a time series perspective is given by Van den Brakel, Smith, and Compton (2008). They describe quality procedures for survey transitions; see also Section 2.5 above. A practical study dealing with response burden is presented by Hedlin, Lindkvist, Bäckström, and Erikson (2008). A practical study about pre-printing effects is presented by Holmberg (2004); see also Section 2.3 above.

2.6.2 Evaluation

An evaluation should always be made after a production round to learn for the future, considering both this survey (if repeated, as described here) and other related surveys. The ambition level should vary with regard to survey frequency and findings. In short-term statistics it is natural to catch the most urgent matters after each single round. All findings can be summarised in more detail after a longer period and then acted upon. The regular annual planning may be a suitable point in time to consider modifications and redesign.

An example of a qualitative type of evaluation is a brief summary of staff experiences made during the production. Debriefing with staff may be a good way both to collect findings and to get suggestions

for improvements. Staff working on editing is such a group where difficulties regarding data collection can be detected and summarised. These include observed difficulties, for instance in filled-in questionnaires and in re-contacts with respondents about spurious values. Suggestions for better wordings in questions and instructions may be obtained here, and also suggestions to improve the editing procedure.

Other evaluations are more quantitative. It is recommended to collect process data (paradata) as a basis for evaluations. Two important types are measures/indicators of quality and data about costs. A quality indicator may be useful for process quality or product quality or both. The goal of an evaluation is both quality assessment for this round and improvements in future rounds. The quality assessment is, of course, used in quality reporting. Measures of response rate over time provide an example of a quality indicator, where the conclusions drawn about the output quality have to be careful and restricted. Similarly, there are rates in editing that are indicators of the process, for instance how many “suspicious” cases turn out to be influential errors. This is not a good indicator of output quality, though.

Analysis of the response rate together with dates of reminders is one way to improve the strategy for contacting respondents: how often, in what way etc. There are several ways to improve the data collection. See Section 3 below for a brief description and several handbook modules: the module “Data Collection – Main Module”, two modules on design of data collection and one on mixed mode.

There is also a handbook module devoted to the topic evaluation: “Evaluation – Evaluation of Business Statistics”.

3. Design issues

3.1 Introductory remarks

There are two basic situations: a new repeated survey and an ongoing repeated survey. The latter situation is typical for statistical offices and their business statistics. The former situation is rarer, but sometimes a new repeated survey is launched. Several design issues are mentioned already in Section 2 in their respective contexts. Here in Section 3 the focus is on design, and especially frame, sampling, and estimation are discussed concentrating on repeated surveys. Most of the design issues are relevant for all types of surveys. Responsive design, sampling, and estimation from a sample are exceptions, useful only for surveys with direct data collection.

3.2 A new repeated survey – some issues

Even if the design of a repeated survey is in many respects similar to the design of a one-time survey, there are some additional issues and possibilities to consider. Many of these should be considered before the first round to achieve the potential gains from the start and to eliminate risks of undoing some work.

It is, as always, important to consider the user needs early. One of the special issues here is to discuss and verbalise the priorities among the accuracies of different estimators. One such example is how to balance between the accuracy of estimation of levels and the accuracy of estimation of change over time.

Coherence and comparability also need to be considered. These considerations may have implications for the choice of statistical unit(s), population delineation, variable definitions, reference times/periods, frames etc. They may also influence a possible co-ordination with other surveys. This is discussed in the next section.

The fact that the survey is made repeatedly means that previous data can be used in data collection, editing, and imputation as discussed above in Section 2. This should preferably be designed from start, since the production system needs to include these previous data and additional procedures. Editing and imputation procedures are, of course, needed also for units without previous data. Hence, several different procedures need to be included with priority rules.

Possibilities for continuous improvements should be built into the system, preferably from the beginning. This means both collecting appropriate paradata and using them. This is considered in Section 3.4 below and in the handbook module “Overall Design – Overall Design”.

3.3 Frame, sampling, and estimation

There are some essential choices in the design. The description is simplified here, mostly because there is information in other topics and modules. Some essential decisions are included, which are typical for repeated surveys. They are, of course, not to be made one by one, but together.

One important decision is about the frequency of frame updates and of new (or renewed) samples. In annual surveys the procedure is usually to update the frame and select a sample once a year. In short term statistics the value of new information and the additional work with new samples have to be balanced. Since business populations change rapidly using a sample for a year (or many months) leads to coverage problems especially at the end of the period. Births, deaths, splits, and mergers are frequent and lead to population changes. Moreover, changes in size measure and economic activity, typically used for stratification, are common, too. If the business register and other sources of data, which are used for frame construction and sample selection, are updated more frequently than annually, a new frame can also be constructed more frequently than annually. See “Statistical Registers and Frames – Main Module” and its directions to further modules in that topic.

Selecting new samples often leads to increased workload in the data collection: contacting new enterprises etc. The first time a new sample is used, the non-response is sometimes larger, since some of the units are included in the survey for the first time. There may also be more problems with measurement errors, since new data providers are included more often. Furthermore, selecting samples more often also leads to increased burden for enterprises that are included and excluded from the survey more often. It may also be possible to continue with the previous sample but utilise the new frame information to handle movements into and out of target population. The sample may then be “complemented” with regard to units that belong to the target population according to the new but not the previous information. Conversely, parts that no longer belong to the target population are removed. Both theoretical and practical issues have to be considered. The sampling procedure and the estimation method with design weights must be known. Some further comments are given below.

Another important design decision is if and how to co-ordinate samples, both over time and with other surveys. The idea of co-ordination between samples over time is to obtain a ‘large’ overlap part of the samples between survey rounds. Large overlap between samples over time normally ensures smaller variance of estimated change. Parts of the sample should, however, be replaced to capture changes in

the population, for example to include new units, and to spread the response burden among the businesses. There may, for instance, be a system with permanent random numbers (PRNs), where:

- Positive co-ordination is used over time. This increases the accuracy of estimators of change. The respondent may find it relatively good to be “in” for a period and then “out” for period.
- Positive co-ordination is used between some related surveys. This makes comparisons easier on the micro level (co-editing) and on the macro level.
- Negative co-ordination is used between many non-related surveys, mainly to reduce response burden.

See the handbook module “Sample Selection – Sample Co-ordination” and several related method modules.

Stratification is usually done by domains, such as kind of activity, and size. The largest size classes, one or more, in each domain are usually totally enumerated while samples are selected from strata with small or medium-sized enterprises. In repeated surveys the choice of size stratification variable is maybe not the seemingly optimal one. A variable that is more stable over time can be preferable to one that leads to a more efficient design but which changes more rapidly. Using size in the design is, of course, important to avoid, as far as possible, problems with outliers.

A further important decision is about principles and handling of outliers: Some type of outlier treatment must be in place in the estimation and included already in the design. There are several methods to handle outliers in the estimation. Many methods modify the weight or the value. Often this means that the variance of the estimator is considerably reduced, but that a bias is introduced; together this normally means that the mean squared error is reduced. With a repeated survey data from previous runs of the survey can be used for identifying outliers, for outlier treatment in the estimation, and possibly for a better sampling design.

A further important estimation decision, which is related to updates of the frame and the sample, is about handling changes of units: Principles are needed for taking merging, splitting, deaths, changed industry etc. into the estimation. Some of these changes will be known in later versions of the register and frame, possibly at the time of estimation but possibly not until later. Hence, the changes may be known for the sample only. It is possible to work with the weights or the values of the statistical units. As mentioned above, there is also the decision how often to update the sample and in which way (fully or partly). See the handbook module “Weighting and Estimation – Main Module”.

Another decision for estimation concerns the possible use of auxiliary information to improve accuracy, for instance with regard to non-response and coverage deficiencies. If the first production round (for preliminary statistics) is very early, specific estimation procedures may be needed, more model-based. If there is cut-off sampling, a model-based estimation is needed. See the handbook module “Weighting and Estimation – Main Module”, where design is discussed.

Finally, in repeated surveys in which there is a controlled overlap of samples between occasions, survey feedback into the business register and the survey frames has to be considered. There is a dependency between samples. For instance, consider the case where information about deaths is quicker from the survey than from the regular sources. If such information is fed into the register and the new frame, the next sample that is drawn will have more updated information than the frame and population as a whole, due to the positive co-ordination with the previous and updating sample. With

standard estimation procedures, survey feedback can lead to bias in estimators. The problem with survey feedback is further discussed in the handbook module “Sample Selection – Sample Coordination”. See also Cox et al. (1995). It is still important to use the updated information in communication with the respondents and to design the handling of the updated information in all surveys in a consistent way.

3.4 Improvements of repeated surveys

Improvements are based on evaluation, mostly of previous production rounds. Some typical warning signals are a high item non-response rate, a lot of editing work for some variables, and lower measures of accuracy than expected. Causes should be searched, for instance some small mistake in the questionnaire or a computer program. The estimation procedure should perhaps be improved, for instance through more auxiliary information. The working procedures should perhaps be adjusted by allocating resources differently. There are many possibilities for improvement. Warnings of quality deficiencies should be traced backwards. Such studies to learn and improve are relevant for editing, for instance, where an editing method, including parameters, can be studied with respect to numbers of suspicious cases, detected errors etc.

There is also a possibility of evaluation and actions within a production round. The design may, for instance, include points in time where the situation is considered, typically with respect to non-response. A method to adjust the data collection procedures (for instance the contact strategy based on achieved response rates) based on the data so far can be included in the design. Such a design is often called a responsive or adaptive design and should be planned in advance. See the handbook modules “Data Collection – Design of Data Collection Part 2: Contact Strategies” and “Overall Design – Overall Design”.

The second aspect is the inclusion of measures or indicators of resources used, both in paradata and analyses for improvement. The aim is to try to increase the efficiency not only for single methods but also for the allocation of resources used in the production. See also the handbook module “Overall Design – Overall Design”.

3.5 Redesign and other considerable changes of a repeated survey

There may be internal reasons to make considerable changes or to redesign a survey, for example due to integration with other surveys or new data collection possibilities. Then evaluations and possibly other information from the survey hitherto should be combined with user needs. Contacts with stakeholders and users about possible changes and about current needs and priorities may be wise. Embedded experiments may be valuable, as described above in Section 2.6, as a start of a redesign, to study possible, desirable, and non-desirable effects.

There are a few different situations with substantial changes. Time series breaks provide an example, for instance when there is a new classification or some other unavoidable external change. Switching from direct data collection to an administrative data source, partially or fully, may be such a substantial change, where output quality components have to be considered carefully. It is important to be pro-active as described above in Section 2.5.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S. (eds.) (1995), *Business Survey Methods*. John Wiley and Sons, New York.

Brunauer, M. and Haitzmann, M. (2010), *Backcasting Methods at Statistics Austria*. (Comparison of the Micro and Macro Approach on basis of the Structural Business Statistics 2005 to 2007. Results within the scope of NACE Revision 2.) Paper presented at the conference Q2010, Statistics Finland.

De la Fuente, A. (2009), *A mixed splicing procedure for economic time series*. Barcelona Economics Working Paper Series, Working Paper No. 415.

Hedlin, D., Lindkvist, H., Bäckström, H., and Erikson, J. (2008), An Experiment on Perceived Survey Response Burden among Businesses. *Journal of Official Statistics* **24**, 301–318.

Holmberg, A. (2004), Pre-printing Effects in Official Statistics: An Experimental Study. *Journal of Official Statistics* **20**, 341–355.

Knottnerus, P. and Van Delden, A. (2012), On variances of changes estimated from rotating panels and dynamic strata. *Survey Methodology* **38**, 43–52.

Morrison, R. L. (2009), Writing and Revising Questionnaire Design Guidelines. Proceeding of Statistics Canada Symposium 2008. Data Collection: Challenges, Achievements and New Directions.

Nordberg, L. (2000), On Variance Estimation for Measures of Change When Samples are Coordinated by the Use of Permanent Random Numbers. *Journal of Official Statistics* **16**, 363–378.

OECD (2007), Data and Metadata Reporting and Presentation Handbook. OECD Publishing. doi: [10.1787/9789264030336-en](https://doi.org/10.1787/9789264030336-en).

Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D. K. (2013), *Designing and Conducting Business Surveys*. John Wiley and Sons, Inc.

Van den Brakel, J. A. (2010), Sampling and estimation techniques for the implementation of new classification systems: the change-over from NACE Rev. 1.1 to NACE Rev. 2 in business surveys. *Journal for Survey Research Methods* **4**, 103–119.

Van den Brakel, J. A. and Renssen, R. H. (2005), Analysis of experiments embedded in complex sampling designs. *Survey Methodology* **31**, 23–40.

- Van den Brakel, J. A. and Roels, J. (2010), Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *Annals of Applied Statistics* **4**, 1105–1138.
- Van den Brakel, J. A., Smith, P. A., and Compton, S. (2008), Quality procedures for survey transitions, experiments, time series and discontinuities. *Journal for Survey Research Methods* **2**, 123–141.

Interconnections with other modules

8. Related themes described in other modules

1. Overall Design – Overall Design
2. Statistical Registers and Frames – Main Module
3. Sample Selection – Sample Co-ordination
4. Data Collection – Main Module
5. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
6. Data Collection – Design of Data Collection Part 2: Contact Strategies
7. Data Collection – Mixed Mode Data Collection
8. Data Collection – Collection and Use of Secondary Data
9. Statistical Data Editing – Editing Administrative Data
10. Statistical Data Editing – Editing for Longitudinal Data
11. Imputation – Imputation for Longitudinal Data
12. Weighting and Estimation – Main Module
13. Weighting and Estimation – Estimation with Administrative Data
14. Evaluation – Evaluation of Business Statistics

9. Methods explicitly referred to in this module

1. Weighting and Estimation – Calibration
2. Weighting and Estimation – Outlier Treatment

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 1. Specify Needs
2. Phase 2. Design
3. Phase 3. Build (especially regarding tests)
4. Phase 4. Collect
5. Phase 5. Process
6. Phase 6. Analyse
7. Phase 7. Disseminate

8. Phase 9. Evaluate

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

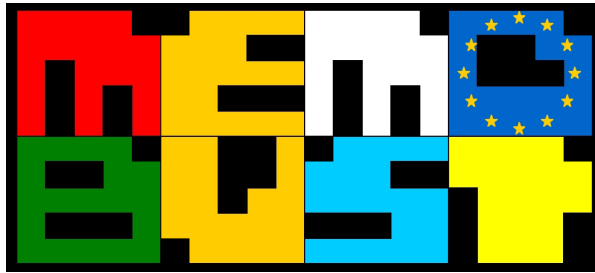
Repeated Surveys-T-Repeated Surveys

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	10-03-2012	all modules together, with sections, from NL review	Eva Elvers and Tiina Orusild	Statistics Sweden
0.1.5	27-03-2012	from GR review	Ditto	Statistics Sweden
0.2	21-06-2012	from Sander, mainly	Ditto	Statistics Sweden
0.2.5	11-03-2013	from HU, templ. glossary	Eva Elvers	Statistics Sweden
0.3	24-05-2013	updates	Eva Elvers	Statistics Sweden
0.3.2	26-06-2013	NL review and glossary	Eva Elvers	Statistics Sweden
0.3.3	23-08-2013	a few adjustments	Eva Elvers	Statistics Sweden
0.3.5	06-11-2013	some clarifications	Eva Elvers	Statistics Sweden
0.3.6	15-11-2013	references, ...	Eva Elvers	Statistics Sweden
0.3.7	22-11-2013	clarifications, NL review	Eva Elvers	Statistics Sweden
0.4	20-12-2013	after EB review; preliminary release	Eva Elvers	Statistics Sweden
0.4.1	08-01-2014	add admin data; harmon.	Eva Elvers	Statistics Sweden
0.4.2	14-01-2014	clarification for admin.	Eva Elvers	Statistics Sweden
0.5	10-02-2014	updates, glossary	Eva Elvers	Statistics Sweden
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:28



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Questionnaire Design – Main Module

Contents

General section	3
1. Summary	3
2. General description.....	3
2.1 Characteristics of the establishment population	3
2.2 The response process	4
2.3 From objectives to a draft questionnaire	5
2.4 Prototype questionnaire	7
2.5 Systematic testing	8
2.6 Questionnaire design standards	9
2.7 Quality	9
2.8 Omnibus surveys	10
2.9 Ad hoc surveys	10
3. Design issues	10
4. Available software tools	11
5. Decision tree of methods	11
6. Glossary	11
7. References	11
Interconnections with other modules.....	13
Administrative section.....	14

General section

1. Summary

In statistical surveys, the questionnaire is the pipeline which enables the flow of desired data. Although questionnaire design is part of the operational phase of a survey, it is critical in terms of survey objectives. It is difficult to compensate at later stages errors made due to an insufficient instrument (Brancato et al., 2006). What must be stressed is the iterative nature of its design and development. The relationship between information demand and the response burden has to be taken into account when introducing new forms and assessing existing ones. The thirst for more and more facts and figures must be balanced against the reporting unit's burden, quality aspects and costs.

As part of the survey process, the questionnaire preparation process, which is by its very nature an iterative process of improvement and development, must also be seen as a permanent and continuous cycle.

This module brings up the general issues connected with the questionnaire preparation for statistical data collection. The context is set for business surveys. The following modules are devoted to specific parts of the questionnaire design, making together with this module a list of modules devoted to the Questionnaire Design topic in this Handbook:

1. Electronic Questionnaire Design.
2. Editing During Data Collection.
3. Testing the Questionnaire.

2. General description

2.1 *Characteristics of the establishment population*

General characteristics of establishment surveys, which distinguish them from household surveys, and further affect questionnaire design, development and testing, include:

- The response process is complex and burdensome, since preparing the required data entails a mixture of organisational and individual tasks. Data are mostly quantitative and their acquisition requires access to business records.
- The predominant role of self-administered questionnaires and the role of the respondent. In the case of establishments, the respondent is usually a representative, who acts as a data provider in an organisational environment.
- The mandatory character of reporting as opposed to mostly voluntarily aspect of household surveys.
- Concepts and definitions are technical, complex, and often based on legal or regulatory considerations. Practices, terminology and standards used by businesses in their daily operations, the accounting standards, for example, are the context that needs to be taken into account. This calls for detailed instructions that accompany questions.
- The distributions of totals of the target population are skewed in favor of larger establishments.

- Timeliness is often given priority over quality.
- The longitudinal character of surveys and overall reluctance to changes in measurement instruments used.

2.2 *The response process*

This section only slightly touches the question of response process models in business surveys; a more deep discussion on the subject the reader can find in the theme module “Response – Response Process”.

When a respondent is asked for information by a questionnaire, a series of activities must be performed before the task is completed. A better understanding of this task can help to make the answering process less burdensome for the respondent by applying the knowledge to improve the questionnaire.

2.2.1 *Traditional response process model*

The study of what happens when an interview is conducted to elicit answers to survey questions laid the foundation for the response model consisting of separate stages. The task approach, which divides the process into stages, paved the way for techniques of detecting problems with questions and improving questionnaires. The model, originating from social surveys, was developed by Tourangeau (1984) and consists of four cognitive steps: comprehension – understanding the question, retrieval – recalling the fact from memory, judgment – assessment of its correctness, reporting – formatting the response. The model provided the basis of cognitive interviewing techniques practices.

2.2.2 *The model for business surveys*

The starting point for the response process model in social surveys is the individual, on whom the whole model is based. Models for establishment surveys, however, must adopt a different perspective to account for the fact that the response takes place in an organisation, which constitutes a ‘universe’, with all social connections within it. Nevertheless, those models are based on the four step cognitive model. To integrate organisational and individual factors, other models have been developed, such as the Hybrid Response Process Model (Sudman, Willimack, Nichols and Mesenbourg, 2000; Willimack and Nichols, 2001, 2010), or the Multidimensional Integral Business Survey Response (MIBSR) model (Bavdaž, 2010a) which distinguishes between the business/organisational and individual/personal levels. In a study by Lorenc (2007) the Socially Distributed Cognition theory was used to study the establishment response process, whereby an establishment is treated as a unit and survey response-related processes are analysed within the framework of representational states of various interactions between persons and tools used, over time. These models can be seen as complementary ways of gaining a better understanding of processes involved in establishments survey response with a view to reducing the measurement error and obtaining valid and reliable data. What follows below is the Hybrid model, which illustrates how the original concept of cognitive steps has been developed and adapted to the needs of establishment surveys:

1. Encoding in memory/Record formation.
2. Selection and identification of respondent.
3. Assessment of priorities.

4. Comprehension of the data request.
5. Retrieval of relevant information from memory and/or existing company records.
6. Judgment of the adequacy of the response.
7. Communication of the response.
8. Release of the data.

2.3 *From objectives to a draft questionnaire*

2.3.1 *Objectives and concepts*

Before the initial stage of the questionnaire construction begins, objectives of the survey must be identified. Consultations with respondents regarding the information demand, translated into concepts and the resulting outline, are an essential foundation for developing an adequate measurement instrument. Concepts, such as the target population and sample design, must be determined. The response process in establishment surveys recognises the important fact that data are contained in business records maintained for business reasons. This, in turn, is related to the issue of data availability and their matching with survey concepts. Consultation studies with data users, subject data experts and survey methodologists at the early stage help to avoid the discrepancy between the intended objective and actually collected data. Exploratory studies are the way to determine the existence of data and the complexity of the process of compiling them. Most often the concepts used are complex and require technical definitions. Cooperation between parties at the interface between methodology and subject fields can lead to a better understanding of technical terms.

2.3.2 *Variables*

Conceptual ideas must be broken down into definitions and lists of name data items, all of which leads in a straightforward way to questions. The longitudinal character of economic surveys and the goal to measure changes in time calls for the stability of variables. This strategy has two consequences. For one thing, it can ease the response burden; on the other hand, previous errors can persist in future survey periods. For this reason, changes in questionnaires should be made with caution and respondents should always be notified.

2.3.3 *Determine data collection modes*

The data collection method chosen determines the layout of the questionnaire. This issue should be resolved before starting the design process. Two major types of data collection modes might be distinguished from the administrative point of view:

- Interviewer-assisted.
- Self-administered.

From a technological perspective they can be classified into:

- Paper-based interviewing.
- Computer-assisted interviewing.

Computer assisted interviewing, so widely used nowadays as to be considered a standard, comprises:

- Interviewer-assisted – CAPI and CATI.
- Self-administered – Web interviewing, CASI.

The data collection topic is covered in several modules on “Data Collection” in this Handbook.

2.3.3.1 Suitability of the interviewer-administered mode

Personal interviewing can be considered suitable for a small sample of respondents, where concepts and questions are moderately or highly complex. Another situation where those methods can be useful are observation or panel surveys. Personal visits can be a follow-up method to mailed questionnaires. The costs of personal interviewing are the main disadvantage, as these methods are very expensive. The personal mode is more suitable for voluntary social surveys, where guidance from an interviewer is necessary to elicit more accurate responses and increase response rates.

Telephone based methods are suitable for interviews involving questions with simple concepts, where the number of items does not exceed 10. With a greater number of items, telephone based methods become problematic: questionnaires with over 40 items are regarded as unsuitable (ABS Forms manual, 2010). Similarly, the duration of the interview is an important factor when choosing a collection mode. The maximum limit for the telephone-based mode is around 20 minutes. Additional factors influencing the choice of the mode include the survey frequency and sample size. Although the front-end costs of telephone interviews are lower compared to personal visits, costs of preparing CATI instruments also keep rising, which requires a trade-off between costs and benefits. Thus, the larger the sample size and the more frequent a survey is, the more cost-effective such techniques are.

2.3.3.2 Suitability of the self-administered mode

Characteristics of the response process in establishment surveys underline the role of business data records. The requested data are stored in records, which has implications for its retrieval: selecting the proper person, the knowledge of business records. Another typical challenge is the need to merge data from different departments of the institution. The use of complex terms and definitions in establishment surveys makes the response additionally burdensome. When all these aspects are considered, it becomes clear that the most suitable mode of data collection in business surveys is the self-administered mode. If this is the case, the weight of communication rests on the questionnaire and its content. Regardless of the technique applied – be that paper-based or electronic – some common elements can be distinguished in the questionnaire: it is what methodologists call languages: verbal, numerical, symbolic and graphical. Questions express the meaning of concepts, while numbers are the characteristic trait of economic surveys. Graphics and symbolic language influence the flow and cognitive burden. Generally, the visual side is the only communicative medium as far as the paper self-administered mode is concerned. In the self-administered mode the role of instructions is of the utmost importance, whenever it is necessary to clarify the meaning of complex definitions. Methodological papers advise placing instructions close to questions rather than using separate booklets. Another recommendation is to formulate instructions as questions, that is, incorporating them into questions content. Placing instructions adjacent to or within the question can improve understanding and making them available for easy reference (Tuttle et al.,2007).

2.3.3.3 *The mixed mode*

The burden of response in establishment surveys and the low motivation for respondents to participate in those surveys are two factors, which motivate statistical agencies to look for ways of easing the response burden and improving the response rates. The common practice is to allow respondents to choose the collection, such as mail, fax, mail out – fax in, web. A growing number of surveys supplement the main mode of collection by other methods (Nicholls, Mesenbourg, Andrews, and De Leeuw, 2000). Another example of mixed mode data collection is making an initial telephone contact to choose the proper person in the establishment as a respondent. Initial contact can also aim to confirm the identification of an establishment and to announce the upcoming survey data collection (Goldenberg et al., 1997). The actual collection is then conducted in a unimodal way and the usage of the initial mix-mode system will reduce nonresponse and has no implications for measurement errors (de Leeuw, 2005). A follow-up contact in the case of nonresponse to elicit response can be yet another example of effective multi-mode collection or sequential multi-mode.

2.4 *Prototype questionnaire*

Major concerns in the questionnaire design process when treated as a whole are:

- introduction – the goals of the survey, status (mandatory or voluntarily), deadline date, collection mode;
- motivation – respondent factors affect every step of the questionnaire construction: goals of the survey must be convincing to the respondent, who should also see benefits resulting from their participation;
- understanding – the logical and concise structure of questions and concepts;
- flow – layout, sections and groups should facilitate an intuitive and clear path from start to completion.

One should emphasise the iterative character of designing, developing and testing questionnaires. A typical iterative cycle starts from developing the initial measurement instrument, which is then reviewed by experts, pretested and, finally, submitted to another revision.

In an effort to work out a consistent image of the surveying agency and to improve data collection instruments that can ease the reporting burden, it is necessary to formulate guidelines for questionnaire design and development. The related aim is also to achieve a coherent “look and feel” of the data collection instrument. These guidelines distinguish several groups of elements that a questionnaire consists of (Morrison, 2007, 2008; Morrison, Dillman and Christian, 2010):

- Question wording – questions should be formulated as sentences ending with question marks, not sentence fragments to be completed; alternatively as imperatives. The question word at the beginning helps to recognise that an answer is expected. It is preferable to have a larger number of simple questions than fewer more complex ones.
- Visual design – the proposed rules can be divided into page layout guidelines and response field options. Theories about the influence of visual design on the question interpretation and comprehension suggest that some layouts contribute to questionnaires being perceived as more friendly and simpler to get through than others. For example, one column format is

easier to follow thanks to its unidirectionality, although two columns may be used for simple numerical data and paper formats. Placing too many graphics and symbols not closely related to the response task is regarded as visual clutter. One way to avoid this is the consistent use of fonts and their attributes for the same purpose throughout the questionnaire. Clutter can also result from vague organisational logic. For example, things placed close to each other seem to be related. This proximity principle can be used to indicate the beginning and the end of one question and the start of the next one. Blank spaces should be used to separate questionnaire items rather than lines, which can break up groups of elements that are, in fact, related. Groupings and separation spaces can also assist the respondent in navigating through the form. The same applies to questions and response options – questions can be linked with answer options by means of leading dots in the case of a paper questionnaire; in an electronic questionnaire shades and colours are available. It is advisable to provide a clear indication of units of required items. Any changes in the flow should be signalled by strong visual cues. Aligning questions helps to perceive the flow to be a natural consecutive path to follow, with answer options placed in one column and along a line.

- Instructions – guidelines generally suggest incorporating instructions into the questionnaire, especially in the case of business surveys, where definitions and descriptions of necessary steps to reach the value required by an item in the questionnaire are of great importance. Reference to separate documents adds burden to the response task and increases the probability of omitting an important detail. Wherever possible, instructions should be incorporated into questions. Bullet lists are more advisable than narrative paragraphs, which produce a more congested impression and require more careful reading.
- Matrices – though users in companies are accustomed to using tabulated data, such as spreadsheets, it is advisable to use tables with caution. They are burdensome and more difficult to comprehend. The decision whether or not to use a table can be made based on the pretesting phase of the questionnaire. At that stage it can be determined whether using a table will not place too much of a demand on respondents' understanding of response options. Properly designed matrices can decrease complexity: a limited number of items, clear navigational path within a table, linkage of rows and columns, using lines to indicate the direction a respondent should follow.

2.5 *Systematic testing*

According to principle 8 of The European Statistics Code of Practice: “In the case of statistical surveys, questionnaires are systematically tested prior to the data collection.” Systematic obligations call for standardised steps to be put into practice. Secondly, testing must be done before the collection phase. Considering all the above, the following steps could be recommended:

- Pre-field testing – differs from field methods in that special conditions are prepared to gather qualitative assessment at the early stage of the design process;
- Field testing – real environment, or rather conditions reflecting the real environment, must be met to evaluate complete questionnaires;

- Evaluation – many business surveys have a longitudinal character. This provides an opportunity for continuous assessment of questionnaires and time for improvements.

2.6 *Questionnaire design standards*

The systematic approach to testing defined in the Code of Practice should be linked to the overall process of questionnaire design. Standards should be applied to practices used for questionnaire design in statistical agencies. Documentation prepared for various stages of questionnaire design and the code of practices provide a coherent and clear picture of a statistical institution as seen from the respondent's perspective. Electronic data collection extends the needs for standards from the visual aspect to the testing protocol. The web data collection environment is characterised by its own dynamic. Standards create the framework for developing and programming techniques to build tools comprising specific components. Treating a questionnaire as an application composed of several components enables developers to determine standards for each component. Specifications for particular components provide standards for question types, field types, function types, validation technique types, layout types. Technical and programming tools are also subject to standardisation in terms of information technology. Guidelines for visual appearance are there to ensure a consistent "look and feel", which involves elements of the screen used for navigation, placement of additional non-content elements, rules for describing fonts, colours, size of text for questions and instructions. Standards for the respondent environment are difficult to describe due to their variability. All of this constitutes a challenge during the testing process. Web browsers are one example. It is not unusual for an application to behave differently when used with different web browsers. One way to solve this issue is to identify operating systems and browsers that most respondents use and then develop testing protocols that are compatible with those platforms.

2.7 *Quality*

Since quality is an essential aspect in the general framework of the "statistical process", the role of the questionnaire should be stressed in this respect. Standards established for quality reports require the inclusion of questionnaires and concise descriptions of the design and testing process. From the perspective of quality reports, three considerations can be mentioned:

- accuracy of statistical output – quality aspects concerning coverage sampling and nonresponse have drawn more attention (Willimack et al., 2004) than measurement errors. Among the sources of measurement errors are the mode of data collection and questionnaire design. The survey measurement instrument is one of the sources of measurement error, which is under direct control of the statistical agency (Bavdaž, 2010b). A good survey instrument can improve reliability and validity of statistical output.
- cost and respondent burden – the quality of respondent's estimates is inferior to a value obtained from records (Willimack et al., 1999). The required data retrieval process can be considered as one of the most burdensome elements of the response process (Willimack et al., 2010). Further, the more burdensome the retrieval process, the more inaccurate reported values are likely to be, ultimately leading to nonresponse.

- user needs and satisfaction – this requires attention at an early stage as well as during the process of assessing the statistical output: as a result, the measurement instrument is under continuous monitoring and assessment.

2.8 *Omnibus surveys*

The omnibus survey is a special kind of survey in which a respondent is asked questions on different topics. The goal of such a survey is to provide multi-subject information collected in a relatively short period of time and at low costs for a group of clients. To satisfy various user's needs the questionnaire must cover a number of different topics. Advantages of such a solution include:

- costs efficiency – clients are charged for questions they want to ask. The sampling and data collection costs are shared between all of them. If the user only wants to ask a few questions then doing this through an omnibus survey can be an effective way to satisfy research objectives at a reasonable cost.
- time efficiency – there is no need to organise resources for all sets of questions separately. Therefore, assuming the field schedule and frequency are flexibly planned there is a chance to get the results quicker than in a custom study.

Among the problematic issues of omnibus surveys are:

- sampling – since the sampling framework is predetermined for all clients who submitted questions for the survey it can be difficult to meet individual criteria and requirements, for example when the target population is to be “small establishments with less than ten employees”. Then the sample might not be large enough to elicit responses from the proper number of respondents for estimations.
- the multi-topic coverage – the goal of the survey is to cover several topics in one questionnaire. One topic section may affect the comprehension of the other. It is difficult to negotiate the order of questions and the impact of a sequence of questions is hard to assess. At the preparation stage different authors devise the questions separately. When the subject is changing, the user should be informed that a new topic is about to be introduced. Therefore, there is a need for an indication of subject change.
- complex questions – since a couple of subjects need to be covered each probably with a set of questions it is inappropriate to include long instructions and definitions. Complex questions introducing skip patterns and many multiple choices are not suitable.

2.9 *Ad hoc surveys*

An ad hoc survey is a survey without any plan for repetition. It is also possible to add ad hoc questions to a questionnaire used in a regular survey. Ad hoc modules included in questionnaires play their role as complements to the main modules. Incorporating additional modules creates an opportunity to provide data on different subject or specific parts of the survey subject. On the other hand adding other modules increases burden imposed on respondents, which may affect the quality of responses.

3. **Design issues**

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

ABS (2010), *Forms Design Standards Manual*. Australian Bureau of Statistics.

Brancato, G., Macchia, S., Murgia, M., Signore, M., Simeoni, G., - Italian National Institute of Statistics, ISTAT, Blanke, K., Körner, T., Nimmergut, A., - Federal Statistical Office Germany, FSO, Lima, P., Paulino, R., - National Statistical Institute of Portugal, INE, and Hoffmeyer-Zlotnik, J. H. P., - German Center for Survey Research and Methodology, ZUMA (2006), *Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System*.

Bavdaž, M. (2010a), The multidimensional integral business survey response model. *Survey Methodology* **36**, 81–93.

Bavdaž, M. (2010b), Sources of measurement errors in business surveys. *Journal of Official Statistics* **26**, 25–42.

Goldenberg, K., Levin, K., Hagerty, T., Shen, T., and Cantor, D. (1997), Procedures for reducing measurement error in establishment surveys. Presented at the American Association for Public Opinion Research, Norfolk, Virginia, May 1997.

de Leeuw, E. D. (2005), To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics* **21**, 233–255.

Lorenc, B. (2007), Using the Theory of Socially Distributed Cognition to Study the Establishment Survey Response Process. Paper presented at the ICES-III, June 18-21, 2007, Montreal, Quebec, Canada.

Morrison, R. (2007), Towards the Development of Establishment Survey Questionnaire Design Guidelines at the U.S. Census Bureau. Paper presented at the ICES-III, June 18-21, 2007, Montreal, Quebec, Canada.

Morrison, R. (2008), Writing Revising Questionnaire Design Guidelines. Component of Statistics Canada Catalogue no. 11-522-X Statistics Canada’s International Symposium Series: Proceedings.

Morrison, R., Dillman, D., and Christian, L. (2010), Questionnaire Design Guidelines for Establishment Surveys. *Journal of Official Statistics* **26**, 43–85.

- Nicholls II, W. L., Mesenbourg Jr., T. L., Andrews, S. H., and De Leeuw, E. (2000), Use of New Data Collection Methods in Establishment Surveys. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Buffalo.
- Sudman, S., Willimack, D. K., Nichols, E., and Mesenbourg, T.L. (2000), Exploratory Research at the U.S. Census Bureau on the Survey Response Process in Large Companies. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 327–337.
- Tuttle, A. D., Morrison, R. L., and Willimack, D. K. (2007), From Start to Pilot: A Multi-Method Approach to the Comprehensive Redesign of an Economic Survey Questionnaire. Presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.
- Tourangeau, R. (1984), Cognitive science and survey methods: a cognitive perspective. In: Jabine, T., Straf, M., Tanur, J., and Tourangeau, R. (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, National Academy Press, Washington, DC.
- Willimack, D. K., Nichols, E., and Sudman, S. (1999), Understanding the questionnaire in business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 889–894.
- Willimack, D. and Nichols, E., (2001). Building an Alternative Response Process model for Business Survey, *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9*.
- Willimack, D. K., Lyberg, L., Martin, J., Japac, L., and Whitridge, P. (2004), Evolution and Adaptation of Questionnaire Development, Evaluation, and Testing Methods for Establishment Surveys. In: Presser, S., et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, Chapter 19, Wiley, New York.
- Willimack, D. and Nichols, E. (2010), Hybrid Response Process Model for Business Surveys. *Journal of Official Statistics* **26**, 3–24.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Electronic Questionnaire Design
2. Questionnaire Design – Editing During Data Collection
3. Questionnaire Design – Testing the Questionnaire
4. Data Collection – Main Module
5. Response – Response Process
6. Response – Response Burden

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Sub-process 3.1

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

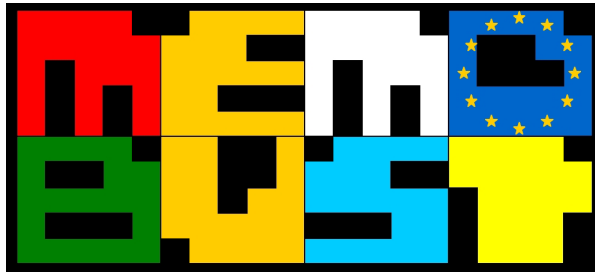
Questionnaire Design-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	04-03-2012	first version	Paweł Lańduch	CSO Poland
0.2	09-04-2013	second version	Paweł Lańduch	CSO Poland
0.3	18-09-2013	third version	Paweł Lańduch	CSO Poland
0.3.1	18-12-2013	minor improvements	Paweł Lańduch	CSO Poland
0.4	21-01-2014	improvements after EB review	Paweł Lańduch	CSO Poland
0.4.1	22-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:26



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Electronic Questionnaire Design

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Response process.....	3
2.2 New design features	4
2.3 Visual design	7
2.4 Usability	8
2.5 Evaluation and testing	10
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	11
Interconnections with other modules.....	13
Administrative section.....	14

General section

1. Summary

The electronic questionnaire can be considered as a complete software system, with a list of requirements the software must meet. This determines the approach to how the questionnaire is designed and tested. One dimension of this approach concerns the questionnaire's objective and its conceptual layer; the other one comprises the technical application of information and software system tools. Thanks to technological development, some aspects of the processing stage of the survey can be performed during earlier stages, such as data collection. The term "computer-assisted" implicates the design stage and the data collection phase. This term is also related to another aspect of the design, namely the question of who is to administer the software application at the data entry stage. Deciding whether the respondent is to be the user or the interviewer influences the preparatory stages of the questionnaire. Also, a successful completion of online forms depends on access to the public network, while the respondent must be equipped with a local computer system. Maintenance of the software is the task of the surveying agency.

2. General description

2.1 Response process

Eliciting responses in surveys can be treated as a task. The task approach analysis has distinguished several steps in this process making the foundation for response process models. This section treats the steps of the response process model as a background for a brief description how electronic technology may affect the response process. There is another module in the Handbook devoted to response process models, namely "Response – Response Process", where more information can be found on this subject. The cognitive approach to improving measurement instruments has gained broad acceptance. The four-step model comprising comprehension, retrieval, judgment and communication for social surveys has been expanded to suit the needs of the response process in business surveys. The aspect of cognition has been augmented to include an organisational frame. The response process in business surveys is more complex than in social surveys. The advent of electronic data collection adds another dimension of burden to the response process. From the cognitive perspective, the application of the electronic mode of data collection can be viewed in the light of its impact on the subsequent steps of the response process model. In the hybrid response model (Sudman et al. 2000, Wilimack and Nichols, 2001) *record formation step*, constituting the top of the model structure, is connected with data maintained by business systems and their management goals and the knowledge of those systems. The *respondent selection and identification step* refers to the cooperative nature of response in establishments. The respondent or rather, in the case of business surveys, the informant or co-ordinator, gathers data from various sources in the organisation. Hence, the need for a tool in the questionnaire software to enable propagation of a part of the questionnaire as well as the import and export of data. Another solution is to enable the option of printing a draft questionnaire to gather pieces of information from multiple sources. The *assessment of priorities* step, which recognises that the response task is treated as a non-productive activity from the point of view of a business, is followed by the task of *comprehension*, which, unlike the paper-based questionnaire, includes additional tasks. For example, limited computer skills can be an impediment in completing the task. Thus, while designing the questionnaire, the user must always remain in the foreground of the process:

it should be a complete tool equipped with clear instructions and an intuitive interface. *Retrieval of relevant information* can require additional assistance from the IT staff, which can be another burdensome factor. The electronic questionnaire contains internal editing, which is designed to monitor the logic and validity of submitted data at the *judgement of the adequacy of the response* stage. If the *Communication of the response* step is to be successful, information should be reported in a proper format. There can be a need to resolve format edits before data submission can be made. *Release of the data* requires a number of tasks to be performed to make sure that the data have been received by the statistical agency. To minimise this additional burden, the electronic questionnaire should enable the respondent to ensure the mandatory reporting has been fulfilled.

2.2 New design features

Screen layout – moving a piece of paper to the computer screen raises the question of how the paper content should be presented on the screen. The mixed mode of collection, used in business surveys, adds another question about whether the paper and its electronic counterpart should be similar in appearance. One option is to put all the questions on a single page. This would most probably require scrolling to navigate through the questionnaire. Another option is to display a group of items or sections on multiple pages. Dillman (2000) suggests that questions in electronic questionnaires should be presented similarly to those in paper counterparts. On the other hand, the use of skip patterns and interactive processing features rule out strict similarity between the two modes. The use of a two column format or a grid poses yet another problem: the expected order of answering (vertical or horizontal), which the user may fail to follow (Abraham et al., 1998). There is a need for connections between the pages with a clear way of navigating the questionnaire. Locating the place where the user actually is and the possibility to freely navigate through the entire questionnaire is a factor making the instrument easier and more comprehensible (Snijders et al., 2007).

Editing – an activity aimed at detecting and correcting errors conducted with paper-based data collection as a post-collection processing, with the advent of electronic data collection has become part of the collection itself. Among the objectives to achieve are better data quality and cost reduction for post-collection editing. This is also an opportunity to reduce burden (Dowling, 2006). Editing rules, called edit checks or edits, are incorporated into the measurement instrument. In the case of an interviewer-administered data collection, he/she is instantly informed about failing an edit check. In a self-administered collection a respondent is notified about errors and should resolve edit rule failures. Questions arise as to what type of edits can be incorporated, searching for a balance between what users find acceptable and what is effective. A further dilemma is how to present messages about data that do not satisfy edit rules contained in the instrument. Another question is when such messages should be presented to the user: immediately after a value was typed in or after the entire portion of data was entered. Two types of edits can be distinguished: edits requiring data to meet editing criteria unconditionally – called hard edits, and soft edits – treated as a warning, which do not prevent the user from finishing and submitting the questionnaire. If there is a high probability of triggering numerous edit checks by a respondent the number of edits incorporated into the questionnaire should be reduced. (Nichols et al., 2006). According to the usability principles as much as possible should be left under the users' control; it is therefore recommended that respondents be allowed to submit data with unresolved edit rules to prevent non-response and respondent's perspective to provide most accurate data they have. Schonlau et al. (2002) advise placing edit messages close to the item, but the study

conducted by Mockovak (2005) demonstrates that especially soft edits are frequently omitted, regardless of the placement of messages. However, not all post-collection editing processes can be moved to data collection editing. For one thing, some corrections can only be made based on an overview of all the collected data; secondly, complicated correction rules may be hard for respondents to understand; finally, they may be difficult to implement in the electronic questionnaire. For information on the data editing process in business surveys, the reader should refer to the topic “Statistical Data Editing”.

Automatic routing – one of the main features of an electronic questionnaire is the use of automatic routings. Unlike paper questionnaires, where respondents can choose the order of questions, electronic ones with automatic routings eliminate routing errors (Leeuw et al., 1998). Skipping questions that do not apply reduces data errors. Previous answers influence the order of consecutive questions. This raises the matter of numbering the questions. Automatic routing can result in a situation where question number 3 is followed by question number 5. One solution is to “grey out” the inapplicable questions (Potaka, 2005), i.e., retain them on the screen without the ability to select them. This requires information for the respondent that greyed out elements do not require answers. Another factor was pointed out in a paper by Abraham et al. (1998). In interactive interviews, questions resulting in a skip pattern can be placed last on the page to make sure that the following questions are in the right order.

Calculations – adding up, subtracting and performing other calculations are expected to be done by computer (Snijkers et al., 2007). This can be viewed as part of keeping data consistency and validity. As such, it is part of edit checks and validation policy. However, it must be clear to the respondents which items are added up. A related issue is where the results of calculations are to be placed. Figures placed at the bottom of the page and difficult to find can cause confusion. Another example of applying automatic calculations is an additional functionality attached to a questionnaire item as a pop-up window where preliminary calculations can be performed to obtain the necessary figure (Snijkers et al., 2007).

“Fill” capability – based on previously provided answers a computer-assisted interview can permit tailoring of the question wording. This functionality can improve question comprehension in interviewer-assisted surveys. In this way the burden imposed on the interviewer is diminished and can contribute to improved measurement. However, there is no empirical research on the effects of easing the burden by using “fills” (Groves and Nichols, 1986). A similar solution can be applied to groups of items or sections, whose results are to be added up and used as elements of other sections – in this case, those sums can be carried over automatically. Such a functionality, however, should follow an explicit logic so that the respondent is aware of the origin of the number (Snijkers et al., 2007).

Progress indicator – in a paper questionnaire the respondent can easily check the completion status by leafing through its content. This possibility is also expected by users in its electronic counterpart (Snijkers et al., 2007). However, being able to observe one’s progress has some disadvantages (Dowling, 2006). If progress is perceived to be slow, the respondent can get discouraged and, in effect, abandon the questionnaire. Further, skipping patterns can be seen as a hindrance in establishing the exact state of completion when the questionnaire is tailored to a specific section of respondents. All in all, it seems that some kind of indicator of completion is desirable. Progress indicators can be presented in graphical formats as well as text.

Instructions – business surveys rely heavily on instructions. The likelihood of respondents using instructions diminishes with the growing effort required to find them. However, even making instructions more noticeable can produce limited results (Willimack, 2008). Nonetheless, respondents should be able to easily find instructions should they need additional explanation. Guidelines on questionnaire design advocate placing instructions close to questions. One method to place instructions in an electronic instrument is to hide them under a hyperlink, which can be clicked to open a pop-up window. Another approach is a hovering text appearing when the respondent is moving the mouse pointer over an element. It seems to be a good idea to follow paper form guidelines by placing essential instructions close to questions, i.e., within the text of the questionnaires. Instructions available by clicking a button should attract attention and be brief and clear (Snijkers et al., 2007).

The navigational path – there are several common guidelines both for paper mail questionnaires as well as electronic instruments. Those concern grouping similar items, separating various sections, using visual features and so on. However, with some aspects of electronic questionnaires, it is not always obvious if they should not be comparable to their paper counterpart. For example, moving backward and forward through a paper questionnaire is easy: all it requires is flipping through the questionnaire booklet. This way the respondent can review the previous answers and possibly correct them. Completing the questionnaire items can be interrupted at any time and resumed later. Hence, incorporating similar functionalities in electronic instruments is desirable when seeking user acceptance. The navigational bar can serve as a tool to locate a desired item. Two ways of navigation – an index of all sections of the questionnaire and a navigation button – are examples of simple and clear navigation (Snijkers et al., 2007). Providing the instrument with an option of saving the current state of work is a way to solve the problem of completing the questionnaire during several sessions. Another question is navigation between fields. This function should be consistent with other computer programs. Two typical methods are used in computer programs: using the “Enter” key and the mouse pointer.

Importing and exporting – data for business surveys are most likely stored in business records. Moreover, before completing a questionnaire item, information must often be gathered from different sources. A questionnaire offering the function of exporting templates for data preparation and importing data from spreadsheets, commonly used in the accounting environment, can facilitate the process of data retrieval and preparation. However, exploratory studies (Hak et al., 2003) indicate that respondents are not familiar with technical terminology, such as importing and exporting, but they find the ability to export the questionnaire or part of it useful. Another option is to make the exchange of data between business systems and statistical agencies automatic, which is called Electronic Data Interchange (Willeboordse, 1998). For EDI the reader may also refer to the “Data Collection” topic in the handbook. Procedures for extracting data from business records must be implemented in respondents’ systems. Establishments are reluctant to devote resources to it (Nicholls et al., 2000). Another related problem is how close statistical definitions meet business concepts, which calls for the need to define formats of data structure and coding conventions. For issues connected with data collection issues the user is referred to “Data Collection” topic.

Printing options – respondents may wish to print either the blank questionnaire or its completed version. This may be in line with their working practices or in order to review the entire questionnaire (Dowling, 2006). Another reason may be archiving purposes or as reference materials for future use (Morrison et al., 2005). This feature can be treated as an additional back-up to saving an electronic

copy. The need for paper copies may also be motivated by the necessity to collect data from different departments or to consult employees from company branches (Snijkers et al., 2007).

Security – the confidential nature of business data raises the question of data security. In the case of electronic questionnaires, this involves restricted access to the questionnaire software and safe transmission to a statistical agency. As for software launched locally, authorisation may be required to submit the data; in a web environment the user must log in to access the questionnaire. In both cases, the respondent must obtain an identification symbol and a password. Because of the compulsory status of business surveys and the need for users to ensure that their data have been submitted successfully, the statistical agency must implement a feature for respondents to verify the status of deliverance. It should be remembered, however, that security requirements are usually in conflict with the ease of use, and contribute to the respondent burden (Dowling, 2006).

Auditing – in computer assisted data collection while the respondent filling in the questionnaire, some sort of information about the process can be gathered. Parallel to the activities of the user connected with completing the questionnaire items, the program can collect administrative data behind the scenes. These automatic data captured during the survey computer data collection are called *paradata* (Couper et al., 2010). The examples include the completion time, keystroke data, software failures. This kind of information can be used to track problems with questions and monitor the ongoing survey process. After the data have been collected paradata can serve for evaluation. The usage of paradata can be the foundation for interactively tailoring the dialog with the respondent (Haraldsen, 2013).

2.3 Visual design

In designing elements of the questionnaire visible on the computer screen and where perception could influence the question-answer process, it is useful to follow the principles of Gestalt psychology (Morrison et al., 2008, p.10):

- proximity – objects close to each other form a group of objects connected with each other in some way;
- similarity – the same font size and colour suggests a relationship;
- Prägnanz – the simpler objects are, the easier to understand and remember.

When planning the arrangement and order on the computer screen and preparing general recommendations for electronic questionnaire design, it is good to take into account the following:

Fonts – consistent use of font size, style and contrast can facilitate understanding and work with the questionnaire. Decisions once taken should be kept throughout the questionnaire. Example: Use of bold font for questions, standard text for a list of answers. The use of various fonts can thus be seen as logical and clear.

Colours – distinguishing answer spaces against the background helps the respondent to recognise where the space for entering data begins. Colours can help distinguish parts of the screen that serve different purposes.

Similarity – questions where the same kind of data is required should be of the same type and size.

Groupings – relations between questions will be emphasised by arranging them in sections, divisions, etc. and consistent numbering and giving titles. Elements in close proximity are perceived as belonging to the same group.

Graphical symbols – graphics plays an important role in questionnaire layout. Placing too many symbols or irrelevant symbols contributes to what is called “visual clutter” and distracts the respondent.

2.4 *Usability*

The term ‘usability’ covers issues connected with how to design products that will be user friendly and understandable for those they are intended to serve. This concern for usability puts the user at the centre of the designing process. The application of electronic instruments for survey data collection has opened new possibilities but has also posed new challenges. Complex branching or editing during data collection are just two examples of the potential of electronic questionnaires that are not available with paper ones. Adding more functions to products increases the list of requirements that have to be met during developing and testing. The desire for better effectiveness and efficiency leads to improved usability and clarity, which contributes to a positive perception of the product, which is not seen as imposing an unnecessary burden. Principles of visual design and the theory of usability can be applied to improve both paper and electronic questionnaires. A paper by Dillman, Gertseva, Mahon-Haft (2005) describes an example of combining the visual design theory and cognitive psychology to improve the usability of a paper questionnaire for business surveys. Cognitive psychology, which describes people’s emotional reactions to various elements and the way visual design conveys meaning and affects comprehension, provides the basis for optimal questionnaire design. Norman (1988) laid the foundation for an approach to designing products, which can also be used to design, develop and test computer-assisted questionnaires. The starting point is the observation that things have their own psychology. The psychology of things manifests itself in the way people react when dealing with products. Based on this observation, general principles of design can be formulated, which can also inform rules for designing electronic questionnaires and, later on, developing and testing. Inspiration for a good design can be drawn from principles of:

- visibility – this principle stresses the need for the user to recognise the purpose of design associated with a particular feature of the product. One example may be applying the principle to questionnaire design, in particular, visual design, font variation or use of colours for different purposes. The logic and consistent use of the same font for the same purpose facilitates understanding and clearly communicates function by means of a visual feature.
- mapping – mapping connects the designing control of the function with the results of its execution. Mapping should be easy to understand. According to the theory, good mapping should be natural in the sense that the function is visible and its result complies with the user’s expectations. The supposed effect is easy to understand when it belongs to the cultural environment and represents a standard operation. If one function is associated with a single purpose and equipped with a clear description, then it is simpler to comprehend.
- feedback – an action returns a signal of its effect. In addition to visibility and mapping, feedback is an important dimension affecting the use of products. Advances in technology have resulted in

many new ways of performing jobs and tasks. New functionalities are constantly being added to existing products. A sense of control over the product is conveyed to the user by feedback.

Several other principles can be listed based on the cognitive approach theory:

- evolutionary road – the designer's perspective and the target user's perspective are different. The gap between them can be bridged through iterative steps. Usable and understandable products are developed through the process of evolution. The product must be submitted to constant evaluation. Before a computer application reaches the user, it must be tested to assess a questionnaire is working properly enough to be used in the field. Of course, the scope of testing is limited by time and cost constraints. By submitting the product to the assessment of end users, a scope for revision is created. Thus, through a series of continuous improvements the tool is becoming more invisible, while the goals it is designed to achieve are becoming more visible. An example of a good computer program pointed out by Norman (1988) is the spreadsheet. Spreadsheets are used to simplify complicated calculations and for this reason are appreciated.
- user-centre design – the theory discusses conceptual models of design: the image of the product is provided by the designer model and the user model. The process places the user at the centre. Sensitivity towards user needs implies avoiding an arbitrary choice of performing the required action. One of the technological development goals is to make the task simple to perform or effective. The simplification can be achieved by providing additional clues which make the task simpler and ease the comprehension burden. Manuals play an important role in this respect. The more complex a product is, the more instructions it requires. This is the case when it comes to business surveys, which tend to be based on intricate, technical definitions and concepts. Interestingly, as exploratory testing demonstrates, users tend to ignore instructions contained in separate files.
- designing for errors - making errors is a natural trait of human behaviour. Owing to time, cost and other constraints, the product itself cannot be perfect. There are various sources of error ranging from memory limitations and automaticity of action to similarity between operations. Built-in rules trigger an action and alert the user when a rule has been violated. The functionality of a computer program hides under commands and actions attached to them. Successful completion of an action depends on effective communication. Errors should be communicated in such a way as to encourage cooperation rather than be perceived as orders. It is advisable not to assume an imposing position towards the user and not to treat errors as a kind of negative behaviour. In other words, the language should be concise and polite and the terminology should be closely related to the subject matter the user is familiar with. Another principle stresses that control should be in the user's hands. Errors can be communicated in two ways: as warnings and as orders. Warnings are often ignored. The balance between the soft and hard treatment of errors is therefore a matter which deserves careful consideration.
- standardisation – standards provide a uniform way of perceiving rules of behaviour and consent so that representations of objects are understood in the same way. A consistent use of colours or symbols, always for the same purpose, is one example of establishing standards. Another one is a clear mapping that connects the visual representation with its meaning. Adopting standards already used in the surrounding world is a natural and cultural constraint; such constraints narrow

down the field of possibilities. Applications of computer technology have not been around for long enough to pervade established standards and change quickly, which is why standards must evolve.

2.5 *Evaluation and testing*

Since the user-computer interaction is a key factor in developing electronic questionnaires, usability testing should be user-oriented. Testing should focus on interaction, where design and layout are the main features to be assessed.

Functionality testing is the second important kind of testing when CAI type questionnaire is used. Different functionality specifications need to be compiled depending on the mode of data collection. Aspects to be considered include such things as whether interviewing is performed by the interviewer or is self-administered, whether questions are asked face to face or by phone. All these decisions affect the testing plan and methods. The testing procedure is labour intensive and it is difficult to be sure if all errors have been found. However, the goal of testing is to obtain enough confidence that the questionnaire is working as described in requirements for implementation and minimise the risk of unexpected behaviour.

The heuristic approach to assessment can help to identify interface elements which need to be revised in order to improve the overall level of user satisfaction. Analysing the questionnaire in terms of cognitive assessment criteria can be a very important step, which can positively affect the whole survey process. General principles formulated by Nielsen (1994) are an example of rules that can be used for purposes of assessing and testing the electronic questionnaire.

The electronic questionnaire is a complex measuring instrument. However, despite its internal complexity, it should have a user-friendly interface. This is why testing procedures require a multi-dimensional approach. One dimension is concerned with the questionnaire as an instrument for collecting statistical facts, which is the purpose of the statistical process. The other one represents the technical perspective, where the questionnaire is treated as a piece of software.

3. *Design issues*

4. *Available software tools*

The Blaise® system is a widely-used, powerful, and flexible tool for computer-assisted data collection and processing. The Blaise language is well-suited to create computer questionnaires, from easy ones to complex instruments and surveys with hierarchical data structures. Blaise® is a registered trademark of Statistics Netherlands.

5. *Decision tree of methods*

6. *Glossary*

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Abraham, S. Y., Steiger, D. M., and Sullivan, C. (1998), Electronic and mail self-administered questionnaires: A comparative assessment of use among elite populations. In *Proceedings of the Section on Survey Research Methods*, 833–841.
- Couper, M., Kreuter, F., and Lyberg, L. (2010), The use of paradata to monitor and manage survey data collection. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 282–296.
- Dillman, D. A. (2000), *Mail and Internet Surveys: The Tailored Design Method* (2nd ed.). John Wiley & Sons Inc., New York.
- Dillman, D., Gersteva, A., and Mahon-Haft, T. (2005), Achieving Usability in Establishment Surveys through the Application of Visual Design Principles. *Journal of Official Statistics* **21**, 183–214.
- Dowling, Z. T. (2006), Web data collection for mandatory business surveys: an exploration of new technology and expectations. Thesis submitted for the degree Doctor of Philosophy in Sociology, Department of Sociology, University of Surrey, United Kingdom.
- Groves, R. M. and Nicholls II, W. L. (1986), The Status of Computer-Assisted Telephone Interviewing: Part II – Data Quality Issues. *Journal of Official Statistics* **2**, 117–134.
- Hak, T., Willimack, D., and Anderson, A. (2003), Response Process and Burden in Establishment Surveys. *Proceedings of the 2003 Joint Statistical Meetings – Section on Government Statistics*, 1724–1730.
- de Leeuw, E. D., Hox, J. J., and Snijders, G. J. M. E. (1998), The effect of computer-assisted interviewing on data quality. *Market Research and Information Technology*, 173–198.
- Mockovak, W. (2005), Comparing the Effectiveness of Alternative Approaches for Displaying Edit-Error Messages in Web Forms. Bureau of Labor Statistics, Statistical Survey Papers.
- Morrison, R. L., Anderson, A. E., and Brady, C. F. (2005), The Effect of Data Collection Software on the Cognitive Survey Response Process. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Morrison, R. L., Stokes, S. L., Burton, J., Caruso, A., Edwards, K. K., Harley, D., Hough, C., Hough, R., Lazirko, B. A., and Proudfoot, S. (2008), *Economic Directorate Guidelines on Questionnaire Design*. U.S. Census Bureau, Washington, DC.
- Nicholls II, W., Mesenbourg, T. L., Andrews, S. H., and de Leeuw, E. (2000), Use of New Data Collection Methods in Establishment Surveys. *Proceedings of the 2nd International Conference on Establishment Surveys*, American Statistical Association, Alexandria, VA, 373–382.
- Nichols, E. M., Murphy, E. D., Anderson, A. E., Willimack, D. K., and Sigman, R. S. (2006), Designing Interactive Edits for U.S. Electronic Economic Surveys and Censuses: Issues and Guidelines. *Statistical Data Editing, Volume No. 3: Impact on Data Quality*, 252–261.
- Nielsen, J. (1994), Heuristic evaluation. In Nielsen, J. and Mack, R.L. (eds.), *Usability Inspection Methods*, John Wiley & Sons, New York, NY.

- Norman, D. A. (1988), *The Psychology of Everyday Things*. (Republished as *The Design of Everyday Things* in 1991 and 2002.) Basic Books, New York.
- Potaka, L. (2005), Comparability and Usability: Key issues in the design of Internet forms for New Zealand's 2006 Census of Population and Dwellings. *Proceedings of the 5th QUEST Workshop*.
- Schonlau, M., Fricker, R. D., and Elliott, M. N. (2002), Conducting Research Surveys via Email and the Web. RAND, Santa Monica, CA.
- Snijkers, G., Onat, E., and Vis-Visschers, R. (2007), The annual structural business survey: Developing and testing an electronic form. In *Proceedings of the Third International Conference on Establishment Surveys, Montreal, Canada*, American Statistical Association, 317–326.
- Sudman, S., Willimack, D. K., Nichols, E., and Mesenbourg, T. L. (2000), Exploratory Research at the U.S. Census Bureau on the Survey Response Process in Large Companies. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 327–337.
- Willeboordse, A. (ed.) (1998), *Handbook on the Design and Implementation of Business Surveys*. Eurostat, Luxembourg.
- Willimack, D. K. (2008), Issues in the Design and Testing of Business Survey Questionnaires: What We Know Now that We Didn't Know Then – and What We Still Don't Know. *Proceedings of the Conference on Reshaping Official Statistics, International Association on Official Statistics, Shanghai, China*.
- Willimack, D. and Nichols, E. (2001), Building an Alternative Response Process model for Business Survey. *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9*.

Interconnections with other modules

8. Related themes described in other modules

1. Data Collection – Main Module
2. Response – Response Process
3. Statistical Data Editing – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

1. Blaise

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

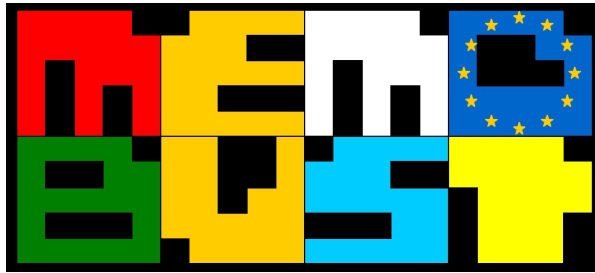
Questionnaire Design-T-Electronic Questionnaire Design

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	29-03-2012	first version	Paweł Lańduch	GUS
0.2	22-04-2012	second version	Paweł Lańduch	GUS
0.3	05-09-2013	third version	Paweł Lańduch	GUS
0.3.1	28-01-2014	minor revisions – EB review	Paweł Lańduch	GUS
0.3.2	29-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:26



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing During Data Collection

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Communication of the need for correction.....	3
2.2 Types of edit rules	4
2.3 The measures to avoid errors.....	5
2.4 Presentation of the messages to the respondent.....	7
2.5 Testing and evaluation of editing strategy.....	9
2.6 Electronic documents	9
2.7 Conclusions	10
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

Data editing is the process of “improving” collected survey data. The improvement involves finding erroneous data and then correcting them. Errors may have happened along the way from the respondent to the survey organisation’s data files for various reasons, intended or unintended. Examples include typing errors, wrongly estimated values, misclassifications. Omission or answer denial can also be a source of measurement error. Up to about 40% of statistical agency’s resources is spent on editing and imputing missing data (De Waal et al., 2011). In mail business surveys the editing process is performed at the post-collection phase of the survey. The advent of computer technology has enabled statisticians to shift data editing to the data collection stage. Some types of data editing tasks can be performed at the data collection phase. Editing was first incorporated into data collection in the CATI mode. The interviewer is assisted by an electronic questionnaire, which is a program running on his computer. The program contains a built-in set of editing rules, called *edit checks or edits*. These rules assess whether the response is allowed by survey criteria or should be discarded, that is whether an edit is satisfied or violated. Mobile computers extend the field of editing to CAPI. The interviewer conducts a face-to-face interview using an interactive computer program with embedded edit checks. Computer self-administered questionnaires also adopt editing rules, in which the editing process is performed by the respondent. The increasing use of the Internet entails a shift to another mode of survey data collection: online data collection. The prevalent self-administered data collection mode in business surveys and the use of computer questionnaires with incorporated edits enable the editing process at the respondent level. This solution results in many benefits: it decreases costs, improves data quality and response rates and lowers the perceived response burden. For the general issues of data editing in business surveys the user is referred to the topic “Statistical Data Editing”.

2. General description

2.1 *Communication of the need for correction*

The goal of editing at the time of data collection is to take advantage of the measurement instrument to improve quality of the data and reduce the costs of the post-collection process. Data typed into the questionnaire are checked for their correctness. This requires to define the conditions that must be met to assume the response is accurate. The response item has a built-in edit rule to inform the user about an error in case the rule is not satisfied. This leads to the definition what is meant by assuming data are erroneous or data are supposed to be suspicious. Typically, validation in software technology, when a reaction is expected from a user or a user should be informed about something, is notified in a dual way: like an error marked in red colour which means the situation is unacceptable and must be changed in order to continue and a warning which notifies the possibility of incorrectness or to draw attention to a certain aspect of working being the consequence of earlier choices. Moving to the editing field rules that must be satisfied unconditionally – called hard edits – prevent the user from going further or from submitting data to the statistical agency. A second kind of edit rules can be called warnings or soft edits. These kind of edits only notify users that an item should be assessed for its adequacy. In this case three types of resolution of that kind of failures can be pointed out: correction, comment or no action. However, no action should be confirmed by respondents as their selection.

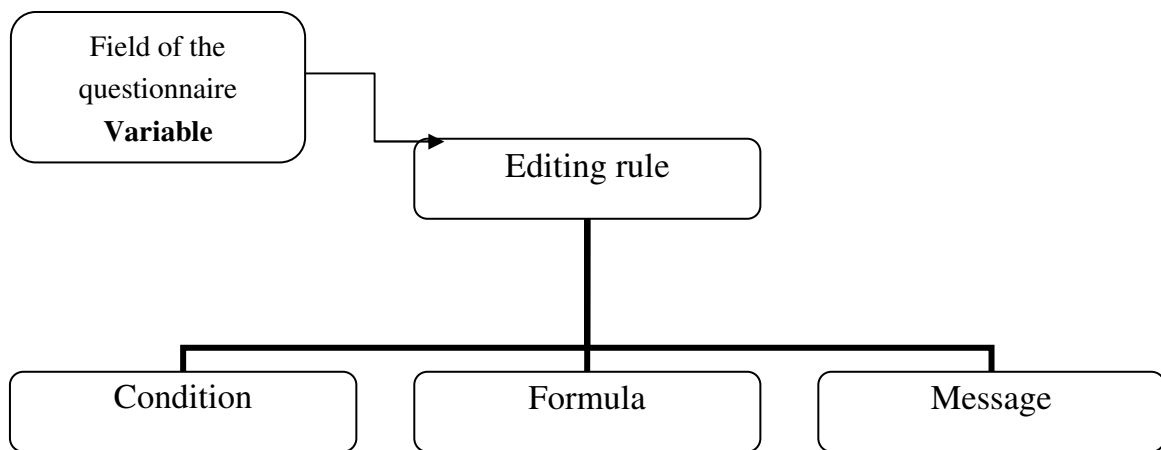
Another treatment of error messages can be pointed out, namely “unsolicited clarification” (Haraldsen, 2013).

2.2 Types of edit rules

An edit rule can be understood as a logical formula triggered when a condition is met under which the variable is tested for its correctness.

IF (variable \in editing set) THEN (testing formula).

Below the schematic diagram of an edit check in the questionnaire is presented:



Correctness may depend on the type of the variable:

- a formula that allows keying only some type of characters, e.g., numeric characters,
- a required item check – the edit rule stipulates that the field cannot be empty: null values are not allowed,
- formulas for numeric fields – Let X be a variable denoting, for example, turnover, then edit rules for instance can check:
 - o $\text{Minimum value} \leq X \leq \text{maximum value}$ (range constraint),
 - o $X \geq 0$ (non-negative number requirement),
- a formula that sets a length limit for text variables, the number of characters to be entered is limited,
- an edit rule allows only a specific pattern in the field, for example an e-mail address must contain the @ character.
- edits that check relationships between two or more values:
 - o balance edit – checking if the sum of selected items equals a total value,
 - o logical formula – various types of relationships between variables (also called inter-item rules), e.g., equality, inequality, greater than, less than, ratio edit, other types of logical relations between two or more variables.

2.3 The measures to avoid errors


The items of electronic instruments and their features can be used as cues to help respondents to complete the responses. There is a possibility of an interactive way of communication between a respondent and a questionnaire. Moreover, the greater usage of a web data collection is an opportunity to tailor the measurement instrument to an individual respondent context. Haraldsen (2013) talks about “questionnaire communication” instead of questionnaire design, stressing the role of the questionnaire as a way to communicate the request for business data. The context is set to technological environment as a shift from paper one-way communication to a dynamic two-way self-administered exchange of information.

- Information from the business register determines the obligation to convey data to various types of surveys. According to size and kind of activity a list of such surveys, devoted only to the distinguished respondent, can be presented after the user logs in to the web portal which is a communication point for data collection by using electronic questionnaires.
- Access to certain modules of the questionnaire can be determined from answers to previous questions. This means that certain skips and filters can activate when the questionnaire is loading. Also, only selected variables can be enabled for editing. Not only can this improve data consistency but also diminish the response burden. Whether a variable is enabled for editing may depend on previous answer(s). Automatic routing can sometimes lead to a gap in the numbering. Below is an example of such a result. A solution can be to use a two-level numbering.



Ulica		ul. Adama Asnyka
6	Jaki jest główny lub przeważający rodzaj działalności zakładu pracy, który jest Pana(i) głównym miejscem pracy?	Administracja budynków
7	Ile godzin zwykle Pan(i) pracuje w ciągu tygodnia w głównym miejscu pracy?	20
8	Czy w tygodniu od 25 do 31 marca 2011r. miał(a) Pan(i) pracę dodatkową?	<input type="radio"/> tak <input checked="" type="radio"/> nie
22	Czy jest Pan(i) użytkownikiem gospodarstwa rolnego lub członkiem gospodarstwa domowego z użytkownikiem?	<input type="radio"/> tak, użytkownikiem <input type="radio"/> tak, członkiem gospodarstwa domowego z użytkownikiem <input checked="" type="radio"/> nie
25	Jak opisałby (opisałaby) Pan(i) swoją sytuację na rynku pracy w tygodniu od 25 do 31 marca 2011r.? (Proszę wybrać tylko jedną odpowiedź)	<input checked="" type="radio"/> pracowałem(am) wyłącznie poza rolnictwem <input type="radio"/> pracowałem(am) głównie poza rolnictwem i dodatkowo w rolnictwie <input type="radio"/> pracowałem(am) głównie w rolnictwie i dodatkowo poza rolnictwem <input type="radio"/> pracowałem(am) wyłącznie w rolnictwie <input type="radio"/> byłem(am) bezrobotny(a) <input type="radio"/> uczyłem(am) się, studiowałem(am) <input type="radio"/> byłem(am) na emeryturze, wcześniej/na emeryturze

- Some values can be chosen only from a predefined set of values. The idea is to take advantage of the meta-data environment. The questionnaire items are sometimes based on classification tables. Examples of such classifications are the Statistical Classification of Economic Activities (NACE), the Classification of Products by Activity (CPA), and a table of units for the questionnaire element. The figure below presents a possible solution of limiting the choice to the table containing the CPA nomenclature and table of units.

This is marked by an icon with a green downturned arrow. There is also a possibility to enter values by keying them, but in this case, if the values are not in the table an error message is triggered.

<input type="checkbox"/>	01	nazwa reprezentanta	<input type="text"/>			
	02	<input type="text"/>		<input type="text"/>		1 <input type="text"/>
	03	<input type="text"/>		<input type="text"/>		2 <input type="text"/>

- In longitudinal surveys, editing during data collection should account for relationships between current data and data from previous periods. The motive for doing so is to improve consistency and enable the respondent to pay attention to which data have been submitted previously. This can lead to lowering variability and avoidance of outliers. Another benefit of this approach is that the respondent is presented with values from earlier periods, which reduces the response burden. Whether historical data should be presented or not is a question that has not been clearly settled. A study by Holmberg (2002) indicates that presenting data from earlier periods has a positive effect. The study has not revealed undesired effects of repeating data from earlier rounds or underestimation. Holmberg advocates this approach in surveys with a high degree of data variability. The fear of conformity to and replication of previously reported data in current rounds was not confirmed. On the other hand, a study by Phillips et al. (1995) recommends a more conservative use of historical data. It stresses respondents' inclination to conform to information from previous survey rounds even if the presented data were spurious.
- In Haraldsen (2013), which is chapter 8 of a book devoted to designing questionnaires for business surveys, one can find useful information on how the technological aspects of business web questionnaires can assist shifting from presenting one and the same general approach to all respondents to a more personalised one. Information about a possible error or a request for a confirmation of data entered can result from the analysis of provided responses. A more active dialog can be based on data generated through processes of the questionnaire completion (paradata), registered behind the scenes. The figure beneath provides an example of attaching an icon with sign i (information) close to the field. By clicking on that icon the respondent is assisted with additional clarification about the item.

Additional information	
Please, provide a total estimated time for required data retrieval	<input type="text"/> 
Please, provide a total estimated time needed for this questionnaire completion	<input type="text"/> 

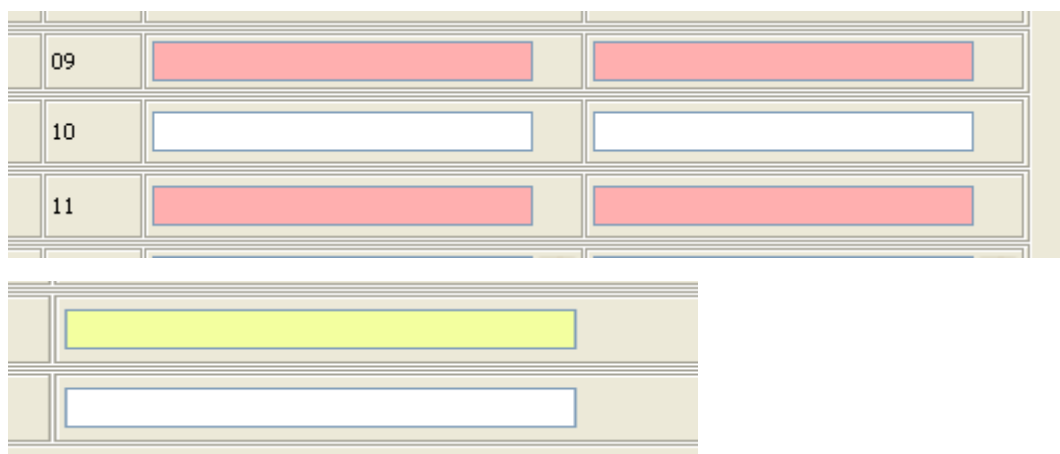
2.4 *Presentation of the messages to the respondent*

Implementation of edit rules into an electronic questionnaire poses a problem how to efficiently signify the error messages to users. The messages are to be recognised and comprehended. The visual side of a user interface includes a suite of elements such as graphics, colours, and fonts. Another aspect involves the phrasing of the error message. Usability test results suggest that words like “error” or “mistake” should be avoided because of their strong judgemental sense. Politeness and sensitivity towards the user is advisable as well as avoidance of jargon, e.g., computer terminology. The wording of error messages should be similar to that used in questions and associated with the subject of the survey. Error messages should be accompanied by the following attributes: item number, item topic and actual response (Murphy et al., 2001). In the interaction between a user and a computer program a message with a red icon signifies an error as a result of the execution or submission of an incorrect value. By analogy, a similar solution can be used in surveys. Another icon with an exclamation mark is used to signify soft errors. Beneath are examples of these icons:



Schonlau et al. (2002) advise placing the message as close as possible to the questionnaire item which it concerns. The message should be displayed either directly above or below the incorrect item. On the other hand, the study conducted by Mockovak (2005) showed no clear significance between different approaches to the placement of messages. Three kinds of solutions were tested. The first two involved placing the message above the item that triggered the edit and directly under that item, respectively. In these cases the error message was displayed after all the items on the page had been completed. In the final solution, the message was placed directly under the item and displayed as soon as the user left the field. The variation in placement and timing of the messages did not have a clear impact on noticing them. It also did not have a significant effect on the resolution of the problem indicated by the message or on following instructions contained in the message text, after the message had been noticed by the respondent. However, participants expressed clear preference for the message under the item.

The following examples present ways of marking erroneous fields using colours.





The image displays two examples of questionnaire forms. The top example is a table with three rows. The first row, labeled '09', has two red rectangular boxes. The second row, labeled '10', has two white rectangular boxes. The third row, labeled '11', has two red rectangular boxes. The bottom example shows a single form with two rows. The first row has a yellow rectangular box, and the second row has a white rectangular box. Both examples are set against a light beige background.



The two examples below are taken from a Polish reporting web portal and present marking the erroneous field by using icons.






roboczegodziny	roboczodni
<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>

<input type="text"/>	
----------------------	---

The following examples present solutions for displaying error messages (examples taken from a Polish reporting web portal). The first two figures provide messages close to the fields which have been marked as erroneous. Clicking on the item displays the message as a pop-up box. The last example presents a solution where a list of errors is gathered in a table exposed at the bottom of the web page.

<input type="text"/>			Błąd	
<input type="text"/>			Liczba strajków musi być wypełniona (czyli >0)	
001122				
wanka@gmail.com				

<input type="text"/>				
<input type="text"/>				Ostrzeżenie
<input type="text"/>				
001122				Imię i nazwisko osoby sporządzającej sprawozdanie powinny być podane.
wanka@gmail.com				

Lista błędów				
Lp.	Typ	Strona	Pole	Opis
1		1. Wstęp	lstrajk	Liczba strajków musi być wypełniona (czyli >0)
2		1. Wstęp	osoba	Imię i nazwisko osoby sporządzającej sprawozdanie powinny być podane
3		2. Karta statystyczna strajku	d1p2_1	[P2_1] poz.21 musi być wypełniona (czyli >0)
4		2. Karta statystyczna strajku	d1p7_2	[P7] poz.7 musi być wypełniona (tylko jedna z poz.71 lub poz.72)
5		2. Karta statystyczna strajku	d1p8_1	[P8_1] poz.81 musi być wypełniona (>0)

Timing of edit rules – The question is how the edit messages should be presented for their maximum effect. The possible solutions can be: present the message immediately after the field has been left, after the page was filled or at the end of the questionnaire entry. Immediate edits allow the respondent to correct the error straight away and can prevent similar mistakes later on (Skelterbery and Davies, 2012). From the other side, edits involving more than one variable raise the issue of waiting with edit execution for last variable completion. Whatever the case, usability studies point to the expectation of users that the form checking can be run iteratively. Another case is to prevent the user from entering some sort of keys, for example permitting only numeric keys. Programming formatting edits are examples of editing to prevent errors. This kind of edit checks should be triggered immediately. Edit rules should also be executed to reflect relationships between two or more variables. Such actions should be deferred, which requires additional functionality, where the user should be able to manually start an editing action as a batch operation. This, in turn, raises the question of whether the editing action should be triggered at the moment of completion or when the questionnaire is submitted over

the internet. Usability tests discourage this last solution, since performing actions that combine multiple functions is perceived as confusing (Anderson et al., 2005).

2.5 *Testing and evaluation of editing strategy*

Usability testing – Generally, usability testing results suggest the need for a good visual questionnaire design that uses fonts of different size and colour for questions and answers, can facilitate the answering process and reduces the completion time (Hansen and Couper, 2004). Though usability tests have their limitations, as they are conducted on a small number of users and try to test an entire questionnaire, not only the editing aspect, they can be a source for best practices for designing edit rules.

Analyses of collected data – Business surveys have a longitudinal nature. This grounds the possibility to evaluate the data collection instrument. A way to evaluate the set of built-in editing rules can be the number of non-response items. An issue when respondents tried to fit values to the upper bound of a range edit when it exceeded the range (Anderson et al., 2005) can be an example for tracking too rigorous edit rules in questionnaires.

User's centred design — Usability principles advocate the basic rule: user needs should be at the centre of the design. All tasks to be performed should be under the user's control. Throughout the response process, during the data entry stage, edit messages can appear several times and in various forms. The user needs to be able to choose the right moment to deal with them and to ensure the action taken is effective, which requires inter-connectivity between edit messages and the item back and forth as desired. The policy on how data with unresolved items submitted will be treated should be included in the instruction manual. It should be clear whether data marked as erroneous can be submitted. In other words, the question is whether strict conformity to edit rules should be required or rather whether users should be allowed more freedom in this matter, which will make them more likely to provide data, thus reducing the rate of non-response. The principle of emotional design (Norman, 1990) states that errors can result from various sources. This calls for a consistent design that accounts for the possibility of various errors. Another purpose of design is to counteract errors.

The burden – Incorporating edit rules into the questionnaire does not necessarily increase response burden (Anderson et al., 2005). Usability studies showed that some automatic checking of data entries are awaited by users to be performed by a computer. If the goal of edit checks is clearly understood by respondents the tolerance and acceptance for them can be easier gained. The limit for the scope of edits can be drawn from usability testing. The aim of edit rules is to improve data quality and not to encourage non-response.

Testing proposals – Skentelbery and Davies (2012) give good examples of testing online edits set-ups in their paper “Editing Challenges for New Data Collection Methods”. They bring up the research stating that for obtaining quality data the paging questionnaire design is the best option bearing in mind that two approaches are possible: paging and scrolling survey design.

2.6 *Electronic documents*

Typically, electronic processing involves implementing algorithms performed by a computer. An electronic questionnaire is simply a computer program. It seems useful to create a universal system, understood as a prototype program that could envelope a set of statistical variables and their validity

rules. In this way, variables and edit rules are combined, which gives shape to the definition layer of the output questionnaire. The questionnaire itself is designed to be a complete electronic document. This is why, a unified system combining the outer and inner part of the questionnaire should be created. The outer part refers to a computer program executed by the respondent, regardless of whether it is executed locally or remotely. The inner part, the core of the system, comprises the questionnaire definitions. The new technology supplies powerful tools that could be used to create such a unified solution. The extensible mark-up language seems to be well suited to the purpose of defining structural documents.

2.7 Conclusions

The goal of incorporating edits into the electronic questionnaire is to decrease measurement error in surveys. In the context of business surveys their unique features should be remembered when adopting a strategy for data collection editing. First, the response process is more burdensome than in social surveys. The data most commonly reside in business records and their retrieval requires time and effort. The response of a single unit may have an influential character. This is related to outliers. Some types of editing may require an aggregate level outlook. These features determine the types of edits used in questionnaires and also the scope of them. Data may be received with unresolved edit checks in order to avoid non-response. The compulsory requirement of edits resolving may be reserved for a “critical” set of items (Anderson et al., 2005). On the other hand the need for continuous evaluation of data collection instruments can be an opportunity for improvements. The crucial principle can be drawn from the usability principles that put the user control of the response process at the core of the design.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Anderson, A., Murphy, E., Nichols, E., Sigman, R., and Willimack, D. (2005), Designing interactive edits for U.S. electronic economic surveys and censuses: Issues and guidelines. Proceedings of UNECE Conference of European Statisticians, Ottawa, Canada, May 2005.
- Hansen, S., Couper, M. (2004), Usability Testing to Evaluate Computer-Assisted Instruments. In *Methods for Testing and Evaluating Survey Questionnaires*, Chapter 17, Wiley, New York.

- Haraldsen, G. (2013), Questionnaire Communication in Business Surveys. In Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D., *Designing and Conducting Business Surveys*, Chapter 8, John Wiley & Sons.
- Holmberg, A. (2002), Pre-printing Effects in Official Statistics, an Experimental Study. *International Conference on Questionnaire Development, Evaluation, and Testing Methods*, Charleston, SC.
- Mockovak, B. (2005), An evaluation of different design options for presenting edit messages in web forms. Bureau of Labour Statistics.
- Murphy, E., Nichols, E., Anderson, A., Harley, M., and Pressley, K. (2001), Building usability into electronic data-collection forms for economic censuses and surveys. *The Federal Economic Statistics Advisory Committee 2001 Conference*.
- Norman, D. (1990), *The Design of Everyday Things*. DoubleDay.
- Phillips, J. M., Mitra, A., Knapp, G., Simon, A., Temperly, S., and Lakner, E. (1995), The Determinants of Acquiescence to Preprinted Information on Survey Instruments. *Proceedings of the Survey Methods Research Section*, American Statistical Association, 1169–1171.
- Schonlau, M., Fricker, R. D., and Elliott, M. N. (2002), Guidelines for designing and implementing internet surveys (chapter five).
- Skentelbery, R. and Davies, C. (2012), Editing Challenges for New Data Collection Methods. Working Paper No. 18, Work Session on Statistical Data Editing, Oslo, Norway, 24-26 September 2012.
- de Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, Hoboken, New Jersey.

Interconnections with other modules

8. Related themes described in other modules

1. Response – Response Process
2. Statistical Data Editing – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

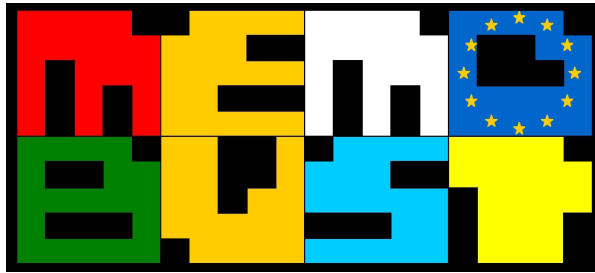
Questionnaire Design-T-Editing During Data Collection

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	13-03-2012	first version	Paweł Lańduch	GUS (Poland)
0.2	31-03-2013	second version	Paweł Lańduch	GUS (Poland)
0.3	10-12-2013	third version	Paweł Lańduch	GUS (Poland)
0.3.1	20-12-2013	preliminary release		
0.4	18-02-2014	version revised after EB review	Paweł Lańduch	GUS (Poland)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:27



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Testing the Questionnaire

Contents

General section.....	3
1. Summary	3
2. General description.....	4
2.1 Iterative itinerary	4
2.2 The response process model as a tool for testing the questionnaire	4
2.3 Pretesting	4
2.4 Field testing	6
2.5 Standards	8
2.6 Software.....	8
3. Design issues	13
4. Available software tools	13
5. Decision tree of methods	13
6. Glossary.....	13
7. References	14
Interconnections with other modules.....	16
Administrative section.....	17

General section

1. Summary

Establishment surveys differ from household surveys. This fact is reflected in the different culture of questionnaire development, evaluation and testing. First of all, the response process is more complex than in household surveys. The extensive adaptation of cognitive methods in social questionnaire development and testing based on Tourangeau's (1984) response process model, in the case of establishments, had to be enhanced by including new dimensions. The four step model, consisting of comprehension, retrieval, judgment and communication focuses on the individual. In the context of establishment surveys, however, the respondent is an informant selected within an organisation. Besides cognition, testing must also take into account the institutional frame, the need for cooperation, and the fact that required data items may be stored in business records. The non-existence of data in business records has to be taken into account as well.

Another consideration in the field of questionnaire testing is the specific nature of business surveys with their technical and intricate concepts and definitions; therefore, comprehension depends considerably on instructions. Closely connected with that and having its own consequences is the predominately self-administered data collection mode. There are other distinguishing features of the establishment population which pose a challenge for testing procedures. Among those are the longitudinal character of business surveys and the subsequent use of resulting data as inputs for other surveys. Yet another problem is a negative attitude to any changes in the questionnaires.

Nonetheless, the need for testing is beyond contention. This requirement is stated clearly in the Eurostat Code of Practice: "In the case of statistical surveys, questionnaires are systematically tested prior to the data collection". Ongoing data collection instruments are also under scrutiny. The goals to achieve include the improvement of the quality of statistical output, the reduction of costs to the surveying agency and to respondents, a decrease in the scope of output variables, an increase in the use of administrative data (Giesen, 2005). The adoption of Computer Assisted Interviewing has prompted redesign efforts to explore new prospects in data collection. This has added a new level of complexity to questionnaire testing. One new dimension is usability testing, which is intended to assess if the testing tool is user friendly and whether the interaction with the computer is intuitive and simple for the respondent. The optimal approach to efficient testing requires the involvement of end users. This leads to a paradox: in an effort to improve the collection instrument and ease the response burden, another burden is imposed on respondents (Willimack, 2005). Adding more burden during the response process, which in itself is burdensome, can hardly meet with the respondent's approval. It is, therefore, important that respondents should be aware of the goals of the testing procedure, which is intended to simplify and ease the response. When the aims of the procedure are clear, additional efforts can be received with a higher degree of approval. On the other hand, the iterative and longitudinal character of business surveys makes it possible to work out a systematic approach to improving the data collection instrument in a step-by-step procedure, which involves incorporating the testing and developing research into ongoing and repeated surveys. Instead of the usual practice of relying on post-collection activities to correct errors, a new paradigm is proposed, encouraging research on improving questionnaire design that leads to "error prevention rather than error correction" (Willimack et al., 2004).

A broad spectrum of recommended practices for developing and testing statistic questionnaires can be found in the Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System (2006). The Handbook adds valuable enhancement to the general subject of developing and testing questionnaires in statistical surveys. A detailed discussion of testing and evaluation questionnaires for establishment surveys can be found in Willimack (2013).

2. General description

2.1 Iterative itinerary

The development, testing and evaluation of questionnaires is part of a continuous process consisting of separate but linked stages along a continuum (Goldenberg et al., 2002). The process can be divided into 2 parts. The first part, including development and testing, and the second part, which comprises the assessment of measurement instruments after they have been used in the data collection process. The path goes through iterative steps, often going back and repeating the same cycle again. The starting point is to work out the survey goals. A draft questionnaire is a translation of concepts and definitions into questions and variables. Considering the precise and complex nature of concepts in establishment surveys the role of subject data experts cannot be neglected. The requirements and objectives of the survey should be combined with design aspects such as layout, technical constrains, instructions. A systematic approach requires guidelines in order to guarantee a consistent “look and feel” of the questionnaire. Draft versions have to be submitted to questionnaire design experts and subject data experts for reviewing. Data users should also play an active role in reviewing questionnaires because of the technical nature of the economic data and their stringency (Willimack et al., 2004). This stage is followed by the pretesting phase. Newly developed questionnaires require several rounds of pretesting. The pretesting process is limited by the costs and time. Beginning with internal staff testing, efficient testing should also involve end users. Findings from pretesting are the basis for further revisions. Ongoing surveys questionnaires can also be submitted for evaluation. The questionnaire may require a redesign to reduce costs, response burden, or change the data collection method. The testing and evaluation process should be based on the establishment response process model (Snijkers et al., 2005). The whole designing and redesigning process is iterative and open.

2.2 The response process model as a tool for testing the questionnaire

The module “Response – Response Process” provides a general discussion about response process models in business surveys. In this place the response process is referred only as the foundation for the cognitive approach to improve and evaluate questionnaires in statistical measurement. The knowledge of consecutive phases the respondent may go through starting from the decision to participate in a statistical survey to its successful completion can reduce the burden and, consequently diminish the measurement error. The response process in establishment surveys is more complex than in social surveys, which makes the matter even more important. Therefore, the response process can be a tool to evaluate the questionnaire and to find ways to improve it. An example of such an implementation can be found in a paper by Giesen (2007).

2.3 Pretesting

Pretesting involves applying testing techniques before the measuring instrument is used in the field in a survey operational stage.

2.3.1 Cognitive pretesting

Cognitive aspects of survey methodology (CASM) provided the basis for extending the interviewing process, typically understood as a way of eliciting answers to questionnaire questions, to a new field of exploration, namely to survey questionnaire testing, in an attempt to reduce measurement error. Cognitive interviewing focuses on the individual who becomes “the subject” of research and on the response thought process which drives the response (Willis, 1999). The general response process model by Tourangeau (1984), who developed its conceptual theoretical framework, consists of four steps: comprehension – associated with question understanding, retrieval – recalling the required information from memory, judgment – decision about the adequacy of the response, reporting – mapping the response to question categories. Cognitive interviewing relies on two methods: think aloud, where the subject is instructed to verbalise the process of arriving at an answer, and verbal probing, where, after having answered the question the subject is “probed” to get to the bottom of the response process. These methods were adapted to social surveys where the respondent means an individual. Cognitive pretesting in establishments surveys derives from the same methods but contains significant differences. First, the response process is much more complex. The four cognitive steps are extended by including additional steps unique to the establishment culture (Sudman et al., 2000). The next important aspect is that required data are contained in business managerial systems, that is in business records, not in a person’s memory. Such data can be dispersed in various parts of a company and gathering them requires cooperation between many persons. The largely quantitative nature of data is another characteristic feature. Building a protocol for cognitive pretesting should take into account all those aspects. Cognitive interviews take place at business locations rather than in laboratory conditions (Goldenberg et al., 2002; Willimack, 2008) owing to respondents’ unwillingness to participate in interviews outside their workplace. What matters here as well is access to records. The process of filling the questionnaire by the respondent during a cognitive interview is complemented by assessing how business records match the required data items. The interview also focuses on the timeline of data requirements and data availability (Goldenberg et al., 2002). The complex nature of the response in business surveys requires the expansion of cognitive interviewing (Freedman and Rutchik, 2002), which consists of pre-survey design visits and cognitive testing of the questionnaire at a business location. Pre-survey design visits should test data availability, record keeping practices, the compatibility between the time data are available and the time they are submitted to a statistical agency, the need for data confidentiality. A data model and a draft questionnaire are then cognitively pretested. “Think aloud” interview cognitively tests questions, instructions and concepts used. The informal unstructured part is used to discuss business records and the questionnaire itself.

2.3.2 Studies of business records

The first step in Tourangeau’s response process model (1984), that is *encoding in memory*, has been complemented in the business response model by *record formation* step (Edwards and Cantor, 1991; Sudman et al., 2000). The step stresses the fact that required data are contained in business systems. This affects the further steps of the response process, such as the selection of the proper respondent with access to data records and the knowledge of those records. Survey questions might be comprehensible but the required data may not be available in business records (Willimack, 2008). Record studies in companies, conducted as pre-survey design visits, can be a useful tool to collect information about the availability of requiring data, the compatibility of record keeping practices with

data collection instruments or the burden connected with retrieving those data from managerial systems (Murlow et al., 2007). Interviewing has a cognitive background. Interviews conducted in the four subsequent stages evolve from initially being focused on the overall goals of the survey, such as concepts and definitions or organisational aspects, to aspects directly connected with records. Valuable findings from such studies provide the knowledge about many aspects of record keeping in establishments. The study by Murlow et al., showed that different data are kept in different places of an establishment and that different people have different degrees of access to company data. Since the confidentiality of company data is a crucial aspect, it is easier to get general information without looking into details. The awareness of many aspects of the survey, Research and Development Survey as a result of the study by Murlow et al., helps to rebuild the questionnaire structure. The results proved to be worth the costs and efforts.

2.3.3 Usability testing

The usability theory stresses cognitive emotional aspects of the communication process and establishes principles of how to make things easier to use. While in the case of the paper questionnaire only visual improvements are possible, electronic questionnaires can be tested not only to assess their visual aspect, but also the user interface – the interconnection between the user and the computer program – and its functionality.

The theoretical background and empirical studies provide guidelines for visual elements of the questionnaire, such as consistence in the use of colours, fonts, spaces for questions and answers, answer options. Draft screens or their paper specimens submitted for assessment can develop design standards. Usability principles and the design of questionnaires are combined in the form of heuristic reviewing principles. A suite of usability tests addresses such aspects as navigation, skip routes, data entry or error correction. A product is usually subject to internal testing, before undergoing site testing by actual respondents. On-site testing, such as observation, can be connected with other cognitive testing methods. The Internet collection mode enables the application of software that records the response process in a real environment. System data logs can store information about user practices and the amount of time spent on the work with the questionnaire. This can provide additional knowledge about the response process.

2.4 Field testing

Field testing differs from pretesting in that it is applied to data collection instruments after they have been used in an operational stage (i.e., when the data have been collected). However, the term is sometimes used interchangeably with pretesting. Field testing can take the form of pilot tests that can be run before the data collection phase or after data collection is complete. What differentiates them from pretests is the greater number of respondents taking part in tests; as a result, the sample scope permits statistical inference; another difference is the iterative character of pretests as opposed to a one-off administration of a pilot test (Willimack et al., 2004). Post-collection questionnaire assessment is also referred to as questionnaire evaluation (QE).

2.4.1 Pilot tests

For a new or a redesigned survey questionnaire, a formal pilot test is the final step seen as a “dress rehearsal” of the measurement instrument (Goldenberg et al., 2002). During this test, all the steps of

the data collection process can be assessed. Leaving the respondent alone with a self-administered questionnaire, without the presence of an interviewer, mirrors the real environment of the response process. Evaluation of a redesigned instrument can be conducted by addressing the pilot questionnaire to a subsample of the target population and using the old form with the rest of the sample (Tuttle et al., 2010). In this particular study the results were gathered by debriefing respondents completing the pilot form and from additional questions designed to assess respondents' attitudes to the new questionnaire and by comparing it with the old one. A pilot questionnaire embedded in the final collection instrument is an opportunity to evaluate the proposed changes to the questionnaire (Willimack, 2008). This helps to avoid obstacles posed by traditional methods of improving questionnaires. For example, an additional question can be included to obtain an evaluation of how labour-intensive a questionnaire is (i.e., how much time is required to complete it).

2.4.2 *Debriefing respondents*

After the required data have been collected, the respondent is contacted one more time. The goal of such a contact is to acquire an assessment of the quality of the gathered data or to evaluate the questionnaire itself. Findings can help to improve the data collection instrument. There are formal and informal methods of conducting debriefings. Formal methods include *response analysis surveys* (RAS). Evaluations are conducted using structured questionnaires which contain questions about data sources and response strategies (Willimack et al., 2004). The feature that differentiates them from pretests is that contacted respondents are chosen from among original survey respondents and debriefings are done after data collection has been completed. This enables generalisations and in the case of ongoing surveys, allows future revisions (Goldenberg et al., 2002). The renewed contact can be performed in person, by telephone or by means of an evaluative questionnaire. Debriefings can also have an informal or ad hoc character. The purpose of a recontact is to get feedback from respondents on questions and questionnaire elements. Findings from a small sample of respondents can provide suggestions as to which elements of the questionnaire should be changed to address reported problems (Willimack et al., 2004). General issues concerning the response process can be addressed using unstructured interviews.

2.4.3 *Debriefing survey staff*

Survey operational personnel can be a valuable resource of knowledge about question-related problems. As intermediaries between respondents and the survey agency, offering help with questionnaire completion and recontacting respondents to resolve data item failures, they have the necessary knowledge to evaluate questionnaires and suggest potential revisions. Debriefing can be conducted both using formal methods, such as focus groups, or through informal methods (Goldenberg et al., 2002). One example of an informal method is observation of conversations between survey technical staff and respondents. Cognitive methods used in social surveys can be applied (Willimack et al., 2004). Data collection instruments can also be evaluated by data checking staff, who observe problems with data editing and who have direct contact with respondents. A cognitive debriefing session is a qualitative method of investigating problems with variables and reasons for problems (Hartwig, 2009). It involves an interview based upon a structured protocol which addresses the overall aspects of the survey and the questionnaire and its elements to improve the questionnaire or the design process. Unstructured discussions about experiences from field work can

also take place. At minimum costs, staff debriefings help to identify many problems and are the basis of long-term work on improving data collection instruments.

2.4.4 Post-collection data evaluation

The need for data comparability in long time series and their subsequent usefulness in constructing economic indicators limits the scope of changes that can be introduced in questionnaires and explains the general reluctance to any changes. Post collection data analysis is routinely conducted to assess data quality (Willimack et al., 2004). Measuring non-response to questionnaire data items, detecting outliers in collected data, the rate of imputation and examining data editing failures are various methods used to evaluate how the questionnaire works in the field (Goldenberg et al., 2002). A large number in non-response items may indicate problems with data availability. Data collected by a questionnaire in a survey can be compared with data from other sources to assess data consistency and quality. Data collection analysis is also useful in assessing whether changes made in questionnaires have improved the quality of collected data (Willimack, 2008). Questionnaire pretesting can lead to changes in questionnaires. Cognitive qualitative pretesting tries to improve questionnaire understanding and ease the burden imposed on respondents. By applying quantitative methods to data collected before and after the changes it is possible to find objective measures of how these changes have worked and whether there is any improvement in data quality. The non-response rate and data edit failures can be example of such measures.

2.5 Standards

Thorough questionnaire testing is a complex and burdensome process. The complex response process in establishment surveys makes the procedure even more challenging. Efficient testing requires the involvement of end users. By setting standards for this process one establishes goals that questionnaires must meet before they are actually used in data collection. DeMaio (2005) provides an example of levels of criteria a data collection instrument is expected to fulfil. The criteria are established for testing questionnaires as well as for other survey-related materials, such as introductory letters and supplemental instructions. The recommendations describe questionnaire requirements both for new surveys and redesigned measurement instruments in social and economic surveys and censuses. The minimal level is testing for proper administration of the questionnaire by an interviewer or by an end user and whether the questions are understandable. Additional recommendation extends requirements for data of great importance, and supplemental materials that are related to the survey. Self-administered electronic questionnaires further require testing that all the components of the software system behave properly and according to the design. Developing and applying guidelines and best practices for layout, visual design, field types, error handling can facilitate the testing process.

2.6 Software

Testing is the process of analysing software to detect differences between the expected and actual state, and to evaluate individual pieces of software. The content of this section was mainly founded on the Certified Tester Foundation Level Syllabus, which is aimed at anyone involved in software testing (ISTQB, 2010).

The main objective of testing is to find defects and errors in a questionnaire, software or documentation. Rigorous testing of software and documentation can reduce the risk of failure in

a production environment and contribute to a high quality product. It is necessary that any defects found in the process should be corrected before allowing the system to operate in a production environment.

- **Testing reveals errors**

Testing may indicate that there are defects, but we are not able to prove through testing that there are no errors. Testing reduces the probability that defects remain unidentified in the software, but even if no defects are found, it is no proof of software correctness.

- **Thorough testing is impossible**

Testing everything (all combinations of inputs and preconditions) is possible only for very trivial cases. By defining the scope of tests, instead of focusing on thorough testing, we focus on risks and priorities.

- **Early testing**

Testing procedures should start as early as possible in the software development cycle. They should also be geared to achieving the defined objectives.

- **Accumulation of errors**

Most defects found during testing prior to release or software failures revealed during production are located in a small number of modules.

- **Pesticide paradox**

If the same tests are repeated continuously on the same set of test cases, no new errors are found. This phenomenon involves the development of resistance to software testing (pesticide paradox¹). To overcome this paradox, test cases must be regularly reviewed and revised. In order to check other parts of software or system to potentially find more errors, one should use new tests.

- **Testing is context-sensitive**

Testing is done differently in different situations. For example, systems critical for safety are tested differently than e-commerce systems.

- **A false notion of correctness**

Finding and removing errors does not help if the system is not suitable for use and does not meet users' needs and expectations.

2.6.1 *The testing process*

The most visible part of testing is conducting tests. However, for tests to be efficient and effective, test planning should also take into account the time spent on test planning, designing test cases, preparing to perform tests and evaluating test execution status.

The basic test process consists of the following main steps:

¹ An analogous phenomenon – insects become resistant if one keeps applying the same insecticide.

- **Planning and supervision**

Test scheduling verifies the test mission, defines the objectives for testing and methods of achieving them. Test supervisions involves repeated comparison of actual testing progress and reporting with the plan and providing information about any deviations.

- **Analysis and design**

Test analysis and design are aimed at transforming general testing objectives into tangible test conditions and test designs.

- **Implementation and execution**

Test implementation and execution are a stage in which test conditions are transformed into test cases and test environment is created.

- **Assessment of the degree of completion and reporting**

At this stage, tests are evaluated in terms of predefined goals and exit criteria and the results are reported. It is specified whether more testing is needed or a change of termination conditions is necessary. Also the final report of the testing process is created.

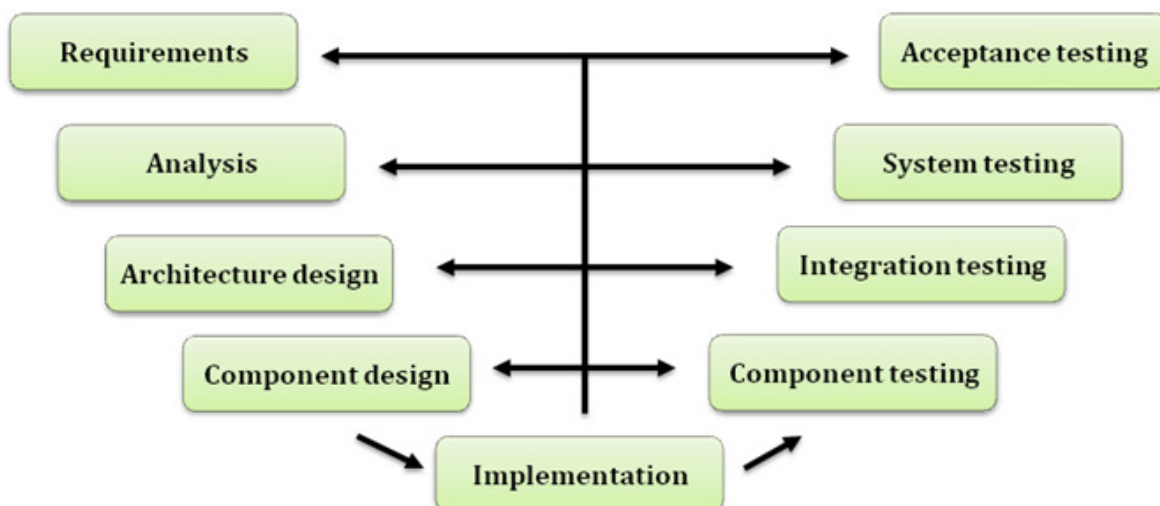
- **Test closure**

As part of closing the testing process, data are collected from completed test activities for future reference.

Although these activities are logically sequential, the process may overlap or occur simultaneously.

2.6.2 The testing phase in software life cycle.

Testing does not make sense in isolation from software development activities they are connected with. Typically, four levels of tests are identified, which correspond to four levels of software:



- **Component testing** (called unit test, module test) is a programming method for testing software by performing tests that verify the functionality of individual components (units) of the program - for example, methods or objects in object-oriented programming and procedures in procedural programming.

- **Integration testing** is performed to detect errors in the interfaces and interactions between modules (assembly testing). For example, we test communication between the module that stores and provides a set of parameters and a module that uses these parameters for initiation, for example, to fill form fields with default values.
- **System testing** is intended to determine whether an integrated system already meets the functional requirements and system requirements contained in the specification.
- **Acceptance testing** is not aimed at detecting errors but obtaining a formal confirmation of software quality.
- **Regression testing** is performed to ensure that the application works after modifications, error correction or expansion (new features). This kind of testing, due to its repeatability, lends itself to automation and can reveal previously undiscovered bugs.

2.6.2.1 Functional tests (black box)

Functional tests are based on functionalities and can be made at each level of testing. It is assumed that the tester doesn't know the structure of the program or its code. Their main features are:

- no prior knowledge of the application structure is required
- data are divided into classes of equivalence
- their purpose is to test the final functionalities

2.6.2.2 Structural tests (white box)

Structural testing can be performed at all levels of the test, but its main use is to test modules and module integration. Their task is to test those parts of the design which have not been tested by functional tests. They are based on the architecture of the application. Their main features are:

- knowledge of the application structure is required
- they cannot be used to reveal missing functionalities
- their purpose is to test the application structure

2.6.2.3 Non-functional tests

Non-functional tests include performance tests, load tests, stress tests, usability tests, cooperation tests, service tests, reliability tests and tests of the ability to work across different platforms. Tests of this type determine how the system works. Their main features are:

- they assume knowledge of the application configuration
- they require multiple test platforms
- they check performance

2.6.3 Automated software tests

2.6.3.1 What do automated tests deliver?

Automated tests are tests carried out with the help of specialised software. They are used to speed up the testing process, allowing you to generate test data and expected results, perform a set of tests with a final evaluation being positive or negative.

The advantages of automatic testing:

- efficient verification of bug fixes
- reuse of prepared tests
- quick reports
- comprehensive analysis of test results
- the use of large volumes of test data
- reduction in the cost of testing

Automatic testing can detect errors in the early stages of software development and protects against re-creation of the same error, which reduces the cost of creating questionnaires. Therefore, automated tests are performed at every stage of the project and invest in software testing.

Many tools are available on the market that make test preparation easier and faster. Presented below are the most popular ones.

2.6.3.2 Tools used for automated testing

Software testing varies depending on types of tests performed:

2.6.3.2.1 Functional testing

Functional tests are designed to test specific functionality including testing of the user interface.

They are used to simulate user behaviour and test questionnaire responses to these behaviours. These tests can be extremely useful at the early stages of the project. Automated tests of this type can be used as a key component of regression testing, especially in complex questionnaires, where manual testing of all functionalities is very time consuming and thus expensive. Functional tests must specify global standard methods of performing tasks such as filling out forms, login procedures etc. for future use in a single line of code that is not duplicated. Thus, for example, a change in the login procedure only requires a revision of a few lines in the log handler rather than making the same changes repeatedly at each test that required login.

Functional tests can be created either by authors of the questionnaire or target users, allowing quick and inexpensive testing under realistic conditions. Tools to enable this type of testing in the case of web questionnaires include **Selenium**, **Neoload** and **Rational Functional Tester**.

2.6.3.2.2 Unit testing (structural)

Unit tests are used to test individual modules of the questionnaire.

A program that performs unit tests verifies the accuracy of data input and output of the questionnaire software and the accuracy of the data processing method. It also checks integration between the modules themselves. It does not allow testing graphic elements of the questionnaire, it only serves to test the logical layer. Since unit tests focus on the logic of the questionnaire, they can only be created by programmers with high technical expertise and knowledge of the code. An example of a tool used to perform unit testing is **NUnit** for NET framework and its counterpart **JUnit** for Java and other tools of the **XUnit** family. In NUnit different classes or groups of classes are tested.

2.6.3.2.3 *Non-functional testing*

Non-functional testing includes, among others, load, performance, usability testing, and operations of the questionnaire on different platforms. These tests determine the hardware requirements and the help desk support necessary to prepare a questionnaire for action in difficult conditions, such as a very large number of users. They are expensive tests because they require experts in various fields (developers, service engineers, testers, users target). Tools to carry out non-functional tests include **NeoLoad**, **LoadRunner**, **HP LoadRunner**.

3. **Design issues**

In the context of questionnaire design one needs to mention systematic errors. While business surveys usually operate on a smaller number of variables than social surveys, their definitions tend to be complex and technical. Information in businesses data systems is organised to help companies achieve business goals but also to meet regulatory requirements. Therefore, their own definitions sometimes do not match those used by statisticians. In order to diminish problems resulting from these differences, we selectively mention two of the many design issues, when developing and testing questionnaires in business surveys:

- Using top-down approach together with bottom-up approach – the theory driven approach, with questions based on theoretical constructs, should be accompanied by explorations of data using by businesses;
- “Borrowing” questions from other surveys – in the light of the questionnaire testing, adapting a question from another survey may save the resources for additional testing, since the testing procedure has been already performed. The feasibility of “borrowing” depends on many factors of the survey design but determining it can be the first step in constructing a new survey or redesigning the current one.

4. **Available software tools**

Section 2.6.3.2 provides a couple of available software tools for various types of automatic software testing.

5. **Decision tree of methods**

Building the decision tree of methods can start with reviewing the contingency table of methods and steps in questionnaire developing, testing and evaluation. A paper by Goldenberg et al. (2002) gives an example of such a table where rows contain various methods used by statistical organisations and columns correspond to the steps specified by the method for questionnaire development, testing and evaluation (QDET).

6. **Glossary**

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Brancato, G., Macchia, S., Murgia, M., Signore, M., Simeoni G., - Italian National Institute of Statistics, ISTAT, Blanke, K., Körner, T., Nimmergut, A., - Federal Statistical Office Germany, FSO, Lima, P., Paulino, R., - National Statistical Institute of Portugal, INE, Hoffmeyer-Zlotnik, J. H. P., - German Center for Survey Research and Methodology, ZUMA (2006), *Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System*.
- Edwards, W. S. and Cantor, D. (1991), Toward a Response Model in Establishment Surveys. In: P. P. Biemer et al. (eds.), *Measurement Error in Surveys*, John Wiley & Sons, New York, 211–233.
- Freedman, S. and Rutchik, R. (2002), Establishments as Respondents: Is Conventional Cognitive Interviewing Enough? *Presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, SC*.
- Giesen, D. (2005), Results of the Evaluation and Redesign of the Dutch Structural Business Statistics Questionnaires. *Proceedings of the 5th QUEST Workshop, 19-21 April 2005*, Statistics Netherlands, Heerlen.
- Giesen, D. (2007), The Response Process Model as a Tool for Evaluating Business Surveys. *Proceedings of the Third International Conference on Establishment Surveys (ICES-3), 18-21 June, Montreal, Canada*, American Statistical Association, Alexandria, VA, 871–880.
- Goldenberg, K. L., Anderson, A. E., Willimack, D. K., Freedman, S. R., Rutchik, R. H., and Moy, L. M. (2002), Experiences Implementing Establishment Survey Questionnaire Development and Testing at Selected U.S. Government Agencies. *Presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, S.C., November, 2002*.
- Hartwig, P. (2009), How to use edit staff debriefings in questionnaire design. *Presented at the European Establishment Statistics Workshop - EESW09*.
- ISTQB (2010) International Software Testing Qualifications Board (ISTQB®) Certified Tester Foundation Level Syllabus, Version 2010 by Thomas Müller (chair), Armin Beer, Martin Klönk, Rahul Verma.
- DeMaio, T. (2005), Standards for Pretesting Questionnaires and Survey Related Materials for U.S. Census Bureau Surveys and Censuses.
- Murrow, J. M., Freedman, S., and Rutchik, R. (2007), Record Keeping Studies – Love 'Em or Leave 'Em. *Paper presented at the Third International Conference on Establishment Surveys - ICES-III, June 18-21, 2007, Montreal, Quebec, Canada*.
- Snijders, G., Onat, E., and Tonglet, J. (2005), The Dutch Annual Business Inquiry: Developing and testing the electronic form. *Proceedings of the 5th QUEST Workshop, 19-21 April 2005*, Statistics Netherlands, Heerlen.
- Sudman, S., Willimack, D. K., Nichols, E., and Mesenbourg Jr., T. L., (2000), Exploratory Research at the U.S. Census Bureau on the Survey Response Process in Large Companies. *Proceedings of*

- the 2nd International Conference on Establishment Surveys*, American Statistical Association, Alexandria, VA, 327–335.
- Tourangeau, R. (1984), Cognitive science and survey methods: a cognitive perspective. In: T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*. National Academy Press, Washington, DC.
- Tuttle, A. D., Morrison, R. L., and Willimack, D. K. (2007), From Start to Pilot: A Multi-Method Approach to the Comprehensive Redesign of an Economic Survey Questionnaire. *Presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.*
- Willimack, D. K. (2005), The Paradox of Respondent Burden in Testing Electronic Instruments for Establishment Surveys. *Proceedings of the 5th QUEST Workshop, 19-21 April 2005*, Statistics Netherlands, Heerlen.
- Willimack, D. K. (2008), Issues in the Design and Testing of Business Survey Questionnaires: What We Know Now that We Didn't Know Then – and What We Still Don't Know. *Proceedings of the Conference on Reshaping Official Statistics*, International Association on Official Statistics, Shanghai, China.
- Willimack, D. K., Lyberg, L., Martin, J., Japac, L., and Whitridge, P. (2004), Evolution and Adaptation of Questionnaire Development, Evaluation, and Testing Methods for Establishment Surveys. In: S. Presser et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*, Wiley, New York, Chapter 19.
- Willimack, D. (2013), Methods for the Development, Testing, and Evaluation of Data Collection Instruments. In: G. Snijders, G. Haraldsen, J. Jones, and D. Willimack (eds.), *Designing and Conducting Business Surveys*, Wiley, Hoboken, New Jersey, 253–301.
- Willis, G. (1999), Cognitive interviewing: a “how to” guide. Research Triangle Institute, <http://appliedresearch.cancer.gov/areas/cognitive/interview.pdf>.

Interconnections with other modules

8. Related themes described in other modules

1. Response – Response Process

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

Questionnaire Design-T-Testing the Questionnaire

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	23-02-2012	first version	Magdalena Homenko	GUS (Poland)
0.2	29-05-2013	major revisions (reviews from Netherlands and Sweden)	Magdalena Homenko Paweł Lańduch	GUS (Poland)
0.3	30-09-2013	minor revisions (reviews from Netherlands and Sweden)	Magdalena Homenko Paweł Lańduch	GUS (Poland)
0.4	30-10-2013	minor revisions (Dutch review)	Magdalena Homenko Paweł Lańduch	GUS (Poland)
0.4.1	18-11-2013	preliminary release		
0.4.2	11-03-2014	minor revisions (EB review)	Paweł Lańduch	GUS (Poland)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:27



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Data Editing – Main Module

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction to statistical data editing	3
2.2 Types of errors.....	5
2.3 Edit rules.....	6
2.4 Overview of methods for statistical data editing	7
3. Design issues	10
4. Available software tools.....	10
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	12
Administrative section.....	13

General section

1. Summary

Data that have been collected by a statistical institute inevitably contain errors. In order to produce statistical output of sufficient quality, it is important to detect and treat these errors, at least insofar as they have an appreciable influence on publication figures. For this reason, statistical institutes carry out an extensive process of checking the data and performing amendments. This process of improving the data quality for statistical purposes, by detecting and treating errors, is referred to as statistical data editing.

2. General description

2.1 Introduction to statistical data editing

Errors are virtually always present in the data files used by producers of statistics. This is true for both data obtained by means of surveys and data originating from external registers. Insofar as these errors result in inaccurate estimates of publication figures, it is important for statistical institutes to detect and treat these errors.

Errors can arise during the measurement process; if this is the case, there will be a difference between the reported value and the actual value. This can occur because the respondent does not know the actual value exactly or at all, or has difficulty finding this value and therefore makes an estimate. Another possible cause is a difference in definitions between the accounting records of businesses and the statistical institute, for example because the financial year differs from the calendar year. Furthermore, it is possible that businesses simply do not have all the information requested by the statistical institute on file. In this case, the respondent will again estimate certain values or not answer all questions. Finally, respondents may also read or understand questions incorrectly. For example, they may report in euros, while they were actually asked to report in thousands of euros (this is an example of a so-called *unit of measurement error*).

Errors may also arise during data processing. At a statistical institute, the collected data typically go through different processes, such as entering, coding, detection, imputation, weighting, and tabulation. All of these processes can introduce errors into the data. An example of this is that the manual entry of data can result in misinterpretations, for example, a '1' is taken for a '7' or vice versa. Similar mistakes can occur when optical character recognition is used to process survey forms automatically. Additionally, there may be errors in the processing software, and good values may incorrectly be seen as errors during the editing process.

The process of detecting and treating errors in a data file to be used for statistical purposes is called *statistical data editing*. Other commonly used terms are *data validation* and *data cleaning*. In traditional survey processing, data editing was mainly a manual activity, intended to check and correct all data items in every detail. Inconsistencies in the data were investigated and, if necessary, adjusted by subject-matter experts, who would consult the original questionnaires or recontact respondents to verify suspicious values. Overall, this was a very time-consuming and labour-intensive procedure. According to estimates in the literature, statistical institutes would spend up to 25% or 40% of their total budget on data editing (Federal Committee on Statistical Methodology, 1990; Granquist, 1995; Granquist and Kovar, 1997).

According to Granquist (1997), statistical data editing should have the following objectives, in descending order of priority:

1. To identify possible sources of errors so that the statistical process can be improved in the future;
2. To provide information about the quality of the data collected and published;
3. To detect and correct influential errors in the collected data.

In EDIMBUS (2007), a fourth objective is added:

4. If necessary, to provide complete and consistent microdata.

In line with the first objective mentioned above, the main aim of recontacts with respondents should not be to merely resolve individual observed errors, but rather to collect information on the causes of these errors. By collecting and analysing this information, a statistical institute has the opportunity to identify potential measures for improving the quality of incoming data in the future. Examples of such measures include improving the design of the questionnaire and, in particular, changing the wording of a question that many respondents found difficult to answer. In the words of Granquist (1997), “editing should highlight, not conceal, serious problems in the survey vehicle.”

Currently at most statistical institutes, statistical data editing is used primarily with the third and fourth of the above goals in mind: correcting errors that have a significant influence on publication totals and providing complete and consistent data. Although it is widely acknowledged in the data editing literature that the information obtained during editing could and should also be used to improve aspects of the statistical process for a repeated survey, the development of practices to achieve this goal still appears to be a rather neglected area. Some statistical institutes have had good experiences with standardised debriefings of editing staff as a device for identifying possible improvements in questionnaire design (Rowlands et al., 2002; Hartwig, 2009; Svensson, 2012). An overview of indicators for assessing the quality of the data before and after editing is given in EDIMBUS (2007).

Over the past decades, statistical institutes have recognised that it is usually not necessary to correct all data in every detail. Several studies have shown that reliable estimates of publication totals can also be obtained without removing all errors from a data set (see, e.g., Granquist, 1997, and Granquist and Kovar, 1997). The main output of most statistical processes consists of tables of aggregated data, which are often estimated from a sample of the population. Hence, small errors in individual records can be accepted, provided that (a) these errors mostly cancel out when aggregated, and (b) insofar as they do not cancel out when aggregated, the resulting measurement error in the estimate is small compared to the total error – in particular the natural variation in the estimate due to sampling.

The notion that not all errors need to be corrected in every detail has led to the development of more efficient editing approaches: in particular selective editing, automatic editing and macro-editing. Section 2.4 introduces these approaches, and also illustrates how they may be combined into an effective data editing process. Before that, we discuss different types of errors in Section 2.2 and edit rules in Section 2.3.

We refer to De Waal et al. (2011) and EDIMBUS (2007) for a more comprehensive description of statistical data editing.

2.2 Types of errors

Different editing methods have been developed for different types of errors. We will consider here the distinction between influential and non-influential errors and the distinction between systematic and random errors.

Influential errors include the errors that have a significant influence on the final publication total. An error can be influential because it was made by a business that naturally has a strong influence on the estimate, i.e., either by a large business or by a smaller one with a large sampling weight. In addition, sometimes an error is so large that it will strongly influence the total, regardless of the size of the business for which the error occurred. A notorious example of a type of error that is usually influential is the above-mentioned unit of measurement error.

It is clear that errors that have a large influence on a publication total can lead to significant bias. For this reason, it is crucial to treat these errors as effectively as possible. An efficient and timely data editing process will have to focus mainly on the detection and treatment of influential errors. The distinction between influential and non-influential errors is particularly useful in business surveys, because these often contain variables with a skew distribution in the population, such as *Turnover*.

Another distinction that is often made is that between *systematic* and *random* errors.¹ These terms do not have universally accepted definitions. In particular, UN/ECE (2000) defines a systematic error as “an error reported consistently over time and/or between responding units”, while EDIMBUS (2007) defines it as “a type of error for which the error mechanism and the imputation procedure are known.” The first definition refers in particular to errors that are caused by persistent response problems, which are ‘not random’ in the sense that they would likely be observed again if the data collection process were repeated. Examples include: the unit of measurement error mentioned in Section 2.1; different definitions used by the statistical institute and the respondent (e.g., gross turnover versus net turnover); persistent problems with data entry or coding at the statistical office. The second definition focuses on the fact that, in many cases, errors of this kind are relatively easy to detect, precisely because they are made in a consistent way. Thus, in many cases, these two definitions of systematic errors agree. In practice, the only systematic errors that can be treated as such are those for which the error mechanism is understood, i.e., errors that are systematic according to the definition of EDIMBUS (2007).

Although the above definitions of systematic errors do not mention bias, it does hold that systematic errors often produce a systematic bias in estimated figures. This is true because these errors are often made in the same way by several respondents. For random errors – i.e., errors that are not systematic as defined in the previous paragraph – the risk of a bias is smaller. On the other hand, random errors are more difficult to detect and correct reliably, precisely because little is known about the underlying causes.

It should be noted that systematic errors may or may not be influential. For instance: the unit of measurement error is usually influential, but an error where a small business with a moderate sampling

¹ Here, the terms ‘systematic’ and ‘random’ are supposed to refer to the mechanism that *causes* an error. This differs from the use of these terms in measurement error models, where they refer to the *effect* of an error on an estimator (an error being systematic to the extent that it introduces bias and random to the extent that it introduces noise). As explained in the main text, these two meanings of ‘systematic’ do overlap to some extent.

weight reports gross turnover instead of net turnover will usually be non-influential. The same holds for random errors.

2.3 Edit rules

To detect errors in observed data, *edit rules* are widely used. These are rules that indicate conditions that should be satisfied by the values of single variables or combinations of variables in a record. Edit rules are also commonly known as *edits* or *checking rules*. If a record does not satisfy the condition specified by an edit rule, the edit rule is said to be failed by that record. Inspection of data items that fail an edit rule is an important technique for finding errors in a data file.

A conceptual distinction should be made between so-called *hard* and *soft* edit rules. Hard edit rules (also known as *fatal* edit rules or *logical* edit rules) are edit rules that must hold by definition, such as

$$\text{Turnover} = \text{Profit} + \text{Costs}.$$

If a hard edit rule is failed by an observed combination of values, then it is certain that at least one of those values contains an error. Soft edit rules (also known as *query* edit rules) indicate whether a value, or value combination, is suspicious. For instance, the soft edit rule

$$\text{Profit} / \text{Turnover} \leq 0.6$$

states that it is unusual for the value of *Profit* to be higher than 60% of the value of *Turnover*. In contrast to hard edit rules, soft edit rules can be failed by unlikely values that are in fact correct. Thus, soft edit failures should trigger a closer investigation of the data items involved, to assess whether the suspicious values are erroneous or merely unusual.

Typically, business surveys involve (mainly) numerical data. For this type of data, some commonly encountered classes of edit rules include the following:

- *Univariate edits / Range restrictions.* These edit rules restrict the range of admissible values for a single variable. A common example is the restriction that a numerical variable may attain only non-negative values, e.g., the edit rule “ $\text{Turnover} \geq 0$ ”. Depending on the context, edits of this type can be either hard or soft.
- *Ratio edits.* These edit rules are bivariate restrictions taking the general form $a \leq x / y \leq b$, where x and y are numerical variables and a and b are constants. An example could be that the ratio of *Turnover* and *Number of Employees* (i.e., the average contribution of one employee to the total turnover of a business) should be between certain bounds. The above-mentioned edit rule “ $\text{Profit} / \text{Turnover} \leq 0.6$ ” is another example of a ratio edit. As the latter example illustrates, some ratio edits contain only a lower bound a or an upper bound b , but not both. Typically, ratio edits are soft edit rules.
- *Balance edits.* These edit rules are multivariate restrictions that relate a set of variables through a linear equality. The above-mentioned edit rule “ $\text{Turnover} = \text{Profit} + \text{Costs}$ ” is an example of a balance edit. The general form of a balance edit is: $a_1x_1 + \dots + a_nx_n + b = 0$, where x_1, \dots, x_n are numerical variables and a_1, \dots, a_n, b are constants. Usually, but not always, balance edits are hard edit rules.

2.4 Overview of methods for statistical data editing

The data editing process that is considered here starts after the data have been collected and entered. It should be noted, however, that nowadays many business surveys use computer-assisted modes of data collection (see the topic “Data Collection”) which often involve electronic questionnaires. With computer-assisted data collection, it is possible to perform part of the editing already at the data collection stage, for instance by building certain edit rules into the electronic questionnaire. We refer to the theme module “Questionnaire Design – Editing During Data Collection” for a discussion of the possibilities.

The specific way that the data editing process is structured will vary by statistic and by statistical institute. However, there is a general strategy that is followed in broad lines in many processes. This general strategy is shown in Figure 1; similar strategies are discussed in De Waal et al. (2011, pp. 17-21) and EDIMBUS (2007, pp. 6-8). It consists of five steps:

1. Deductive editing;
2. Selective editing;
3. Automatic editing;
4. Interactive editing (manual editing);
5. Macro-editing.

In the remainder of this section, we give a brief outline of each of these steps. More detailed descriptions can be found in the accompanying modules on methods for statistical data editing.

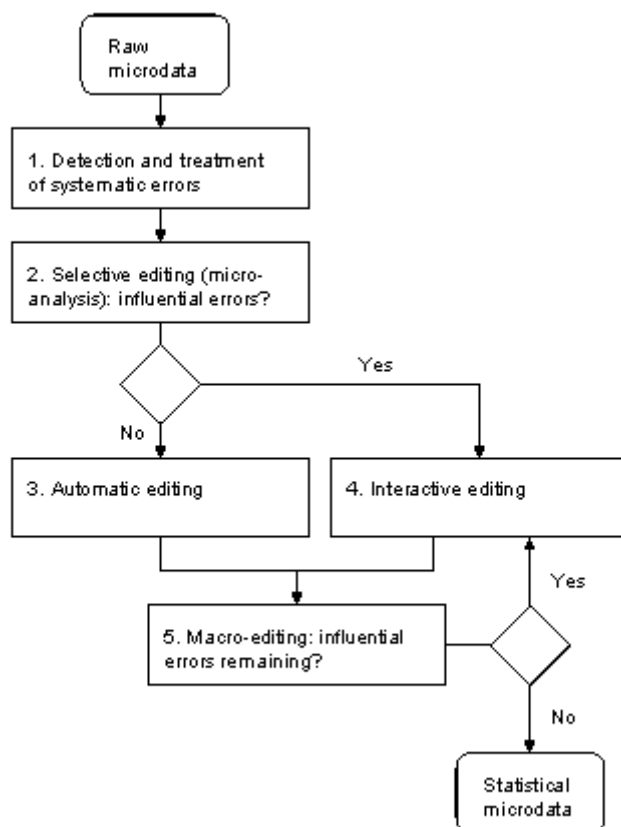


Figure 1. Example of a data editing process flow

In the first phase of the data editing process, identifiable systematic errors are detected and treated. As stated in Section 2.2, these systematic errors can lead to significant bias. Moreover, these errors can often be automatically detected and treated easily and very reliably. It is highly efficient to treat these errors at an early stage. In the remainder of the data editing process, it may then be assumed that the data contain only random errors. The detection and treatment of systematic errors is discussed in the method module “Statistical Data Editing – Deductive Editing”.

After the identifiable systematic errors have been edited automatically, a decision can be taken to begin *manual editing*, i.e., manual detection and treatment of errors. This process step is performed by editors or analysts who are usually supported in this regard by software that allows, for example, edit rules to be applied to the data and values to be changed interactively. This form of editing (also known as *interactive editing*) is described in the method module “Statistical Data Editing – Manual Editing”.

As mentioned above, manual editing is usually expensive and time-consuming. It is therefore better to restrict the manual work only to records that likely contain influential errors, so that the specialists’ limited time can be used where it is most effective. The other records, with less important errors, can either be left unedited or, alternatively, be edited automatically (see below). Limiting interactive editing to those records that likely contain influential errors which cannot be reliably resolved automatically is known as *selective editing* or *micro-selection*. Methods that can be used in this step are discussed in the theme module “Statistical Data Editing – Selective Editing”. It should be noted that the selective editing step by itself does not treat any errors; it merely assigns records to different forms of further treatment.

Most selective editing methods make use of anticipated values for the variables in a record to identify the most suspicious values in the observed data. Observed values that deviate strongly from the anticipated values may be caused by influential errors. In determining the anticipated values, information is used from sources other than the actual data file. Oftentimes, edited data from a previous period for the same statistic is used for this purpose. As such, selective editing can proceed on a record-by-record basis, and hence it is possible to start the selection process for manual editing during the data collection period, as soon as the first records are received. This is in fact the main advantage of selective editing over macro-editing, a different selection method to be discussed below.

Records that are not selected for manual editing can be processed by *automatic editing* instead. The automatic treatment of random errors and other errors for which the cause cannot be established usually takes place in two steps. First, the best possible determination is made of what values in a record are incorrect. This is trivial if a value does not fall in the permissible range according to a univariate edit, such as a negative number of employees or an improperly missing value. As such, the value is then certainly incorrect. In many cases, however, inconsistencies can occur for which it is not immediately clear which value or values are responsible. If, for example, the hard balance edit

$$\text{Total Costs} = \text{Personnel Costs} + \text{Capital Costs} + \text{Transport Costs} + \text{Other Costs}$$

is not satisfied, then it is clear that (at least) one of the reported values must be erroneous, but it is usually not obvious which one. The problem of identifying the erroneous values in an inconsistent record is known as the *error localisation problem*.

In automatic editing of business survey data, the error localisation problem for random errors is usually solved by applying the *Fellegi-Holt paradigm*, which states: a record should be made

consistent by changing the fewest possible items of data (Fellegi and Holt, 1976). Methods for automatic error localisation based on the Fellegi-Holt paradigm are discussed in the method module “Statistical Data Editing – Automatic Editing”.

Once the erroneous values have been detected, they are replaced with better values by means of *imputation*. Automatic imputation relies on (explicit or implicit) mathematical models that use information from the correctly observed values to predict the values that were incorrectly observed or missing. We refer to the topic “Imputation” for a discussion of this subject.

Instead of applying automatic editing, one may also choose not to edit the records that are not selected for interactive treatment by the selective editing procedure. In fact, one may argue that it is not necessary to edit these records, because they will not contain any influential errors, assuming that the selective editing procedure works as intended. Nevertheless, there are reasons why automatic editing may be of use in practice (see also De Waal and Scholtus, 2011). Firstly, it is often desirable to resolve at least all obvious inconsistencies (values that fail hard edit rules), even when these are not influential as such. This is especially true if the microdata are to be released to external users. Secondly, automatic editing provides a relatively inexpensive way to test the quality of a selective editing procedure. If the selection procedure is working correctly, then the records that are not selected for interactive treatment should require only minor adjustments with little influence on a publication figure. Thus, if many influential adjustments are made during automatic editing, this may indicate that the design of the selective editing procedure needs to be improved.

In the final phase of the process in Figure 1, provisional publication figures are calculated and analysed using historical data or external sources. This analysis is called *macro-editing* or *output editing*. If the aggregate figures are implausible, the underlying individual records are examined by, for example, further analysing outliers or influential records and adjusting these as necessary. In Figure 1, this is indicated by the arrow leading back from macro-editing to interactive editing. The errors detected at this stage may be errors that were not found in earlier phases of the data editing process or errors that were actually introduced by the process. In macro-editing, the detection of errors begins at an aggregated level, but the adjustment always takes place in the underlying microdata, i.e., the records of individual respondents. As soon as the provisional figures are considered plausible, the statistical data editing process is completed. For more information on this step, see the module “Statistical Data Editing – Macro-Editing”.

In the macro-editing step, as well as during selective editing and manual editing, mathematical techniques for outlier detection are often applied. An extensive discussion of outlier detection in the context of statistical data editing can be found in EDIMBUS (2007).

The process in Figure 1 should be viewed as a prototype. In practice, not all of the steps will be undertaken for all statistics, or a different order of process steps may be used. For instance, it was already mentioned that automatic editing is not always included in the process. Another example is that the selection of records for manual editing is often partly based on other criteria than only whether a record contains influential errors. As such, important or complex businesses are frequently identified as crucial, meaning that their data are always inspected manually. Examples of such businesses could be those that are individually responsible for a significant portion of turnover in their sector. See, e.g., Pannekoek et al. (2013) for a further discussion of the design of an editing process.

Many business surveys have a longitudinal aspect. Sometimes, a panel of units is followed over time during multiple rounds of the same survey. Even for cross-sectional business surveys, the largest units in the population are usually observed in each survey round. This implies that during a particular survey round, at least for part of the responding units, historical data are available. These historical data may be used in various ways during several steps of the editing process; for example, they are often used to determine anticipated values for selective editing. We refer to the theme module “Statistical Data Editing – Editing for Longitudinal Data” for more details on this aspect of statistical data editing.

Finally, it should be noted that, traditionally, applications of statistical data editing have been aimed mainly at survey data. More recently, the use of administrative data for statistical purposes has become increasingly important. These data require an editing process that is in some respects different from the typical editing process for survey data. For instance, for statistics based on administrative data, often all the data (or a large proportion thereof) become available at the same time. In that case, it is not necessary to use micro-selection methods, and we can start immediately with output editing. We refer to the theme module “Statistical Data Editing – Editing Administrative Data” for a discussion of editing in the context of statistics based on administrative data.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.

De Waal, T. and Scholtus, S. (2011), Methods for Automatic Statistical Data Editing. Paper presented at the 2011 KSS International Conference on Statistics and Probability, Busan.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.

http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Federal Committee on Statistical Methodology (1990), *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18, U.S. Office of Management and Budget, Washington, D.C.

- Fellegi, I. P. and Holt, D. (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.
- Granquist, L. (1995), Improving the Traditional Editing Process. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, 385–401.
- Granquist, L. (1997), The New View on Editing. *International Statistical Review* **65**, 381–387.
- Granquist, L. and Kovar, J. (1997), Editing of Survey Data: How Much Is Enough? In: L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, 415–435.
- Hartwig, P. (2009), How to Use Edit Staff Debriefings in Questionnaire Design. Paper presented at the 2009 European Establishment Statistics Workshop, Stockholm.
- Pannekoek, J., Scholtus, S., and van der Loo, M. (2013), Automated and Manual Data Editing: A View on Process Design and Methodology. *Journal of Official Statistics* **29**, 511–537.
- Rowlands, O., Eldridge, J., and Williams, S. (2002), Expert Review Followed by Interviews with Editing Staff – Effective First Steps in the Testing Process for Business Surveys. Paper presented at the 2002 International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, South Carolina.
- Svensson, J. (2012), Editing Staff Debriefings at Statistics Sweden. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- UN/ECE (2000), *Glossary of Terms on Statistical Data Editing*. United Nations, Geneva.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Editing During Data Collection
2. Data Collection – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Statistical Data Editing – Editing Administrative Data
6. Statistical Data Editing – Editing for Longitudinal Data
7. Imputation – Main Module

9. Methods explicitly referred to in this module

1. Statistical Data Editing – Deductive Editing
2. Statistical Data Editing – Automatic Editing
3. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Statistical data editing

Administrative section

14. Module code

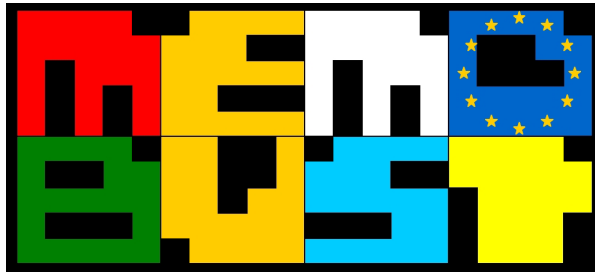
Statistical Data Editing-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	09-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	20-04-2012	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	19-06-2012	minor improvements	Sander Scholtus	CBS (Netherlands)
0.3.1	16-07-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.4	31-10-2013	minor improvements based on comments by Italian reviewer and Editorial Board	Sander Scholtus	CBS (Netherlands)
0.4.1	31-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:10



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Deductive Editing

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Introduction to deductive editing.....	3
2.2 Correction rules for subject-matter related errors.....	4
2.3 The unit of measurement error	5
2.4 Identifying new systematic errors	8
3. Preparatory phase	9
4. Examples – not tool specific.....	9
4.1 Example: Correction rules for the statistic Building Objects in Preparation.....	9
4.2 Example: Simple typing errors.....	10
5. Examples – tool specific.....	11
6. Glossary.....	13
7. References	13
Specific section.....	15
Interconnections with other modules.....	16
Administrative section.....	18

General section

1. Summary

Data collected for compiling statistics frequently contain obvious systematic errors; in other words, errors that are made by multiple respondents in the same, identifiable way (see “Statistical Data Editing – Main Module”). Such a systematic error can often be detected automatically in a simple manner, in particular in comparison to the complex algorithms that are needed for the automatic localisation of random errors (see the method module “Statistical Data Editing – Automatic Editing”). Furthermore, after a systematic error has been detected, it should be immediately clear which adjustment is necessary to resolve it. For we know, or think we know with sufficient reliability, how the error came about.

A separate deductive method is needed for each type of systematic error. The exact form of the deductive method varies per type of error; there is no standard formula. The difficulty with using this method lies mainly in determining *which* systematic errors will be present in the data, before these data are actually collected. This can be studied based on similar data from the past. Sometimes, such an investigation can bring systematic errors to light that have arisen due to a shortcoming in the questionnaire design or a bug in the processing procedure. In that case, the questionnaire and/or the procedure should be adapted. To limit the occurrence of discontinuities in a published time series, it can be desirable to ‘save up’ changes in the questionnaire until a planned redesign of the statistic, and to treat the systematic error with a deductive editing method until that time.

2. General description of the method

2.1 Introduction to deductive editing

In this module, we focus on methods for detecting and treating so-called systematic errors. As mentioned in “Statistical Data Editing – Main Module”, a systematic error is commonly defined as an error with a structural cause that occurs frequently between responding units. A well-known type of systematic error is the so-called *unit of measurement error* which is the error of, for example, reporting financial amounts in units instead of the requested thousands of units.

Systematic errors can introduce substantial bias in aggregates, but once detected, systematic errors can easily be treated because the underlying error mechanism is known. It is precisely this knowledge of the underlying cause that makes the treatment of systematic errors different from random errors. Treating systematic errors based on knowledge of the underlying error mechanism is called *deductive editing*. Systematic errors can often be identified by examining frequently occurring edit rule failures. Deductive methods are therefore mainly effective for data for which many edit rules have been defined.

Deductive editing of systematic errors is an important first step in the editing process. It can be done automatically and reliably at virtually no costs. Moreover, the rest of the editing process can proceed more efficiently after the systematic errors have been resolved. Deductive editing is in fact a very effective and probably often underused editing approach.

Any systematic error for which the cause is understood with sufficient certainty can be resolved deductively. In the case of incorrect assumptions about the error mechanism, however, deductive

editing may introduce a bias in the estimators. In practice, a deductive method might also be used to resolve certain random errors, for reasons of efficiency, provided that the introduced bias is negligible. An example of this is the deductive resolution of rounding errors (see Scholtus, 2011).

De Waal and Scholtus (2011) make a further distinction between *generic* and *subject-related* systematic errors. Errors of the former type occur for a wide variety of variables in a wide variety of surveys and registers, where the underlying cause is always essentially the same. Apart from the unit of measurement error, other examples include *simple typing errors*, such as interchanged or mistyped digits (Scholtus, 2009) and *sign errors*, such as forgotten minus signs or interchanged pairs of revenues and costs (Scholtus, 2011). For an example that involves a simple typing error, see Section 3.2 below. Generic errors can often be detected and treated automatically by using mathematical techniques.

Subject-related systematic errors are specific to a particular questionnaire or survey. They may be caused by a frequent misunderstanding or misinterpretation of some question such as reporting gross values rather than net values. Another example is that, for some branches of industry, staff is frequently classified as belonging to an incorrect department of the responding enterprise. Subject-related systematic errors are usually detected and treated by applying correction rules that have been specified by subject-matter experts.

The remainder of this text is organised as follows. Section 2.2 further discusses the use of correction rules for subject-related systematic errors. Section 2.3 discusses techniques that treat possibly the most notorious of generic systematic errors, the unit of measurement error. Section 2.4 discusses methods for identifying new systematic errors.

2.2 Correction rules for subject-matter related errors

Subject-matter related errors can often be detected and treated by means of deterministic checking rules. Such rules state which variables are to be considered erroneous when the edits are failed in a certain way. Often, deterministic checking rules also describe how the erroneous variables should be adjusted. In that case, these rules are commonly referred to as *correction rules*.

The general form of a correction rule is as follows:

if (*condition*) **then** (*correction*).

Here, *condition* indicates a combination of values in a record that is not allowed. Subsequently *correction* describes the adjustment that is made to the record to resolve the inconsistency.

An example of a correction rule is:

if (*Number of Temporary Employees* > 0 **and** *Costs of Temporary Employees* = 0)
then *Number of Temporary Employees* := 0. (1)

This rule detects an inconsistency that occurs when a business reports to have employed temporary staff without reporting associated costs. In this example, the inconsistency is treated deductively by making the number of temporary employees equal to zero.

In general, a correction rule is intended to resolve an inconsistency that can be resolved in a unique way on logical and/or content-related grounds, under a certain assumption. If the assumption is valid, the deductive editing method always reproduces the true values. For instance, the correction rule (1)

operates under the assumption that the variable *Costs of Temporary Employees* is reported more accurately than the variable *Number of Temporary Employees*. Making such assumptions in a valid way generally requires subject-matter knowledge and knowledge of the data collection process.

Correction rules are attractive because of their simplicity. However, they may only be used when no important nuances are lost with such a simple approach. If the data do not satisfy the assumptions made, then deductive editing may lead to biased estimators. For instance: if in the above example it happens that some businesses actually forget to report the costs of temporary employees, then, after applying the correction rule (1), we may underestimate the total number of temporary employees for businesses in the target population.

Another potential drawback of using correction rules is that a large collection of correction rules may be difficult to maintain, especially when the collection has grown over a long period of time. In particular, it then becomes difficult to grasp the effects of adding a new correction rule, or removing an old one, or changing the order in which the rules are applied to the data. For this reason, it is usually not recommended to try to treat all possible errors in a rule-based manner, because this would require a very complex set of correction rules. Broadly speaking, deductive editing should be limited to the treatment of systematic errors only. For the treatment of random errors, there exist other methods that are more powerful and less difficult to maintain (see “Statistical Data Editing – Automatic Editing”).

2.3 *The unit of measurement error*

Business surveys usually contain instructions to the reporter that all financial amounts must be rounded to thousands of euros (dollars, pounds, etc.), that all quantities must be rounded to thousands of units, et cetera. Some respondents ignore these instructions and, consequently, report values that are a factor 1000 larger than they actually mean. It is clear that, if these *thousand-errors* are not corrected, the resulting estimates for the figures to be published will be too high. The thousand-error is a commonly encountered special case of the more general unit of measurement error, which occurs whenever respondents report values that are consistently too high or too low by a certain factor.

We refer to a *uniform* unit of measurement error if all variables (of a certain type) in a record are too large by the same factor. It is known that, in practice, records with *partial* unit of measurement errors also occur. A partial unit of measurement error could arise, for instance, if several departments of a business each fill in part of a questionnaire independently. Partial unit of measurement errors are generally more difficult to detect than uniform ones.

Traditional methods for detecting unit of measurement errors usually work by comparing one or more reported amounts with reference values. The type of reference data used and the way in which the comparison takes place varies per statistic and per statistical office. Examples of reference data are: a statement from the same respondent from an earlier period, the median value of a number of similar respondents in an earlier period or the same period, and available register data about the respondent.

A widely used method computes the ratio of the unedited value and the reference value. If this ratio is larger than a lower bound, or lies between certain bounds, then it is concluded that the unedited value contains a unit of measurement error. Once a unit of measurement error has been detected, it is treated deductively by dividing all relevant amounts by an appropriate factor. It is often assumed for convenience that all unit of measurement errors are uniform.

For instance: in the Dutch Short Term Statistics, thousand-errors are detected as follows (Hoogland et al., 2011). The total turnover indicated by the respondent for period t , say x_t , is compared to the turnover from the most recent period for which a statement from the respondent is available, up to a maximum of six previous periods. The stated turnover for this earlier period must also not be equal to zero. A thousand-error is detected in x_t if the following applies:

$$|x_t| > 300 \times |x_{t-i}|, \quad \text{for some } i \in \{1, \dots, 6\}.$$

If no data from the respondent from an earlier period are available, then the median of the turnover from the previous period in the stratum of the respondent is used instead. The stratification is based on economic activity and number of employees. A thousand-error is detected in x_t if the following applies:

$$|x_t| > 100 \times \text{stratum median}(x_{t-1}).$$

If a thousand-error is detected by either formula, then it is resolved by dividing the total turnover and all the sub-items by 1000.

Table 1 shows an example of a record with a thousand-error that was found in this way.

Table 1. Example of a uniform thousand-error

	reference data	unedited data	data after treatment
<i>first sub-item turnover</i>	3,331	3,148,249	3,148
<i>second sub-item turnover</i>	709	936,142	936
<i>total turnover</i>	4,040	4,084,391	4,084

It should be noted that the above-described method assumes that the reference value is not affected by unit of measurement errors. Thus, the reference value should either be based on previously edited data, or it should be calculated in a way that is robust to the presence of (some) unit of measurement errors.

Clearly, the choice of bounds in the detection method for unit of measurement errors is important. There is a trade-off here between the number of missed errors (observations that are supposedly correct, but actually contain unit of measurement errors) and the number of false hits (observations that supposedly contain unit of measurement errors, but are actually correct). If previously edited data are available, then a simulation study can be conducted to experiment with different bounds. See Pannekoek and De Waal (2005) for an example of such a simulation study.

In manual editing, unit of measurement errors are often detected using a graphical aid. As an illustration, Figure 1 shows a scatter plot of unedited values of turnover (on the y axis) against reference values (on the x axis), with both variables plotted on a logarithmic scale (using the logarithm to base 10). A cluster of thousand-errors can clearly be identified near the line $y = x + 3$.

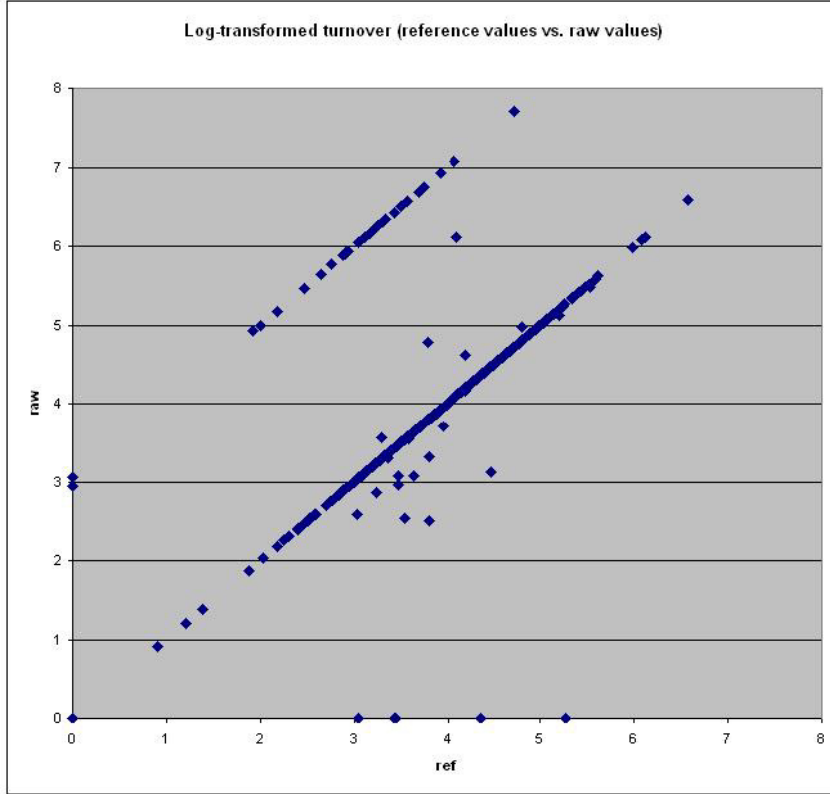


Figure 1. A scatter plot displaying thousand-errors on a logarithmic scale

Elaborating on this graphical approach, Al-Hamad et al. (2008) proposed an alternative automatic method for detecting unit of measurement errors. They considered the difference between the number of digits in the unedited value and the reference value:

$$diff = \left| \lceil \log_{10} x \rceil - \lceil \log_{10} x_{ref} \rceil \right|, \quad (2)$$

where $\lceil a \rceil$ denotes the smallest integer larger than or equal to a . Using (2), different types of unit of measurement errors may be detected by identifying records with a certain value of $diff$. For example, a thousand-error corresponds to $diff = 3$. It should be noted that this method can also detect unit of measurement errors in the reference data, because the absolute value is taken in (2).

Di Zio et al. (2005) proposed a more complex method for detecting unit of measurement errors, by explicitly modeling both the true data and the error mechanism. They used a so-called finite mixture model to identify different clusters within the data set. Each cluster contains records that are affected by a particular type of unit of measurement error; there is also one cluster of records without unit of measurement errors.

Compared with the traditional methods for detecting unit of measurement errors, the approach of Di Zio et al. (2005) has several interesting features. First, it does not require reference data, because the model is fitted directly to the unedited data. However, reference values may also be included in the model if they are available. Second, the method provides diagnostic measures of its own performance, which can be used to identify observations with a significant probability of being misclassified. A selection of doubtful cases may then be checked by subject-matter experts. Finally, this method provides a natural way to detect partial unit of measurement errors. A drawback of the method is that it

may not always be possible to fit an appropriate model to the data set, especially for data sets with many variables or irregular structures. Di Zio et al. (2007) consider an extension of this approach that can accommodate more general models.

2.4 Identifying new systematic errors

New systematic errors can be identified by analysing edit rule failures. If an edit rule is frequently failed, this can be an indication of the presence of a systematic error in the relevant variables. A further analysis of the records that fail the edit rule, in which the questionnaire is also examined, can bring the cause of the error to light. Once the error has been identified, it is generally quite simple to draw up a deductive method to automatically detect and treat the error.

Detecting new systematic errors can only take place once sufficient data have been collected. The results are therefore usually too late to be used in the production process of the current survey cycle. If the analysis produces new deductive editing methods, then these can be built into the editing process for the data in the next survey cycle.

As far as systematic errors are concerned, prevention is better than cure. Sometimes it is possible to improve the design of the questionnaire so that far fewer respondents make a certain type of error. If many respondents make the same kind of error, this can in fact be an indication that a certain question is not presented clearly enough. In some cases, it is also possible to adapt the processing procedure to ensure that a certain processing error no longer arises. In principle, this approach should be preferred to that of making deductive adjustments afterwards. However, because there are practical objections to the constant adaptation of the questionnaire, one may choose initially to build in a deductive editing method, and to use the accumulated knowledge of systematic errors later in a redesign of the questionnaire. (See also the module “Repeated Surveys – Repeated Surveys”.) Moreover, some systematic errors appear to be impossible to prevent, no matter how well the questionnaire is designed. This is, for instance, the case with the unit of measurement error.

To illustrate the identification of a new systematic error, we consider the data collected in 2001 for the Dutch Structural Business Statistics for Wholesale. In this data set, there are (among many other variables) five variables on labour costs, which should satisfy the following edit rule:

$$x_1 + x_2 + x_3 + x_4 = x_5. \quad (3)$$

Here, x_5 represents the variable *total labour costs*. The other four variables are the sub-items of this total. Table 2 shows several records that do not satisfy edit rule (3).

Table 2. Examples of inconsistent partial records in the Dutch SBS for Wholesale 2001

	record 1	record 2	record 3	record 4
x_1	1,100	364	1,135	901
x_2	88	46	196	134
x_3	40	34	68	0
x_4	42	0	42	0
x_5	170	80	306	134

It is striking that, for all records in Table 2, it holds that $x_2 + x_3 + x_4 = x_5$. This suggests that these reporters have ignored the first sub-item x_1 in the calculation of x_5 . A closer look at the questionnaire (see Figure 2) reveals why this could have happened: there is a gap between the answer box for x_1 and the other boxes. As a result, from the design of the questionnaire alone, it is ambiguous whether x_1 should be part of the sum or separate from the rest. Most respondents understand from the context what the intention is, but in several dozen records, we found the same error as in Table 2.

Arbeidskosten	
D.4	Brutolonen en -salarissen van het bij vraag B.1 opgegeven personeel
	Sociale lasten, bestaande uit:
D.5	Werkgeversaandeel sociale voorzieningen
D.6	Pensioenlasten
D.7	Overige sociale lasten
D.8	Totaal arbeidskosten

Figure 2. Part of the questionnaire used for the Dutch SBS Wholesale (until 2005)

We can draw up a deductive method that resolves this error. A more structural solution consists of removing the cause of the error by adapting the questionnaire. This has already been done: the questionnaire from Figure 2 was replaced for the Dutch Structural Business Statistics of 2006. On the new questionnaire, the answer boxes are spaced evenly.

3. Preparatory phase

4. Examples – not tool specific

4.1 Example: Correction rules for the statistic Building Objects in Preparation¹

The Dutch quarterly statistic Building Objects in Preparation (BOP) follows the development of the total construction value of new contracts at architectural firms in the Netherlands. In 2007, a new editing process was designed for this statistic.

When filling in the BOP questionnaire, the reporter must answer several questions about each building object separately. The reporter must tick a box indicating whether the building object concerns a residence (r), a combined-purpose building (c ; this means that the building is used for other purposes as well as residential purposes) or neither of these (o for other). Another question concerns n , the total number of dwellings in the building. For a combined-purpose building, the percentage of floor area intended for residential use (p) is also requested.

¹ This example is adapted from a report written in Dutch by Mark van der Loo and Jeroen Pannekoek (Statistics Netherlands).

The statement contains an error if zero, two, or three of the boxes for r , c , and o have been ticked. In that case, the type of building object has not been clearly specified. In certain situations, this error can be treated deductively based on the values of n and p .

If the value indicated for n is greater than zero and if, moreover, p is equal to 100% or is not filled in, then it is obvious that the building object is a residence. If n is larger than zero and furthermore if p is not equal to 0 or 100%, it is obvious that the building object is a combined-purpose building. And, finally, if neither n nor p has been filled in, or if they have been given the value of 0, then it is highly probable that the building object falls in the category ‘other’. These interpretations follow from the assumption that the statement must be rendered correct by changing as few values as possible.

We write $r = T$ if the box for residence has been ticked, and otherwise $r = F$, and we do the same for c and o . The following correction rule expresses the deductive assertions made in the previous paragraph in formal notation:

```

if  $(r,c,o) \in \{ (T,T,T) , (T,T,F) , (T,F,T) , (F,T,T) , (F,F,F) \}$ 
  then
    if  $( p = \text{'empty'} \text{ or } p = 100\% ) \text{ and } n > 0$ 
      then  $(r,c,o) = (T,F,F)$ 
    if  $0\% < p < 100\% \text{ and } n > 0$ 
      then  $(r,c,o) = (F,T,F)$ 
    if  $( p = \text{'empty'} \text{ or } p = 0\% ) \text{ and } ( n = \text{'empty'} \text{ or } n = 0 )$ 
      then  $(r,c,o) = (F,F,T)$ .

```

This is a small part of the editing process for the statistic BOP.

In the implementation of the editing process for BOP, the derivation of the correction always takes place separately from the actual application of the correction. Initially, in the above example, only an indicator is created that specifies for each record whether a deductive correction is applicable, and if so, which one. Only in the next step are the values of r , c and o changed in the record. As such, the editing process is transparent, so that it is clearly visible afterwards exactly what changes have been made to each record.

4.2 Example: Simple typing errors

We consider a fictitious survey in which the values of *Turnover*, *Costs*, and *Profit* are asked from businesses. By definition, these variables are related through the following edit rule:

$$\textit{Turnover} - \textit{Costs} = \textit{Profit}. \quad (4)$$

The first column of Table 3 shows a record that is inconsistent with respect to (4). The inconsistency can be resolved by adapting any one of the three variables. Moreover, under the assumption that only one variable contains an error, its true value can be computed by inserting the observed values of the other variables into equation (4). The other columns of Table 3 show the three consistent versions of the original record that can be produced by adapting one of the variables (the adapted value is shown in bold in each column).

Table 3. Example of a record with a simple typing error

	record	adjustment 1	adjustment 2	adjustment 3
<i>Turnover</i>	252	315	252	252
<i>Costs</i>	192	192	129	192
<i>Profit</i>	123	123	123	60

Intuitively, the solution in which *Costs* is adapted is the most attractive, since it has the nice interpretation that two adjacent digits were interchanged by mistake. That is to say, it seems much more probable that the true value of 129 was changed to 192 at some point during the collection and processing of the data, than the case that 315 was changed to 252 or 60 to 123. Therefore, we could draw up the following rule for deductive editing: if a record does not satisfy (4), but it can be made consistent by interchanging two adjacent digits in one of the observed values (and, moreover, this can be done in a unique way), then the inconsistency should be treated in this way.

Interchanging two adjacent digits is an example of a simple typing error. Other examples include:

- adding a digit (for example, writing ‘1629’ instead of ‘129’);
- omitting a digit (for example, writing ‘19’ instead of ‘129’);
- replacing a digit (for example, writing ‘149’ instead of ‘129’).

Common features of all simple typing errors are that they only affect one value at a time, and that they produce an observed erroneous value which is related to the unobserved true value in a way that is easy to recognise.

In the example from Table 3, the simple typing error could be detected by using the fact that the variables should satisfy edit rule (4). In general, a survey may contain variables that are related by many equalities and also by other types of edit rules. Moreover, the equalities may be interrelated, so that variables have to satisfy different edit rules simultaneously. Scholtus (2009) described a deductive method for detecting and treating simple typing errors in this more general setting.

Simple typing errors are generic errors, because they occur in many different surveys and they are not content-related. This type of error is easy to make and can therefore occur frequently in practice. A review of data from the Dutch Structural Business Statistics for Wholesale in 2007 revealed, for example, that nearly 10% of all inconsistencies in linear equalities could be explained by one of the four typing errors mentioned above (Scholtus, 2009).

5. Examples – tool specific

The R package `deducorrect`, which can be downloaded for free at <http://cran.r-project.org>, contains an implementation of deductive editing methods for several generic errors:

- sign errors and interchanged values;
- simple typing errors (as defined in Section 3.2);
- rounding errors (very small inconsistencies with respect to equality constraints).

The underlying methodology is described by Scholtus (2011) for sign errors and rounding errors, and by Scholtus (2009) for simple typing errors. To illustrate the use of `deducorrect`, we work out an example. Consider a data set of 11 variables that should satisfy the following edit rules:

$$\left\{ \begin{array}{l} x_1 + x_2 = x_3 \\ x_2 = x_4 \\ x_5 + x_6 + x_7 = x_8 \\ x_3 + x_8 = x_9 \\ x_9 - x_{10} = x_{11} \end{array} \right.$$

The following record is inconsistent with respect to these edit rules; in fact, it does not satisfy the second, fourth, and fifth constraints:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1452	116	1568	161	323	76	12	411	19979	1842	137

We shall use the `deducorrect` package to treat this record for simple typing errors. First, we load the package:

```
> library(deducorrect)
```

Next, we create an object of type “editmatrix” containing the system of edit rules:

```
> E <- editmatrix( c("x1 + x2 == x3",
+                    "x2 == x4",
+                    "x5 + x6 + x7 == x8",
+                    "x3 + x8 == x9",
+                    "x9 - x10 == x11") )
```

We also have to read in the record that we want to treat as a data frame:

```
> x <- data.frame( x1 = 1452, x2 = 116, x3 = 1568, x4 = 161,
+                  x5 = 323, x6 = 76, x7 = 12, x8 = 411,
+                  x9 = 19979, x10 = 1842, x11 = 137 )
```

To check whether simple typing errors can be found in this record, we use the function `correctTypos` provided by the package:

```
> sol <- correctTypos(E, x)
```

The object `sol` is a list which contains the results of the search for simple typing errors. We first check the status of the record:

```
> sol$status
      status
1 corrected
```

The status ‘corrected’ means that the record could be made consistent with respect to all edit rules by only treating simple typing errors. Other possible statuses are: ‘valid’ for a record that was consistent in the first place, ‘invalid’ for an inconsistent record in which no typing error could be detected, and ‘partial’ for a record that could be made consistent with respect to some, but not all edit rules by treating simple typing errors.

The list `sol` also contains the adjusted version of the record and a table of the suggested adjustments:

```
> sol$corrected
      x1  x2   x3  x4  x5 x6 x7  x8   x9  x10 x11
1 1452 116 1568 116 323 76 12 411 1979 1842 137

> sol$corrections
      row variable   old  new
1     1         x4   161 116
2     1         x9 19979 1979
```

Thus, `correctTypos` has detected two simple typing errors in this example: the value of x_4 should be 116 instead of 161 (interchanged adjacent digits), and the value of x_9 should be 1979 instead of 19979 (added digit). By treating these errors, a consistent record is obtained with respect to all edit rules.

We refer to Van der Loo et al. (2011) for more details on the `deducorrect` package.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Al-Hamad, A., Lewis, D., and Silva, P. L. N. (2008), Assessing the Performance of the Thousand Pounds Automatic Editing Procedure at the ONS and the Need for an Alternative Approach. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.
- De Jong, A. (2002), Uni-Edit: Standardized Processing of Structural Business Statistics in the Netherlands. Working Paper, UN/ECE Work Session on Statistical Data Editing, Helsinki.
- De Waal, T. and Scholtus, S. (2011), Methods for Automatic Statistical Data Editing. Paper presented at the 2011 KSS International Conference on Statistics and Probability, Busan.
- Di Zio, M., Guarnera, U., and Luzi, O. (2005), Editing Systematic Unity Measure Errors through Mixture Modelling. *Survey Methodology* **31**, 53–63.
- Di Zio, M., Guarnera, U., and Rocci, R. (2007), A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error. *Computational Statistics & Data Analysis* **51**, 2573–2585.
- Hoogland, J., van der Loo, M., Pannekoek, J., and Scholtus, S. (2011), *Data Editing: Detection and Correction of Errors*. Methods Series Theme, Statistics Netherlands, The Hague.
- Pannekoek, J. and de Waal, T. (2005), Automatic Edit and Imputation for Business Surveys: the Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* **21**, 257–286.
- Scholtus, S. (2009), Automatic Correction of Simple Typing Errors in Numerical Data with Balance Edits. Discussion Paper 09046, Statistics Netherlands, The Hague.
- Scholtus, S. (2011), Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data. *Journal of Official Statistics* **27**, 467–490.

Van der Loo, M., de Jonge, E., and Scholtus, S. (2011), Correction of Rounding, Typing, and Sign Errors with the deducorrect Package. Discussion Paper 201119, Statistics Netherlands, The Hague.

Specific section

8. Purpose of the method

Detecting and treating errors in a deductive manner

9. Recommended use of the method

1. The method should be used, in principle, only for detecting and treating systematic errors.
2. Deductive editing is most effective when it is applied at the very beginning of the editing process, before any other form of editing has been used.

10. Possible disadvantages of the method

1. Deductive editing should only be used to treat errors for which the error mechanism is known with sufficient reliability. Deductive adjustments based on invalid assumptions can produce biased estimators.
2. It may be difficult to maintain a large collection of deterministic correction rules over a long period of time. In particular, it becomes difficult to grasp the consequences of adding or removing a correction rule, or changing the order in which the rules are applied, when faced with a large collection of rules.

11. Variants of the method

1. Each type of systematic error requires its own particular variant.

12. Input data

1. A data set containing unedited microdata.
2. If relevant, a data set containing reference data

13. Logical preconditions

1. Missing values
 1. Allowed, but an assumption has to be made on their interpretation (e.g., “consider all missing values to be equal to zero unless evidence to the contrary is found”).
2. Erroneous values
 1. Allowed; in fact, the object of this method is to detect and treat some of them.
3. Other quality related preconditions
 1. n/a
4. Other types of preconditions
 1. n/a

14. Tuning parameters

1. If relevant, a collection of edit rules for the microdata.

2. Other parameters, depending on the particular variant / type of error.

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. A data set containing partially edited microdata, which is an updated version of the first input data set.

17. Properties of the output data

1. Ideally, the data set should contain no more systematic errors, only random errors.

18. Unit of input data suitable for the method

Incremental processing by record

19. User interaction - not tool specific

1. User interaction is not needed during an execution of deductive editing.

20. Logging indicators

1. All adjustments that are introduced by each deductive editing method should be flagged as such. This helps to keep the editing process transparent and it also provides input for future analyses of the editing process itself.

21. Quality indicators of the output data

1. The quality of deductive editing can be assessed in a simulation study. This requires a data set that has been edited by experts to a point where the edited data may be considered error-free. In the simulation study, the original data are edited again using deductive editing methods. The quality of a deductive editing method may then be measured in terms of its success in detecting systematic errors in the original data set.
2. Alternatively, one could also perform a simulation study by introducing artificial systematic errors into an existing data file. The quality of a deductive editing method may then be measured in terms of its success in identifying these artificial errors.

22. Actual use of the method

1. Several forms of deductive editing are used in the production process for Structural Business Statistics at Statistics Netherlands (see De Jong, 2002).

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Repeated Surveys – Repeated Surveys
2. Statistical Data Editing – Main Module

24. Related methods described in other modules

1. Statistical Data Editing – Automatic Editing

25. Mathematical techniques used by the method described in this module

1. n/a

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1. R package `deducorrect`

28. Process step performed by the method

Statistical data editing

Administrative section

29. Module code

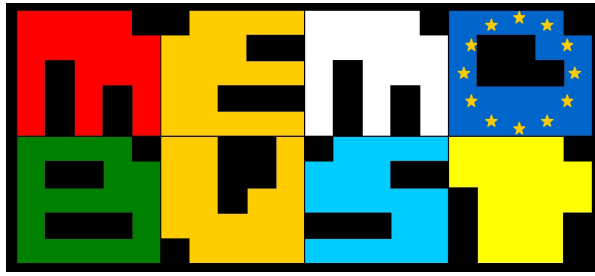
Statistical Data Editing-M-Deductive Editing

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	22-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	20-04-2012	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.2.1	16-07-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.3	04-09-2013	minor improvements	Sander Scholtus	CBS (Netherlands)
0.3.1	09-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:11



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Selective Editing

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Selective editing	3
2.2 Score function.....	3
2.3 The selection rule	4
2.4 How to compute the threshold.....	5
2.5 Dealing with errors remaining in data: a probability sampling approach to selective editing 7	
3. Design issues	8
4. Available software tools.....	8
5. Decision tree of methods	9
6. Glossary.....	9
7. References	9
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

The experience of NSIs in the field of correction of errors has led to assume that only a small subset of observations is affected by influential errors, i.e., errors with a high impact on the estimates, while the rest of the observations are not contaminated or contain errors having small impact on the estimates. Selective editing is a general approach to the detection of errors, and it is based on the idea of looking for important errors in order to focus the treatment on the corresponding subset of units to reduce the cost of the editing phase, while maintaining the desired level of quality of estimates. In this section a general description of the framework and the main elements of selective editing is given.

2. General description

2.1 *Selective editing*

The experience of NSIs in the field of correction of errors has led to assume that only a small subset of observations is affected by influential errors, i.e., errors with a high impact on the estimates, while the rest of the observations are not contaminated or contain errors having small impact on the estimates (Hedlin, 2003). This assumption and the fact that the interactive editing procedures, like for instance, recontact of respondents, are resource demanding, have motivated the idea at the basis of selective editing, that is to look for important errors (errors with an harmful impact on estimates) in order to focus the expensive interactive treatments (follow-up, recontact) only on this subset of units. This should reduce the cost of the editing phase maintaining at the same time an acceptable level of quality of estimates (Lawrence and McKenzie, 2000; Lawrence and McDavitt, 1994). In practice, observations are ranked according to the values of a *score function* expressing the impact of their potential errors on the target estimates (Latouche and Berthelot, 1992), and all the units with a score above a given threshold are selected.

2.2 *Score function*

The score function is an instrument to prioritise observations according to the expected benefit of their correction on the target estimates. According to this definition, it is natural to think of the score function as an estimate of the error affecting data. The estimate is generally based on comparing observed values with predictions (sometimes called *anticipated values*) obtained from some explicit or implicit model for the data. In the case of sample surveys, the comparison should also include the sampling weights in order to properly take into account the error impact on the estimates. An additional element often considered in the context of selective editing, is the *degree of suspiciousness*, that is an indicator measuring, loosely speaking, the probability of being in error. The necessity of this element arises from the implicit assumption of the intermittent nature of the error in survey data, i.e., the assumption that only a certain proportion of the data are affected by error, or, from a probabilistic perspective, that each measured value has a certain probability of being erroneous (Buglielli et al., 2011). Some authors do not introduce this element, others implicitly use it in their proposals. Norberg et al. (2010) state that several case studies indicate that procedures based only on the comparison of observed and predicted values without the use of a degree of suspiciousness tend to generate a large proportion of false alarm.

Several score functions are proposed in literature, the difference being mainly given by the kind of prediction and the use of ‘degree of suspiciousness’.

Among the different methods used to obtain predictions it is worthwhile to mention the use of information coming from a previous occasion of the survey (Latouche and Berthelot, 1992), regression models (Norberg et al., 2010), contamination models (Buglielli et al., 2011). A detailed review can be found in De Waal et al. (2011).

As far as the degree of suspiciousness is concerned, a common drastic approach consists in introducing it in the score function through a zero-one indicator that multiplies the difference between observed and predicted values, where zero and one correspond to consistency or inconsistency respectively with respect to some edit rules. In this case it is assumed that errors appear only as edit failures and observations that pass the edits are considered error-free without uncertainty (Latouche and Berthelot, 1992). More refined methods to estimate the probability of being in error can be found in Norberg et al. (2010) and Buglielli et al. (2011). In the first case a nonparametric approach based on quantiles is used, while in the second a latent model based on a mixture of normal (or lognormal) distributions is proposed.

Prediction and suspiciousness can be combined to form a score for a single variable, named *local score*. A local score frequently used for the unit i with respect to the variable Y_j is

$$S_{ij} = \frac{p_i w_i |y_{ij} - \tilde{y}_{ij}|}{\hat{T}_{Y_j}}$$

where p_i is the degree of suspiciousness, y_{ij} is the observed value of the variable Y_j on the i th unit, \tilde{y}_{ij} is the corresponding prediction, w_i is the sampling weight, and \hat{T}_{Y_j} is an estimate of the target parameter.

Once the local scores for the variables of interest are computed, a global score to prioritise observations is needed.

Several functions can be used to obtain the global score (see Hedlin, 2008); an example is the sum of squares $GS_i^{(2)} = \sum_j S_{ij}^2$.

In some cases, some variables can be considered to be more important than others. Such situations can be dealt with by multiplying the local scores by weights stating their relative importance.

2.3 The selection rule

Once the observations have been ordered according to their global score, it is important to build a rule in order to determine the number of units to be reviewed.

A first rule can be suggested by budget constraints. In this case, it is obvious to choose the first n^* observations, in the given ordering, such that the budget constraints are satisfied.

A more interesting and complex approach is to select the subset of units such that the impact on the target estimates of the errors remaining in the unedited observations is negligible, that is in fact the core of selective editing. Since the true values are unknown, this bias cannot be evaluated and an approximation is used. This approximation can be expressed in terms of the weighted differences

between the raw values y_{ij} and the anticipated values \tilde{y}_{ij} for the variable Y_j in the units i not selected for interactive treatment (EDIMBUS, 2007).

Let T_{Y_j} be the target quantity related to the variable Y_j (for instance the total), the estimated bias is given by

$$EB_j(t) = \frac{\left| \sum_{i \notin E_t} w_i (y_{ij} - \tilde{y}_{ij}) \right|}{\hat{T}_{Y_j}},$$

where w_i is the sampling weight of the i th unit, \hat{T}_{Y_j} is an estimate of the target quantity T_{Y_j} , and E_t is the set of units to be selected. This set is composed of all the units having a global score $GS > t$, where t is a threshold value such that $EB_j(t)$ is below a predefined value.

An alternative measure known as the *estimated relative bias* is obtained by replacing the estimate of the total at the denominator of EB with the standard error of the estimate \hat{T}_{Y_j} . With this measure, the error due to the non-sampling error left in data is compared with the sampling error. The reasoning underlying is that there is no need to edit observations because the ‘noise’ due to their errors is overwhelmed by the sampling error.

We remark that when edited values are available, they can be used as anticipated values, in this case the estimated bias and the estimated relative bias are the absolute pseudo bias and the relative pseudo-bias introduced by Latouche and Berthelot (1992) and Lawrence and McDavitt (1994), respectively.

It is worthwhile to note the similarity between the terms appearing in the sum defining the estimated bias and the local score function. The main difference is in the parameter related to the suspiciousness. In fact in the estimated bias all differences between observed values and corresponding predictions are considered as they were determined by errors, while in the score functions, where the degree of suspiciousness is included, this is not assumed with certainty.

2.4 How to compute the threshold

There are two approaches: 1) through a simulation study, 2) by using a model.

2.4.1 Simulation approach

This approach is based on the availability of raw and edited data comparable with the data on which selective editing has to be applied. The idea is to simulate the selective editing procedure considering the edited data as if they were the ‘true’ data. Often data from a previous cycle of the same survey are used for this purpose.

The approach can be described by the following steps (De Waal et al., 2011).

- Compute the global scores for the raw data and order (decreasingly) the observations.
- Determine a subset E of units composed of the first p units and replace their raw values with the corresponding edited values.
- Compare the estimates computed using the completely edited data set and the raw data where the subset E is obtained according to step 2.

- Repeat steps 2 and 3 with different values of p until the difference between the two estimates is negligible. Let p^* be the first index such that this condition is fulfilled.
- The threshold t is the value of the GS corresponding to the p^* -th unit.

Remarks:

- The assumption of this approach is that the edited data can be considered as ‘true’ data. This is a limitation because it can be rarely assumed.
- The simulation approach is frequently applied to data of a previous survey occasion to obtain a threshold value to be used for the current survey. It is worthwhile to note that in this case we assume that the error mechanism and the data distribution are the same in the two occasions.
- The method cannot be applied when you deal with the first wave of a survey.

2.4.2 Model based approaches

In this context, some of the main elements of the problem are modelled through a probability distribution: the true data distribution, the error mechanism, the score functions.

The introduction of a model may be useful to give estimates of the error left in data after the revision of the selected units and thus to ease the determination of a threshold for the selection of units to be reviewed.

A first attempt can be found in Lawrence and McKenzie (2000). By denoting with a the threshold value, they assume that the difference between the observed and the predicted value for the non-selected observations follows a uniform distribution in the interval $(-a, a)$, i.e., $U(-a, a)$. The threshold a is determined so that the bias due to not editing a set of units is low if compared to the sampling error.

A conservative solution is $a = \sqrt{\frac{3k}{n}} SE(\hat{Y})$, where $kSE(\hat{Y})$, $k < 1$ is the upper bound for the bias and n is the total number of observations.

The intermittent nature of the error is taken into account in Arbués et al. (2011). The search of a good selective editing strategy is stated as an optimisation problem in which the objective is to minimise the expected workload with the constraint that the expected error of the aggregates computed with the edited data is below a certain constant.

A model based approach is also adopted by Buglielli et al. (2011). They propose to consider (log)- true data y_i^* as realisations from a multivariate Gaussian distribution with mean vector possibly dependent on a set of error-free covariates: $\tilde{y}_i \sim N(\mu_i, \Sigma)$. Errors are supposed to act on a subset of data by inflating the variance, i.e., the covariance matrix of the contaminated data is $\lambda\Sigma$ where λ is a numerical factor greater than one. The intermittent nature of the error is reflected by a Bernoullian random variable with parameter π taking values zero or one depending on whether an error occurs in a unit or not, respectively. This approach naturally leads to a latent class model formulation, where observed data (y) can be viewed as realisation from a mixture of two Gaussian probability distributions associated to contaminated and error-free data:

$$f_Y(y) = (1 - \pi)N(y; \boldsymbol{\mu}, \Sigma) + \pi N(y; \boldsymbol{\mu}, (\boldsymbol{\lambda} + 1)\Sigma).$$

In this context, the parameter π represents the mixing weight of the mixture and can be interpreted as the *a priori* probability of errors in data. The estimated conditional distribution of true data given observed ones is used to build an appropriate score function. More precisely, for a given variable of interest, a relative (local) score function is defined in terms of difference between the observed value and the expectation of the “true” value conditional on the observed one (the prediction). This approach allows to interpret the score function as the expected error, and to relate the threshold for interacting reviewing to the accuracy of the estimates of interest. A global score can be defined in many ways combining the different local score functions. In Buglielli et al. (2011) the global score is defined as the maximum of the single local scores. This ensures that the accuracy of the estimates is kept under control simultaneously for all the variables of interest.

In practice the steps to perform selective editing within this framework are similar to the ones detailed in the simulation approach, with the difference that the predicted value is obtained by using an explicit model, and that the score directly gives an estimate of the error contaminating each observation.

Remarks:

- The introduction of a model for the error mechanism allows to formalise the problem and hence to have a statistical interpretation of the elements characterising selective editing. Furthermore, using a latent class model implies the advantage that no edited data are required, and the bias of the simulation approach due to considering edited data as true data is avoided.
- The main drawback is that the validity of the conclusions depends on the validity of the model assumptions.

2.5 *Dealing with errors remaining in data: a probability sampling approach to selective editing*

Ilves and Laitila (2009) and Ilves (2010) propose a two-step procedure for selective editing. Their proposal is motivated by the fact that the non-selected observations may still be affected by errors resulting in a biased target parameter estimator \hat{T}_Y . To obtain an unbiased estimator a sub-sample is drawn from the unedited observations (below threshold for global scores), follow-up activities with recontacts are carried through and the bias due to remaining errors is estimated.

The estimated bias is used to make the target parameter estimator \hat{T}_Y unbiased. If our target parameter is the total of the population, the bias-corrected estimator is obtained by subtracting the estimated bias from the HT estimator of the total computed on edited (selected by the selective editing procedure) and unedited (non-selected) observations. Formulas for the variance and a variance estimator are derived by using a two-phase sampling approach. The procedure is discussed in general without specifying a particular selective editing technique, but sampling with probabilities proportional to scores seems to be the obvious choice.

3. Design issues

In the following some important elements concerning the design of a selective editing procedure are reported.

- Selective editing can be applied only to numerical variables. This implies that selective editing is mainly applied to business surveys.
- Selective editing is useful when accurate interactive editing can be performed.
- Selective editing can be applied at the early stages of data collection. This kind of application is named *input editing*. The methods used in this context apply to each incoming record individually, classifying each record as critical or non-critical. The advantage of input editing is that time-consuming task procedures as interactive editing and follow-up are started as soon as possible, with positive effects on response burden and the timeliness of the results. The disadvantage is that the parameters needed for the selection of influential errors should be estimated before data are available. This can be performed only when data from previous survey occasions are available (or strong a priori knowledge is disposable), and the assumptions are that the situation is not changed from the previous surveys to the actual one. On the contrary, the approach consisting in applying selective editing when almost all the data are available is named *output editing*. The disadvantage is clearly related to the timeliness of the results because time consuming task as interactive editing or follow-up are moved to a later stage of the process. The advantage is that all the parameters needed for the selection of influential errors are estimated on the data at hand, so they refer to the actual distribution of data with a potential benefit effect on the precision of selection.
- It is advisable to apply selective editing after the process of detection and correction of systematic errors (see “Statistical Data Editing – Main Module”). Actually, also systematic errors can lead to significant bias but they can often be automatically detected and corrected easily and very reliably. It is highly efficient to correct these errors at an early stage.
- The application of selective editing should be limited to the subset composed of the most important target variables.
- Once one observation is selected, all the variables should possibly be revised, not only the ones considered in the score function.
- Sampling weights are important to estimate the impact of errors on the final estimates. When an input editing approach is chosen, initial sampling weights may be used.

4. Available software tools

- SeleMix is an R-package for selective editing based on contamination models (Di Zio and Guarnera, 2011) freely available on the website <http://cran.r-project.org/>.
- Selekt is a set of SAS-macros for selective editing, allowing “traditional” hard and soft edits as well as a nonparametric approach based on quantiles to produce measures of suspicion. Selekt works with one and two-stage samples and several sets of domains in output. (Norberg et al., 2010; Norberg, et al., 2011).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Arbués, I., Revilla, P., and Saldaña, S. (2011), Selective Editing as a Stochastic Optimization Problem. UN/ECE Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9-11 May 2011.
- Buglielli, T., Di Zio, M., Guarnera, U., and Pogelli, F. R. (2011), Selective Editing of Business Survey Data Based on Contamination Models: an Experimental Application. NTTS 2011 New Techniques and Technologies for Statistics, Bruxelles, 22-24 February 2011.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- Di Zio, M. and Guarnera, U. (2011), SeleMix: an R Package for Selective Editing via Contamination Models. *Proceedings of the 2011 International Methodology Symposium, Statistics Canada. November 1-4, 2011, Ottawa, Canada*.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Hedlin, D. (2003), Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics* **19**, 177–199.
- Hedlin, D. (2008), Local and Global Score Functions in Selective Editing. UN/ECE Work Session on Statistical Data Editing, Wien.
- Ilves, M. and Laitila, T. (2009), Probability-Sampling Approach to Editing. *Austrian Journal of Statistics* **38**, 171–182.
- Ilves M. (2010), Probabilistic Approach to Editing. Workshop on Survey Sampling Theory and Methodology Vilnius, Lithuania, August 23-27, 2010.
- Latouche, M. and Berthelot, J. M. (1992), Use of a Score Function To Prioritise and Limit Recontacts in Business Surveys. *Journal of Official Statistics* **8**, 389–400.
- Lawrence, D. and McDavitt, C. (1994), Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics* **10**, 437–447.
- Lawrence, D. and McKenzie, R. (2000), The General Application of Significance Editing. *Journal of Official Statistics* **16**, 243–253.
- Norberg, A. et al. (2010), *A General Methodology for Selective Data Editing*. Statistics Sweden.
- Norberg, A. et al. (2011), *User’s Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing*. Statistics Sweden.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Statistical Data Editing – Automatic Editing
3. Statistical Data Editing – Manual Editing
4. Statistical Data Editing – Macro-Editing
5. Imputation – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

14. Module code

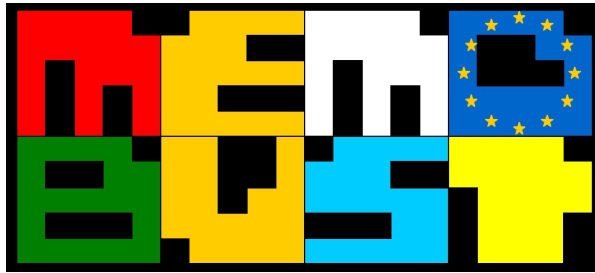
Statistical Data Editing-T-Selective Editing

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	08-02-2012	first version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.2	19-03-2012	second version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.3	06-04-2012	third version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.3.1	04-10-2013	preliminary release		
0.4	15-10-2013	changes according to the EB comments	Di Zio Marco, Guarnera Ugo	Istat (Italy)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:11



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Automatic Editing

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Introduction to automatic editing	3
2.2 Edit rules.....	4
2.3 The error localisation problem	6
2.4 Solving the error localisation problem: the method of Fellegi and Holt	7
2.5 Solving the error localisation problem: other methods	10
3. Preparatory phase	11
4. Examples – not tool specific.....	11
5. Examples – tool specific.....	12
6. Glossary.....	13
7. References	13
Specific section.....	16
Interconnections with other modules.....	18
Administrative section.....	20

General section

1. Summary

The goal of automatic editing is to accurately detect and treat errors and missing values in a data file in a fully automated manner, i.e., without human intervention. Methods for automatic editing have been investigated at statistical institutes since the 1960s (Nordbotten, 1963). In practice, automatic editing usually implies that the data are made consistent with respect to a set of predefined constraints: the so-called *edit rules* or *edits*. The data file is checked record by record. If a record fails one or more edit rules, the method produces a list of fields that can be imputed so that all rules are satisfied.

In this module, we focus on automatic editing based on the (generalised) Fellegi-Holt paradigm. This means that the smallest (weighted) number of fields is determined which will allow the record to be imputed consistently. Designating the fields to be imputed is called error localisation. In practice, error localisation by applying the Fellegi-Holt paradigm often requires dedicated software, due to the computational complexity of the problem.

Although the imputation of new values for erroneous fields is often seen as a part of automatic editing, we do not discuss this here, because the topic of imputation is broad and interesting enough to merit a separate description. We refer to the theme module ‘Imputation’ and its associated method modules for a treatment of imputation in general and various imputation methods.

2. General description of the method

2.1 Introduction to automatic editing

For efficiency reasons, it can be desirable to edit at least part of a data file by means of automatic methods (see “Statistical Data Editing – Main Module”). Assuming that all systematic errors with a known structural cause have already been treated using methods for deductive editing (see the method module “Statistical Data Editing – Deductive Editing”), the task remains to also detect and treat random errors. In the literature on data editing, the problem of identifying the erroneous values in a record containing only random errors is known as the *error localisation problem*. Compared to detecting systematic errors, solving the error localisation problem is usually more difficult and requires complex methodology.

Broadly speaking, there are two approaches to solving the error localisation problem. The first approach uses outlier detection techniques in combination with an implicit or explicit statistical model for the data under consideration. Records corresponding to data points that do not fit the model well are supposed to contain errors, and within such a record, the values that contribute most to the ‘outlyingness’ of that record are identified as erroneous; see, e.g., Little and Smith (1987) and Ghosh-Dastidar and Schafer (2003). This approach appears to be mainly suitable for editing low-dimensional data (data sets containing a small number of variables). Moreover, if there are edit rules that define consistency constraints for the variables in the data set, these cannot be used under this approach. In particular, the edited data will not necessarily satisfy the edit rules. For these reasons, this approach is not ideal for automatic editing in business surveys at statistical offices, where one typically encounters data sets with many variables and many edit rules. In fact, it is seldom used in this context.

In the remainder of this module, we shall focus on the second approach. Under this approach, a set of edit rules is defined for the data set. A record is called *consistent* – and is considered to be error-free – if it satisfies all edit rules. For inconsistent records, the erroneous values are identified by solving a mathematical optimisation problem.

The remainder of this section is organised as follows. Section 2.2 considers edit rules. In Section 2.3, the error localisation problem is formulated as a mathematical optimisation problem. Sections 2.4 and 2.5 describe techniques for solving this optimisation problem.

2.2 Edit rules

Edit rules are introduced in a more general context in “Statistical Data Editing – Main Module”. Here, we focus on aspects of edit rules that are relevant to automatic editing in particular.

A record of data can be represented as a vector of fields or variables: $x = (x_1, x_2, \dots, x_n)$. The set of values that can be taken by variable x_i is called its domain. Examples of variables and domains are *size class* with domain {'small', 'medium', 'large'}, *number of employees* with domain {0,1,2,...}, and *profit* with domain $(-\infty, \infty)$.

Edit rules indicate conditions that should be satisfied¹ by the values of single variables or combinations of variables in a record. For the purpose of automatic editing, all edit rules must be checkable per record, and may therefore not depend on values in fields of other records. However, they may contain parameters based on external sources (for instance, quantiles of univariate distributions in a reference data set that has been edited previously), provided that these parameters are set prior to the start of the editing process.

For automatic editing of numerical data, it is convenient to assume that all edit rules are written as linear relationships such as

$$Turnover \geq 0$$

or

$$Profit + Costs = Turnover.$$

The general form of a linear edit rule for a record (x_1, x_2, \dots, x_n) is as follows:

$$a_{j1}x_1 + \dots + a_{jn}x_n + b_j \geq 0 \tag{1}$$

or

$$a_{j1}x_1 + \dots + a_{jn}x_n + b_j = 0, \tag{2}$$

¹ Edit rules of this type are sometimes called ‘validity rules’. In some applications, edit rules are specified instead in the form of ‘conflict rules’, which means that they indicate conditions that are satisfied by invalid combinations of values. For instance, an edit rule stating that the variable *turnover* should be non-negative can be written either as the validity rule ‘*turnover* ≥ 0 ’ or as the conflict rule ‘*turnover* < 0 ’. Clearly, both formulations are equivalent. The choice of validity or conflict rules should not lead to difficulties, provided that one of the forms is used consistently.

where j numbers the edit rules, a_{ji} are numerical coefficients and b_j are numerical constants. It should be noted that a *ratio edit* – i.e., a bivariate edit rule of the form

$$x_1/x_2 \geq a,$$

where a denotes a numerical constant and x_1 and x_2 are constrained to be non-negative – can also be expressed as a linear edit rule. Namely, the ratio edit can be rewritten as

$$x_1 - ax_2 \geq 0.$$

For categorical data, an edit rule can identify as admissible any combination of values from the domains of the categorical variables. Categorical edit rules are often written in if-then form, for example:

if *Gender* = ‘male’ **then** *Pregnant* = ‘no’.

Finally, mixed data and mixed edit rules, containing both categorical and numerical variables also occur in practice. Mixed edit rules are also often written in if-then form. For example:

if *Size Class* = ‘small’ **then** *Number of Employees* < 10.

For automatic processing, it can be convenient to require that the if-part of a mixed edit only contains categorical variables, while the then-part only contains numerical variables. The above-mentioned example is written in this form. Many types of mixed edits can be rewritten in this simple form, although this may require the introduction of auxiliary variables; see De Waal (2005).

In the remainder of this module, we focus on numerical data and linear edits, because these are most common to business surveys. We refer to De Waal et al. (2011) for a discussion of automatic editing of categorical or mixed data. A numerical variable x_i is said to be *involved* in an edit rule of the form (1) or (2) if it holds that $a_{ji} \neq 0$. Clearly, whether a record fails or satisfies an edit rule only depends on the values of the variables that are involved in that edit rule.

In manual editing, subject-matter specialists often distinguish between *hard* and *soft* edit rules. As mentioned in “Statistical Data Editing – Main Module”, hard edit rules are rules that must hold by definition, while soft edit rules only indicate whether a value, or value combination, is suspicious. A soft edit rule can occasionally be failed by unlikely values that are in fact correct.

In nearly² all methods for automatic editing, no distinction can be made between hard and soft edit rules: all rules are treated as hard edit rules. Thus, in automatic error localisation, all records that fail one or more edit rules are viewed as certainly inconsistent. Hence, formulating edits for the purpose of automatic editing should be done with care (Di Zio et al., 2005). If too many soft edit rules are defined, or soft edit rules that are too strict, there is a danger of *overediting*: the unjustified adaptation of correct values. On the other hand, if too few edit rules are defined, or soft edit rules that are not strict enough, then certain errors might be left in the data after automatic editing.

² In fact, to our best knowledge, all methods for automatic editing that are currently in use at statistical offices do not distinguish between hard and soft edit rules. The method of Freund and Hartley (1967) uses soft edit rules, but it has the important drawback that it cannot handle hard edit rules; hence, it is not recommended to be used in practice. Scholtus (2013) has described a method that incorporates both hard and soft edit rules, but at the time of writing, this method remains to be tested in practice.

2.3 The error localisation problem

For a given record and a collection of edit rules, it is straightforward to verify which values in the record are missing and whether any of the edit rules are failed. However, given that some of the edit rules are failed, determining which values in the record are actually causing the edit failures is much less straightforward. On the one hand, most edit rules involve more than one variable, and on the other hand, most variables are involved in more than one edit rule.

In order to solve the error localisation problem automatically, one has to choose a guiding principle for finding errors. The most commonly used guiding principle for error localisation is the so-called *Fellegi-Holt paradigm*, first formulated by Fellegi and Holt (1976). According to this paradigm, one should minimise the number of observed values that have to be adjusted in order to satisfy all edit rules. This paradigm is often used in a generalised form, for which each variable is given a *reliability weight* $w_i \geq 0$. A high value of w_i indicates that the variable x_i is expected to contain few errors. The generalised Fellegi-Holt paradigm now states that one should search for a subset of the variables E with the following two properties:

- The variables x_i ($i \in E$) can be imputed with values that, together with the observed values of the other variables in the record, satisfy all edit rules.
- Among all subsets that satisfy the first property, E has the smallest value of $\sum_{i \in E} w_i$.

The original Fellegi-Holt paradigm is recovered from this more general form by taking all reliability weights equal, for instance all equal to 1.

A distinctive feature of the (generalised) Fellegi-Holt paradigm is that it does not take the size of the differences between the original and imputed values into account in any way. In fact, the method of Fellegi and Holt only provides a list of variables that can be imputed to satisfy all edit rules, but it does not provide the actual values to impute. These have to be determined in a separate step. This might seem like a drawback, but it actually has the advantage that an appropriate imputation method can be chosen independently of the method used for error localisation. Methods for imputation are discussed in the topic “Imputation”.

Some authors have suggested other guiding principles for error localisation that do look at the size of the adaptations. Casado Valera et al. (1996) proposed to minimise the sum of the squared differences between the observed values and the adjusted values, under the restriction that all edit rules are satisfied by the adjusted values. This leads to a quadratic optimisation problem, which can be solved using standard software. A different formulation of the error localisation problem as a quadratic optimisation problem was proposed by Freund and Hartley (1967).

To illustrate the difference between these principles, we consider a very small example. Suppose that there are two edit rules:

$$\text{Turnover} = \text{Profit} + \text{Costs},$$

$$\text{Turnover} \geq 0,$$

and suppose that we are presented with the following inconsistent record:

$$(\text{Turnover}, \text{Profit}, \text{Costs}) = (-30, 10, 20).$$

Under the Fellegi-Holt paradigm (in its original form, without reliability weights), the optimal solution is to adjust only the value of *Turnover*, because both edits can be satisfied without changing the values of the other variables. After imputation, this certainly yields

$$(\textit{Turnover}, \textit{Profit}, \textit{Costs}) = (30, 10, 20),$$

because the value to impute for *Turnover* is uniquely determined by the edits in this example.

On the other hand, if we minimise the unweighted sum of the squared differences between observed and adjusted values, the optimal solution changes all values:

$$(\textit{Turnover}, \textit{Profit}, \textit{Costs}) = (0, -5, 5).$$

This happens because, under this minimisation criterion, it is optimal to distribute the total adjustment required by the edit rules over as many different variables as possible.

Assuming that errors occur with a low probability and in isolated values, the Fellegi-Holt paradigm appears to be a sensible choice, because it distorts as few of the observed values as possible. Methods that try to distribute the total adjustment over many different variables, such as the quadratic minimisation approach, are less suitable in this context. However, the latter type of method can be useful in the context of micro- or macro-integration, where many small inconsistencies in data from different sources have to be resolved, while preserving patterns that occur in the original data as much as possible. We refer to the topics “Micro-Fusion” (in particular the method module “Micro-Fusion – Reconciling Conflicting Microdata”) and “Macro-Integration” for these subjects.

In order to solve the error localisation problem according to the Fellegi-Holt paradigm, we have to find the smallest subset of the variables that can be imputed so that all edit rules become satisfied. Several methods have been proposed for this. Section 2.4 presents the original method of Fellegi and Holt (1976) for numerical data. Section 2.5 briefly mentions several other methods. These sections contain material that is somewhat more technical than the rest of this module.

2.4 Solving the error localisation problem: the method of Fellegi and Holt

For a given record that fails certain edit rules, we want to determine the smallest subset of the variables that can be imputed so that all edit failures are resolved. A naïve way to solve this problem might proceed as follows: “It is clear that a subset of the variables E can only be a feasible solution to the error localisation problem if every failed edit rule involves at least one variable in E , i.e., if the failed edit rules are ‘covered’ by these variables. Therefore, let us choose the smallest set of variables with this property.” Unfortunately, although ‘covering’ the original failed edits is a necessary condition for a set of variables to be a feasible solution to the error localisation problem, it is not a sufficient condition in general. We will demonstrate this by means of a small example.

Consider the following two numerical edit rules: $x_1 \geq x_2$ and $x_2 \geq x_3$. The unedited record $(x_1, x_2, x_3) = (4, 5, 6)$ fails both edits. Since the variable x_2 is involved in both edit rules – that is to say, the failed edits are ‘covered’ by x_2 –, we might try to obtain consistency with respect to the edit rules by changing only the value of x_2 . This turns out to be impossible, because the imputed value would have to satisfy $4 \geq x_2$ and $x_2 \geq 6$.

Fellegi and Holt (1976) showed that, in order to determine whether a set of variables can be imputed to satisfy all edits simultaneously, it is necessary to derive so-called *implied edits* from the original set of edits. An implied edit is an edit rule that can be derived from the original edit rules by logical reasoning. For numerical data, the number of implied edits that can be derived from even a single original edit is actually infinite; e.g., if $x_1 \geq x_2$ is an edit rule, then so is $\lambda x_1 \geq \lambda x_2$ for any $\lambda > 0$. Fortunately, for the purpose of solving the error localisation problem, it is not necessary to derive all possible implied edits from the original set of edits, but only the so-called *essentially new implied edits* (see below). By adding the essentially new implied edits to the original set of edit rules, one obtains a so-called *complete set of edits*. For a complete set of edit rules, it does hold that any subset of the variables which ‘covers’ all failed edit rules is a feasible solution to the error localisation problem.

In the example above, the complete set of edits consists of the two original edit rules and the (only) essentially new implied edit $x_1 \geq x_3$. The latter edit rule is also failed and it does not involve the variable x_2 , which shows that imputing only x_2 does not solve the error localisation problem. On the other hand, the three failed edits are ‘covered’ by $\{x_1, x_3\}$, and it is easy to see that imputing new values for x_1 and x_3 is indeed a feasible solution to the error localisation problem. In fact, imputing any combination of values with $x_1 \geq 5$ and $5 \geq x_3$ leads to a consistent record in this example.

In general, for a given set of edit rules of the forms (1) and (2), essentially new implied edits are constructed by selecting one of the variables, say x_g , as a so-called *generating variable*. We consider all pairs of edit rules that involve the generating variable, i.e., all pairs (s, t) with $a_{sg} \neq 0$ and $a_{tg} \neq 0$. If one of the edits, say edit s , is an equality, then we may solve this equality for x_g :

$$x_g = \frac{-1}{a_{sg}} (a_{s1}x_1 + \dots + a_{s,g-1}x_{g-1} + a_{s,g+1}x_{g+1} + \dots + a_{sn}x_n + b_s).$$

An implied edit is now obtained from the pair (s, t) by substituting this expression for x_g in edit rule t . This new edit rule is an essentially new implied edit, unless it happens to be identical to an existing edit rule, in which case it is redundant.³

If both edits are inequalities, then we apply a technique called *Fourier-Motzkin elimination* (Williams, 1986; De Waal et al., 2011). First, we check whether the coefficients a_{sg} and a_{tg} have opposite signs, i.e., whether $a_{sg}a_{tg} < 0$. If this is not the case, then this pair does not contribute an essentially new implied edit. Hence, we may assume without loss of generality that $a_{sg} < 0$ and $a_{tg} > 0$. This means that edit rule s can be written as an upper bound on x_g , given the values of the other variables:

$$x_g \leq \frac{-1}{a_{sg}} (a_{s1}x_1 + \dots + a_{s,g-1}x_{g-1} + a_{s,g+1}x_{g+1} + \dots + a_{sn}x_n + b_s).$$

³ To give an example of a redundant edit, suppose that we already have the edit rule ‘ $x \geq 3$ ’ and we derive a new edit rule stating that ‘ $2x \geq 6$ ’. Since the second edit rule is identical to the first one after simplification, it does not provide any new information and is therefore redundant.

Similarly, edit rule t can be written as a lower bound on x_g :

$$x_g \geq \frac{-1}{a_{tg}} (a_{t1}x_1 + \dots + a_{t,g-1}x_{g-1} + a_{t,g+1}x_{g+1} + \dots + a_{tn}x_n + b_t).$$

Combining the two bounds and removing x_g , we obtain the implicit condition

$$\begin{aligned} & \frac{-1}{a_{sg}} (a_{s1}x_1 + \dots + a_{s,g-1}x_{g-1} + a_{s,g+1}x_{g+1} + \dots + a_{sn}x_n + b_s) \\ & \geq \frac{-1}{a_{tg}} (a_{t1}x_1 + \dots + a_{t,g-1}x_{g-1} + a_{t,g+1}x_{g+1} + \dots + a_{tn}x_n + b_t) \end{aligned}$$

which can be written in the general form (1) as

$$a_1^*x_1 + \dots + a_n^*x_n + b^* \geq 0,$$

with $a_i^* = a_{tg}a_{si} - a_{sg}a_{ti}$ ($i=1, \dots, n$) and $b^* = a_{tg}b_s - a_{sg}b_t$. This is an essentially new implied edit that is derived from the pair of inequality edits (s, t) , unless it happens to be redundant (see footnote 3).

It should be noted that, both for equalities and inequalities, the essentially new implied edit generated by this procedure does not involve the generating variable (i.e., the coefficient $a_g^* = 0$). This is in fact the defining property that makes an implied edit ‘essentially new’: it adds information to the existing edit rules by eliminating one of the variables.

According to the method of Fellegi and Holt (1976), a complete set of edits may be constructed by repeatedly applying the above-mentioned procedure of generating essentially new implied edits, using all variables in turn as generating variables, until no more (non-redundant) new edits can be derived. At that point, a complete set of edits has been generated.

Having obtained a complete set of edits, one can solve the error localisation problem for any given record in the following manner:

- Select all edits from the complete set of edits that are failed by the original record.
- Find the smallest (weighted) subset of the variables with the property that each selected (original or implied) edit involves at least one of them.

The first step amounts to evaluating the edits for a given record. The second step entails solving a set-covering problem, which is a well-known mathematical problem for which standard algorithms are available (see, e.g., Nemhauser and Wolsey, 1988). We shall work out a small example with the Fellegi-Holt method in Section 4.

A crucial element of the Fellegi-Holt method is the fact that a complete set of edits is ‘sufficiently large’ to reduce the error localisation problem to a set-covering problem. For a proof of this fact, see Fellegi and Holt (1976). For an explanation of what is meant by ‘sufficiently large’ from the viewpoint of logic, see Boskovitz et al. (2005).

The method discussed in this section works for numerical variables, but an analogous method exists for categorical variables. The only difference lies in the procedure for generating essentially new

implied edits. We refer to Fellegi and Holt (1976) and De Waal et al. (2011) for a description of the Fellegi-Holt method for categorical data.

2.5 *Solving the error localisation problem: other methods*

An important drawback of the method of Fellegi and Holt discussed in Section 2.4 is that the complete set of edits can be extremely large, especially with numerical data. In many practical applications, generating a complete set of edits is simply not technically feasible.⁴ For this reason, other algorithms have been developed that solve the error localisation problem without generating a complete set of edits. We can distinguish several classes of such algorithms.

Algorithms based on vertex generation

It is known from the literature that the optimal solution to the error localisation problem for a given record always corresponds with one of the vertices of an appropriately defined polyhedron; see, e.g., Theorem 3.1 in De Waal et al. (2011). Hence, in principle, the error localisation problem can be solved by generating all vertices of that polyhedron and identifying the optimal one. This approach has been elaborated in several error localisation algorithms. See, among others, Sande (1978), Kovar and Whitridge (1990), Fillion and Schiopu-Kratina (1993), Todaro (1999), and De Waal (2003). Tools for automatic editing that use algorithms based on vertex generation include GEIS (Kovar and Whitridge, 1990), Banff (Banff Support Team, 2008), CherryPi (De Waal, 1996), and AGGIES (Todaro, 1999).

Branch-and-bound algorithm

De Waal and Quere (2003) describe how the error localisation problem may be solved by means of a branch-and-bound algorithm. For a record containing n numerical variables, there are 2^n potential solutions, since each variable is either fixed to its original value or imputed. Basically, the branch-and-bound algorithm systematically considers all potential solutions and checks which of these are feasible. In order to do this, the algorithm generates relevant essentially new implied edits ‘on the fly’, but it does not construct a complete set of edits. Finally, the algorithm selects the feasible solution with the smallest sum of reliability weights. A similar branch-and-bound algorithm can be used for categorical or mixed data. We refer to De Waal and Quere (2003), De Waal (2003), and De Waal et al. (2011) for more details. Tools for automatic editing that use the branch-and-bound algorithm include SLICE (De Waal, 2005) and the R package `editrules` (De Jonge and Van der Loo, 2011).

Algorithms based on cutting planes

With this approach, to solve the error localisation problem for a given record, one starts by finding the minimal subset of the variables that ‘covers’ all original edit rules that are failed. As we have seen above, this solution may be infeasible. In that case, the algorithm generates new constraints, so-called

⁴ One exception occurs when all edit rules are ratio edits: it can be shown that, for a data set with n variables, the complete set of edits contains at most $n(n-1)/2$ non-redundant ratio edits. Thus, for ratio edits, the Fellegi-Holt method is usually feasible; see Winkler and Draper (1997).

cutting planes, and adds these to the original set of edit rules. Next, a minimal covering set of variables is determined for the new problem. Again, this solution may be infeasible, in which case more cutting planes need to be generated. In this iterative manner, the algorithm continues until it finds a feasible solution to the error localisation problem. For more details, we refer to Garfinkel et al. (1988), Ragsdale and McKeown (1996), and De Waal et al. (2011).

Algorithms for mixed integer programming

Finally, it is also possible to formulate the error localisation problem according to the Fellegi-Holt paradigm as a mixed integer programming problem; see, e.g., Riera-Ledesma and Salazar-González (2003). This type of problem can be solved by commercially available solvers.

De Waal and Coutinho (2005) compared the performance of several different algorithms for error localisation. They did not find a strong preference for one particular algorithm. Note that ‘performance’ here refers simply to computational efficiency. All of the above algorithms try to solve the same error localisation problem and hence, in theory, should find the same solution.⁵

3. Preparatory phase

The method discussed in this module is only considered appropriate for identifying random errors. Therefore, it is important to treat systematic errors, such as unit of measurement errors, before applying this method. Methods for detecting and treating systematic errors are discussed in the method module “Statistical Data Editing – Deductive Editing”.

In addition, automatic editing is usually applied in combination with a form of selective editing: the most influential errors are edited manually by subject-matter experts, while the other, non-influential errors are resolved automatically. Selective editing and manual editing are discussed in the theme module “Statistical Data Editing – Selective Editing” and the method module “Statistical Data Editing – Manual Editing”, respectively. We refer to “Statistical Data Editing – Main Module” for a discussion on how to combine different editing methods into one editing process. See also Pannekoek and De Waal (2005) for suggestions on how to set up an automatic editing strategy in practice.

4. Examples – not tool specific

To illustrate the method of Fellegi and Holt discussed in Section 2.4, we work out an example based on Fellegi and Holt (1976). In this example, there are four numerical variables. We do not use different reliability weights. The original set of edit rules consists of two edits:

$$x_1 - x_2 + x_3 + x_4 \geq 0 \tag{3}$$

and

$$-x_1 + 2x_2 - 3x_3 \geq 0. \tag{4}$$

⁵ In practice, the error localisation problem according to the Fellegi-Holt paradigm may have several equivalent optimal solutions, particularly if many variables have the same reliability weight. When this occurs, different implementations of these algorithms may differ in the way they choose between equivalent solutions.

By a repeated application of Fourier-Motzkin elimination, it is possible to derive the following essentially new implied edits from (3) and (4):

$$x_2 - 2x_3 + x_4 \geq 0, \quad (5)$$

$$x_1 - x_3 + 2x_4 \geq 0, \quad (6)$$

and

$$2x_1 - x_2 + 3x_4 \geq 0. \quad (7)$$

It is not possible to generate more essentially new implied edits from (3)–(7), so these five edit rules together constitute a complete set of edits. This means that we can now solve the error localisation problem for any record by solving an appropriate set-covering problem.

Consider the record $(x_1, x_2, x_3, x_4) = (3, 4, 6, 1)$. By checking the edit rules (3)–(7), it is seen that this record fails edits (4), (5), and (6). Thus, in order to solve the error localisation problem, we have to find the minimal subset of variables that ‘covers’ these three edit rules. By inspection, we see that the variable x_3 is involved in edit rules (4), (5), and (6). Thus, in this example, x_3 can be imputed to satisfy all the edit rules. Since $\{x_3\}$ is the only single-variable set with this property, changing the value of x_3 is in fact the optimal solution to the error localisation problem for this record. [Note that the single-variable sets $\{x_1\}$ and $\{x_2\}$ cover the original failed edit (4), but not the implied failed edits (5) and (6).] A consistent record can be obtained by imputing, for instance, the value $x_3 = 1$.

5. Examples – tool specific

The R package `editrules`, which can be downloaded for free at <http://cran.r-project.org>, contains an implementation of the branch-and-bound algorithm of De Waal and Quere (2003). To illustrate the use of `editrules` for automatic editing, we work out the example from Section 4 in R code.⁶

First, we load the package:

```
> library(editrules)
```

Next, we create an object of type “editmatrix” containing the two original edit rules:

```
> E <- editmatrix(c("x1-x2+x3+x4 >= 0", "-x1+2*x2-3*x3 >= 0"))
```

We also have to read in the record that we want to edit as a data frame:

```
> x <- data.frame(x1 = 3, x2 = 4, x3 = 6, x4 = 1)
```

Now, the error localisation problem is solved to optimality by giving the following command:

```
> le <- localizeErrors(E, x)
```

This command runs the branch-and-bound algorithm to solve the error localisation problem and stores the results in a new object called `le`. The results can be inspected by calling attributes of this object.

⁶ Version 2.5 of the `editrules` package was used to run the code in this example.

```
> le$status
  weight degeneracy user system elapsed maxDurationExceeded
1      1           1 0.05      0      0.13                FALSE
```

The attribute `le$status` contains background information on the performance of the algorithm. In this example, an optimal solution has been found with the sum of the reliability weights equal to 1 (as can be seen in the column `weight`). Since we have not specified the reliability weights in this example, R has used the default choice: all weights equal to 1. Other reliability weights can be specified by providing the function `localizeErrors` with an optional argument `weight`. The entry ‘1’ in the column `degeneracy` in `le$status` shows that the optimal solution is unique.

To see which variables have to be changed according to the optimal solution, we inspect the attribute `le$adapt`.

```
> le$adapt
      x1      x2      x3      x4
1 FALSE FALSE TRUE  FALSE
```

This command prints a boolean data frame with the value ‘TRUE’ for variables that have to be changed, and the value ‘FALSE’ for the other variables. In this example, the optimal solution is to change only the value of variable x_3 . This solution is identical to the one found in Section 4 by applying the method of Fellegi and Holt.

We refer to De Jonge and Van der Loo (2011) for more details on the `editrules` package.

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Banff Support Team (2008), Functional Description of the Banff System for Edit and Imputation. Technical Report, Statistics Canada.
- Boskovitz, A., Goré, R., and Wong, P. (2005), Data Editing and Logic. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ottawa.
- Casado Valero, C., Del Castillo Cuervo-Arango, F., Mateo Ayerra, J., and De Santos Ballesteros, A. (1996), Quantitative Data Editing: Quadratic Programming Method. Presented at the COMPSTAT 1996 Conference, Barcelona.
- De Jonge, E. and van der Loo, M. (2011), Manipulation of Linear Edits and Error Localization with the Editrules Package. Discussion Paper 201120, Statistics Netherlands, The Hague.
- De Waal, T. (1996), CherryPi: a Computer Program for Automatic Edit and Imputation. Working Paper, UN/ECE Work Session on Statistical Data Editing, Voorburg.
- De Waal, T. (2003), *Processing of Erroneous and Unsafe Data*. PhD Thesis, Erasmus University, Rotterdam.

- De Waal, T. (2005), SLICE 1.5: a Software Framework for Automatic Edit and Imputation. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ottawa.
- De Waal, T. and Coutinho, W. (2005), Automatic Editing for Business Surveys: an Assessment for Selected Algorithms. *International Statistical Review* **73**, 73–102.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- De Waal, T. and Quere, R. (2003), A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics* **19**, 383–402.
- Di Zio, M., Guarnera, U., and Luzi, O. (2005), Improving the Effectiveness of a Probabilistic Editing Strategy for Business Data. Report, ISTAT, Rome.
- Fellegi, I. P. and Holt, D. (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.
- Fillion, J. M. and Schiopu-Kratina, I. (1993), On the Use of Chernikova's Algorithm for Error Localization. Report, Statistics Canada.
- Freund, R. J. and Hartley, H. O. (1967), A Procedure for Automatic Data Editing. *Journal of the American Statistical Association* **62**, 341–352.
- Garfinkel, R. S., Kunnathur, A. S., and Liepins, G. E. (1988), Error Localization for Erroneous Data: Continuous Data, Linear Constraints. *SIAM Journal on Scientific and Statistical Computing* **9**, 922–931.
- Ghosh-Dastidar, B. and Schafer, J. L. (2003), Multiple Edit/Multiple Imputation for Multivariate Continuous Data. *Journal of the American Statistical Association* **98**, 807–817.
- Hoogland, J. and Smit, R. (2008), Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.
- Kovar, J. and Whitridge, P. (1990), Generalized Edit and Imputation System; Overview and Applications. *Revista Brasileira de Estadística* **51**, 85–100.
- Little, R. J. A. and Smith, P. J. (1987), Editing and Imputation of Quantitative Survey Data. *Journal of the American Statistical Association* **82**, 58–68.
- Nemhauser, G. L. and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization*. John Wiley & Sons, New York.
- Nordbotten, S. (1963), Automatic Editing of Individual Statistical Observations. In: *Conference of European Statisticians Statistical Standards and Studies No. 2*, United Nations, New York.
- Pannekoek, J. and de Waal, T. (2005), Automatic Edit and Imputation for Business Surveys: the Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* **21**, 257–286.
- Ragsdale, C. T. and McKeown, P. G. (1996), On Solving the Continuous Data Editing Problem. *Computers & Operations Research* **23**, 263–273.
- Riera-Ledesma, J. and Salazar-González, J. J. (2003), New Algorithms for the Editing and Imputation Problem. Working Paper, UN/ECE Work Session on Statistical Data Editing, Madrid.

- Sande, G. (1978), An Algorithm for the Fields to Impute Problems of Numerical and Coded Data. Technical Report, Statistics Canada.
- Scholtus, S. (2013), Automatic Editing with Hard and Soft Edits. *Survey Methodology* **39**, 59–89.
- Todero, T. A. (1999), Overview and Evaluation of the AGGIES Automated Edit and Imputation System. Working Paper, UN/ECE Work Session on Statistical Data Editing, Rome.
- Williams, H. P. (1986), Fourier's Method of Linear Programming and Its Dual. *The American Mathematical Monthly* **93**, 681–695.
- Winkler, W. E. and Draper, L. A. (1997), The SPEER Edit System. In: *Statistical Data Editing*, Volume 2: *Methods and Techniques*, United Nations, Geneva.

Specific section

8. Purpose of the method

Localising errors in microdata without human intervention

9. Recommended use of the method

1. The method should be used for error localisation in microdata containing only random errors. Any systematic errors that may occur in the original microdata have to be resolved beforehand, using deductive editing methods (see the method module “Statistical Data Editing – Deductive Editing”).
2. If it is known beforehand that certain variables contain more errors than others, then this information should be included in the form of reliability weights (see item 14).
3. The quality of the error localisation strongly depends on the specification of the edit rules. The set of edit rules should be sufficiently powerful to detect the majority of errors, but not so strict that the method results in overedited data.

10. Possible disadvantages of the method

1. In general, it is not possible to construct a set of edit rules that always leads to the correct solution. Thus, the edited data may still contain some errors, although the edited records are consistent with the edit rules. For this reason, automatic editing should not be applied to crucial records, e.g., records belonging to very large businesses. In addition, the quality of automatic editing is lower for records that contain many errors. Both disadvantages can be circumvented by always using automatic editing in combination with a form of selective editing. We refer to “Statistical Data Editing – Main Module” for a discussion on how to incorporate automatic editing in an overall editing strategy.

11. Variants of the method

1. The original method of Fellegi and Holt as described in Section 2.4.
2. Other methods as described in Section 2.5. These methods find the same solution as the original method of Fellegi and Holt, but they use different search algorithms. Examples include:
 - 2.1 Algorithms based on vertex generation;
 - 2.2 Algorithms based on branch-and-bound;
 - 2.3 Algorithms based on cutting planes;
 - 2.4 Algorithms based on (mixed) integer programming.

12. Input data

1. A data set containing unedited microdata.

13. Logical preconditions

1. Missing values
 1. Allowed; they will be considered as erroneously missing, i.e., available for imputation.
2. Erroneous values
 1. Allowed; in fact, the object of this method is to decide which values in a record are erroneous.
 2. It is assumed that the data contain only random errors; systematic errors should be removed beforehand by means of deductive editing.
3. Other quality related preconditions
 1. n/a
4. Other types of preconditions
 1. It is assumed that all edit rules may be interpreted as hard edit rules.

14. Tuning parameters

1. A collection of edit rules for the microdata at hand.
2. A set of reliability weights may be provided for the variables in the data set. By default, all reliability weights are equal to 1.
3. A maximum number of variables to impute may be set to reduce the computational workload. The error localisation problem will not be solved for records that cannot be imputed consistently by changing at most the specified maximum number of variables.

15. Recommended use of the individual variants of the method

1. For variant 1 (the original method of Fellegi and Holt), most of the work lies in the generation of a complete set of edits. Once this complete set is available, the error localisation problem can be solved for any record in a straightforward manner. If the complete set of edits is too large to be generated, this variant of the method cannot be used.
2. For the other variants, the work lies in solving a separate error localisation problem for each individual record. In this case, it is usually necessary to specify a maximum number of variables to impute (see item 14), unless the data set contains few variables (say less than 10).

16. Output data

1. For each record in the microdata, the method attempts to yield a list of variables that can be imputed to obtain a consistent record with respect to the edit rules. For some records, the method may not return such a list, because it could not find a feasible solution to the error localisation problem.

17. Properties of the output data

1. For each record for which the method returns a solution, the variables listed in the solution can be imputed so that the resulting record is consistent with respect to the edit rules. Moreover, they constitute the smallest (weighted) set of variables that has this property.
2. The original values of the variables that are listed for imputation have to be considered as erroneous in all further processing. The natural next step is to impute new values for these variables by means of some imputation method. It should be noted that the imputation step is not a part of the error localisation method itself.
3. For some records, the method may not find a solution. These records have to be processed interactively by subject-matter experts (see the method module “Statistical Data Editing – Manual Editing”).

18. Unit of input data suitable for the method

Incremental processing by record

19. User interaction - not tool specific

1. Ideally, there is no user interaction other than setting parameters and reading in input data at the beginning, and processing output data at the end.

20. Logging indicators

1. The number of records for which the method found/did not find a solution.
2. The computing time per record.

21. Quality indicators of the output data

1. The quality of automatic editing can be assessed in a simulation study. This requires a data set that has been interactively edited by experts to a point where the edited data may be considered error-free. In the simulation study, the original data are edited again using automatic editing. The quality of automatic editing may then be measured in terms of the similarity of the automatically edited data to the interactively edited data.

22. Actual use of the method

1. The method is used at Statistics Netherlands in the production process for structural business statistics. This application uses the tool SLICE, which contains an implementation of the branch-and-bound algorithm of De Waal and Quere (2003). See Hoogland and Smit (2008) for more details.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Micro-Fusion – Main Module
2. Statistical Data Editing – Main Module

3. Statistical Data Editing – Selective Editing
4. Imputation – Main Module
5. Macro-Integration – Main Module

24. Related methods described in other modules

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Statistical Data Editing – Deductive Editing
3. Statistical Data Editing – Manual Editing

25. Mathematical techniques used by the method described in this module

1. Fourier-Motzkin elimination

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1. GEIS
2. Banff
3. CherryPi
4. AGGIES
5. SLICE
6. R package `editrules`

28. Process step performed by the method

Statistical data editing

Administrative section

29. Module code

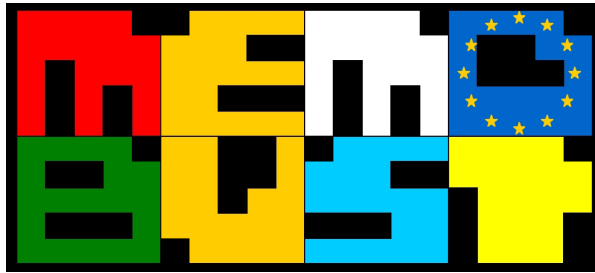
Statistical Data Editing-M-Automatic Editing

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	20-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	20-04-2012	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.2.1	16-07-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.3	04-09-2013	minor improvements	Sander Scholtus	CBS (Netherlands)
0.3.1	09-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:11



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Manual Editing

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Introduction and historical notes	3
2.2 The use of recontacts	4
2.3 Potential problems	5
3. Preparatory phase	6
3.1 The editing staff.....	6
3.2 Editing instructions.....	6
3.3 Error messages	7
3.4 Efficient edit rules for manual editing.....	7
4. Examples – not tool specific.....	9
5. Examples – tool specific.....	10
6. Glossary.....	10
7. References	10
Specific section.....	12
Interconnections with other modules.....	14
Administrative section.....	15

General section

1. Summary

In manual editing, records of microdata are checked for errors and, if necessary, adjusted by a human editor, using expert judgement. Nowadays, the editor is usually supported by a computer program in identifying data items that require closer inspection – in particular combinations of values that are inconsistent or suspicious. Moreover, the computer program enables the editor to change data items interactively, meaning that the automatic checks that identify inconsistent or suspicious values are immediately rerun whenever a value is changed. This modern form of manual editing is often referred to as ‘interactive editing’.

If organised properly, manual/interactive editing is expected to yield high quality data. However, it is also time-consuming and labour-intensive. Therefore, it should only be applied to that part of the data which cannot be edited safely by any other means, i.e., some form of selective editing should be applied (see “Statistical Data Editing – Selective Editing”). Furthermore, it is important to use efficient edit rules and to draw up detailed editing instructions in advance.

2. General description of the method

2.1 Introduction and historical notes

Manual editing is the traditional way to perform data editing. Other data editing methods, in particular automatic editing techniques, did not emerge until the 1960s, and their application has only become widespread from the 1980s onward. Even today, practically all surveys at statistical offices and elsewhere include some form of manual editing. Manual editing is in fact widely viewed as an essential part of any data editing process.

Ideally, a person who performs manual editing – an *editor* – should be an expert who has extensive knowledge of the survey subject, the survey population, and the kind of errors that are likely to occur in the survey data. If necessary, he or she may recontact a respondent to check whether a suspicious value is correct, or to obtain a new value for a data item that was originally missing or incorrect. The editor may compare a survey unit’s data to reference data, such as data on the same unit from a previous survey or from an external register, or data on similar units. Finally, he or she may have access to other sources of information, for instance through internet searches.

In its ideal form, manual editing is expected to yield high quality data. In particular, it should lead to better results than automatic editing. However, it should be clear that the quality of manual editing depends strongly on the competence and training of the available editors. In certain less than ideal situations, the quality of manually edited data need not be significantly higher than that of automatically edited data, and it may even be lower (EDIMBUS, 2007).

Traditionally, manual editing was performed directly on the original paper questionnaires. Later, mainframe computers were used to check the data for inconsistencies and other violations of edit rules. To this end, the information on the questionnaires first had to be keyed in by typists. A list of edit failures identified by the computer was printed out on paper and used by the editors as a guide for making manual adjustments on the original questionnaires. When all questionnaires had been edited, the adjusted data were re-entered into the mainframe computer by typists and the edit checks were run

again, to see the effect of the proposed adjustments on the edit failures. Often, the automated checks revealed that the adjusted data still failed some of the edit rules, and another round of manual editing was required. It was not unusual that five, ten, or even more iterations of automatic checking and manual adjusting were needed before all questionnaires were considered sufficiently edited (Granquist, 1997; Van de Pol, 1995).

The advent of the microcomputer in the 1980s made it possible to integrate automatic checking and manual treatment of errors, thereby improving the data editing process in several ways (Bethlehem, 1987). From now on, the information on the questionnaires had to be keyed in only once.¹ After that, all adjustments could be made by the editors directly on the captured data. This obviously benefited the efficiency and timeliness of the editing process. A second improvement was that the editors could now get immediate feedback on the adjustments they made, because the automatic edit checks could be rerun instantaneously whenever the value of a data item was changed. This made it much easier for them to find adjustments that satisfied the edit rules. In addition, each record/questionnaire could now be edited separately, by one editor, until all violations of edit rules had been either removed or explained. This improved form of manual editing is called *interactive editing*.

Interactive editing requires a survey-processing system that provides the above-mentioned interaction between automated checks and manual adjustments. Well-known examples of survey-processing systems are *Blaise* (see, e.g., Blaise, 2002) and *CSPro* (see, e.g., CSPro, 2008). Pierzchala (1990) discusses general requirements of computer systems for interactive editing.

In today's statistical practice, interactive editing has effectively replaced all older forms of manual editing. Hence, the terms 'manual editing' and 'interactive editing' have become more or less interchangeable. In the remainder of this module, they shall be used as synonyms.

2.2 The use of recontacts

In the previous subsection, possible actions were listed that an editor may take when confronted with a record that requires review. One of these possible actions is recontacting the respondent. At first glance, a recontact may appear to be the natural way of obtaining better values for data items that were reported erroneously during the original field work, as well as items that were originally missing. Actually, depending on the survey, it may not be possible to contact the original respondents. For instance, if an external register is used as a data source and questions are raised about the quality of the incoming data, then the statistical office can usually only contact the supplier of the data set. Direct contact with the individual entities in the register is usually not possible in this case.

However, even when recontacts are possible, this approach can be considered problematic for several reasons. First of all, recontacts clearly increase the burden on respondents, whereas many statistical institutes are trying to reduce the response burden. In addition, recontacts tend to slow down the editing process and can therefore adversely influence the timeliness of statistics. Finally, if one considers that a respondent was not able to give a correct answer in the original survey – supposedly while filling in a meticulously designed questionnaire or talking to a highly qualified interviewer –,

¹ A more recent development is that data often arrive at the statistical office already in digital form, so that no keying is necessary at all. This is true for nearly all registers and for electronic questionnaires. For a discussion of the implications of electronic data collection for the editing process, see the theme module "Questionnaire Design – Editing During Data Collection".

then it is not at all obvious that he/she will give the correct response when talking to an editor. According to EDIMBUS (2007): "...respondents' ability to report should not be overestimated. In fact, if the structure of the questions does not fit their understanding, no amount of badgering will get the 'correct' answers out of them."

Following Granquist (1997), if recontacts are used during interactive editing, their main purpose should be to reveal problems that *cause* respondents to give erroneous answers, rather than merely correcting the individual errors that occurred. When used this way, recontacts can provide important insights into respondents' behaviour – in particular their ability to understand the concepts and definitions used in the survey. They may also reveal differences between what is asked in the survey and what kind of information is readily available in the survey units' accounting systems. These insights may be used as a basis for improvements at the data collection stage in subsequent surveys (see, e.g., Hartwig, 2009; Svensson, 2012).

2.3 Potential problems

There are several potential problems associated with interactive editing. The most important of these are the risks of *overediting* and *creative editing*.

According to Granquist (1995), overediting occurs when "the share of resources and time dedicated to editing is not justified by the resulting improvements in data quality." Manual editing is in fact a very labour-intensive and time-consuming activity, even in its modern, interactive form. Moreover, statistical output is typically affected by all kinds of errors (Bethlehem, 2009), including sampling error, selective unit non-response, coverage errors, measurement errors, etc. Only a subset of these can be treated during data editing: in particular, measurement and processing errors and, to a lesser extent, errors in the survey frame. Therefore, as soon as the data have been edited to a point where the influence of the latter types of errors on the statistical output is negligible compared to other sources of error (e.g., the sampling variance), manual editing should be stopped to prevent overediting. This notion – which was suggested already by Nordbotten (1955) – has received much attention since the 1980s. It has led to the development of methods for selective editing (see the theme module "Statistical Data Editing – Selective Editing") and macro-editing (see the theme module "Statistical Data Editing – Macro-Editing").

Another aspect of overediting is that if the editing process is continued too long, it may actually start to do more harm than good. In general, not all values that appear to be implausible are also incorrect. Hence, replacing all unusual combinations of values by more plausible ones would lead to a data set that does not reflect the natural variability of characteristics in the population. Overediting may therefore adversely influence the quality of the statistical output. An important part of the 'art' of manual editing is understanding which implausible values to adjust and which to leave as they are. This requires expert judgement and, in some cases, a recontact.

A second potential problem is the risk of creative editing: editors inventing their own, often highly subjective, editing procedures. Creative editing often involves complex adjustments of reported data items, done for the sole purpose of making the data consistent with a set of edit rules. Granquist (1995) remarks that creative editing may "hide serious data collection problems and give a false impression of respondents' reporting capacity."

To reduce the risk of overediting and creative editing, it is important to design efficient edit rules and to provide the editors with good editing instructions. These issues are discussed in the next section.

3. Preparatory phase

In this section, several issues will be discussed that are related to the design of manual editing. These are: the desired characteristics of the editing staff (Section 3.1); the use of editing instructions to rationalise the manual editing process (Section 3.2); the design of error messages (Section 3.3); the design of efficient edit rules for manual editing (Section 3.4).

3.1 The editing staff

As mentioned in Section 2.1, the quality of manual editing strongly depends on the competence of the individual editors that are involved. A good editor should have the following characteristics:

- He/she has a large knowledge of the survey subject and of survey methodology. Since most of this knowledge is rather specialised, it has to be acquired through experience and training.
- He/she is communicative and responsive. This is particularly important if recontacts are used. Granquist (1995) remarked that if recontacts are done by telephone, “the editors also become telephone interviewers, needing adequate training and monitoring as in regular telephone interview surveys.”
- He/she is responsible and able to work accurately.
- Preferably, he/she should have an analytical mind, with an interest in problem-solving.

3.2 Editing instructions²

Editing instructions are an important aid in rationalising the manual editing process. They should contain at least the following components:

- A description of the purpose of the survey and the intended statistical output. In addition, the data collection phase and relevant data processing steps prior to editing should be briefly described.
- If relevant, instructions on the order in which the selected records should be treated. If manual editing is used in combination with selective editing (see “Statistical Data Editing – Selective Editing”), then an explanation is needed about the selection criteria and their interpretation. If manual editing is used in combination with macro-editing (see “Statistical Data Editing – Macro-Editing”), then detailed analysis instructions are needed regarding the selection of individual records that need further review.
- An overview of the types of errors that can occur in the data. Common errors in business surveys include classification errors with respect to NACE code or size class (i.e., errors in the survey frame), measurement errors, and processing errors.
- Suggestions about additional sources of information – such as auxiliary registers, sector organisations, and the internet – which should be consulted when following up data that have

² This subsection is to a large extent based on Hoogland et al. (2011).

been flagged by edit rules (see below). For example, many businesses nowadays have websites that contain relevant information for verifying potential NACE code errors.

- For each common type of error, an indication of how the error can be treated. Deterministic correction rules may often be specified for treating systematic errors (see also ‘Deductive Editing’). Clear instructions on this point can prevent the occurrence of creative editing.
- Instructions on how to log the editing actions taken during interactive editing. The survey-processing system should provide a comments field for this. Editors should be encouraged to provide details about the reasons for the adjustments they make. This information can be useful for improving the data collection process as well as the editing process itself.
- Instructions on specific follow-up actions that may be needed for certain types of errors. In particular, in case a NACE code or size class error is detected, it should be clear whether and how this must be communicated to the administrator of the survey frame.

3.3 *Error messages*

As mentioned in the main theme module, an important technique for finding errors in microdata is the inspection of data items that fail *edit rules*. Edit rules (edits for short) describe restrictions that should be satisfied by the data. Edits can be hard (meaning that they have to hold by definition, so that any failure corresponds to an error in the data) or soft (meaning that they are expected to hold for most survey units, but they can sometimes be failed by correct data items).

When edit rules are implemented in a computer system, an error message has to be associated with each edit rule. This message contains the information that the computer system gives to the editor about the unit and variables that are flagged by the edit rule as being (suspected to be) in error. The purpose of the error message is to give sufficient information for a rational follow-up of error flags. It also forms a basis for (process) data about the data collection and production processes.

The content of an error message generally consists of:

- Identifying properties of the flagged unit.
- The name of the flagged variable(s). For the purpose of manual editing, this should be a descriptive name rather than a technical one; e.g., not *TURNOVE100000* but *Total net turnover from domestic sales*.
- The code of the edit rule that was failed.
- A verbal description of the edit rule that was failed or, equivalently, a verbal description of the suspected error.
- If relevant and available, suggestions for auxiliary data that may be consulted in a follow-up of the error flag.

3.4 *Efficient edit rules for manual editing*

Typically, a large part of the work done during manual editing concerns the follow-up of soft edit failures. For this reason, it is important to formulate soft edit rules that are as efficient as possible. Here, an edit rule is considered efficient to the extent that it detects suspected errors that turn out to be

actual errors during manual follow-up, and inefficient to the extent that it detects suspected errors that turn out to be correct. (A measure of efficiency known as the hit rate will be introduced below.)

According to Norberg (2011), most edits that are used in practice consist of three components: an *edit group*, a *test variable*, and an *acceptance region*. The edit group defines the subset of the units to which the edit should be applied. The test variable is a known function of the observed variables that is evaluated by the edit. Finally, the acceptance region describes for which values of the test variable the edit will be satisfied. (Equivalently, one could define a *rejection region* that describes for which values of the test variable the edit will be failed.) Using these components, an edit may be written in one of the general forms

if (*unit* \in *edit group*) **then** (*test variable* \in *acceptance region*)

or

if (*unit* \in *edit group* **and** *test variable* \notin *acceptance region*) **then** *error*.

Both formulations are equivalent. Human editors often find it slightly easier to work with the first formulation (Van de Pol, 1995). In a computer implementation, the second formulation can easily be extended to associate a unique error code and error message to each edit rule.

For a simple example, consider the following edit rule:

if *Size class* = 'small' **then** $0 \leq \text{Number of employees} < 10$.

For this edit, the edit group can be defined as "all units for which *Size class* = 'small'". The test variable is identical to one of the observed variables, *Number of employees*. The acceptance region consists of the interval [0, 10). A computer implementation of this edit could further specify the following actions:

if (*Size class* = 'small' **and** (*Number of employees* < 0 **or** *Number of employees* \geq 10))
then (*error_code_E1* := "failed";
error_message_E1 := "The number of employees does not match the size class.")

The first statement in the then-part assigns the error code "failed" to the current record for this edit (identified here by E1). The second statement gives an error message describing the nature of the current edit failure to the human editor. Of course, the precise implementation of these actions will depend on the survey-processing system.

To give another example, consider the following conditional ratio edit:

if (*Economic activity* = X **and** *Size class* = 'medium')
then $a < \text{Total turnover} / \text{Number of employees} < b$.

Here, the edit group consists of "all medium-sized units with *Economic activity* X", the test variable is defined as the ratio of the observed variables *Total turnover* and *Number of employees*, and the acceptance region is given by the interval (a,b).

Norberg (2011) notes that, for the editing to be efficient, one should choose edit groups that are homogeneous with respect to the test variable. In some cases, the choice of an edit group may be natural (e.g., the first example given above). If this is not the case, suitable edit groups may be derived from an analysis of previously edited data. Norberg (2012) suggests to use classification or regression trees for this. In addition, the acceptance region should reflect the natural variability of the test

variable within the edit group (Norberg, 2012). Again, previously edited data may be analysed (e.g., using box plots) to find suitable acceptance regions. It may be worthwhile to transform a test variable so that its distribution becomes more amenable to summary in the form of an acceptance region (e.g., so that the transformed test variable is approximately normally distributed, or at least symmetrical). Moreover, in repeated surveys, the acceptance regions should be regularly updated.

Outlier detection techniques are often used in the construction of soft edit rules. We refer to EDIMBUS (2007) for a discussion of outlier detection in the context of statistical data editing. Methods that may be used to construct soft edit rules in repeated surveys are discussed in the theme module “Statistical Data Editing – Editing for Longitudinal Data”.

At the design stage, it is useful to assess the efficiency and effectiveness of a proposed set of edits E by means of simulation. This requires historical data that have been fully edited, as well as the original, unedited version of the same data set. Interesting indicators for an edit $e \in E$ include the *failure rate* (the proportion of records in the unedited data that fail edit e) and the *hit rate* (the proportion of edit failures with respect to e in the unedited data that are associated with adjustments in the edited data). Note that for all hard edit rules, the hit rate should be 1. These indicators are local, i.e., defined for one edit at a time. Similar global indicators can be defined for the set of edits E as a whole. It is also interesting to assess to what extent the edits are ‘overlapping’, in the sense that the same error is often detected by multiple edits. Ideally, there should be as little overlap as possible between the edits.

Furthermore, making the assumption that the edited historical data do not contain any errors, one can evaluate the *missed error rate* (the proportion of errors in the original data that were not flagged by any edits in E) and an estimate of the measurement bias due to untreated errors if editing were based on E . See EDIMBUS (2007) and Silva et al. (2008) for formal definitions of these and other indicators.

The Office for National Statistics in the United Kingdom and Southampton University have developed a tool called Snowdon-X which “can be used to understand how current edits are working within the survey and also the impact on quality of any changes to the edit rules” (Skelterbery et al., 2011). Snowdon-X evaluates the indicators mentioned above as well as many other indicators. See Silva et al. (2008) for more details on Snowdon-X.

Note that the failure rate and hit rate of edits can and should be evaluated also during regular production. On the other hand, evaluating the missed error rate requires edited historical data. For repeated surveys, suitable historical data sets are available in theory, if not always in practice (Lindgren, 2012). For a one-off survey, as well as the first cycle of a survey that will be repeated, the situation is different. Often in this case, a small pilot study is conducted beforehand. The data from this study can be used to test the effects of different editing approaches, including experiments with different formulations of edit rules. In addition, experts should be consulted that have had experience with similar surveys in the past.

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bethlehem, J. G. (1987), The Data Editing Research Project of the Netherlands Central Bureau of Statistics. Report 2967-87-M1, Statistics Netherlands, Voorburg.
- Bethlehem, J. G. (2009), *Applied Survey Methods*. Wiley Series in Survey Methodology, John Wiley & Sons, New Jersey.
- Blaise (2002), *Blaise for Windows 4.5 Developer’s Guide*. Statistics Netherlands, Heerlen.
- CSPPro (2008), *CSPPro User’s Guide*, version 4.0. U.S. Census Bureau, Washington, D.C.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Granquist, L. (1995), Improving the Traditional Editing Process. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, 385–401.
- Granquist, L. (1997), The New View on Editing. *International Statistical Review* **65**, 381–387.
- Hartwig, P. (2009), How to Use Edit Staff Debriefings in Questionnaire Design. Paper presented at the 2009 European Establishment Statistics Workshop, Stockholm.
- Hoogland, J., van der Loo, M., Pannekoek, J., and Scholtus, S. (2011), *Data Editing: Detection and Correction of Errors*. Methods Series Theme, Statistics Netherlands, The Hague.
- Lindgren, K. (2012), The Use of Evaluation Data Sets when Implementing Selective Editing. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- Norberg, A. (2011), The Edit. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Norberg, A. (2012), Tree Analysis – A Method for Constructing Edit Groups. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- Nordbotten, S. (1955), Measuring the Error of Editing the Questionnaires in a Census. *Journal of the American Statistical Association* **50**, 364–369.
- Pierzchala, M. (1990), A Review of the State of the Art in Automated Data Editing and Imputation. *Journal of Official Statistics* **6**, 355–377.
- Silva, P. L. N., Bucknall, R., Zong, P., and Al-Hamad, A. (2008), A Generic Tool to Assess Impact of Changing Edit Rules in a Business Survey – An Application to the UK Annual Business Inquiry Part 2. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.

- Skentelbery, R., Finselbach, H., and Dobbins, C. (2011), Improving the Efficiency of Editing for ONS Business Surveys. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Svensson, J. (2012), Editing Staff Debriefings at Statistics Sweden. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.
- Van de Pol, F. (1995), Data Editing of Business Surveys: an Overview. Report 10718-95-RSM, Statistics Netherlands, Voorburg.

Specific section

8. Purpose of the method

Detecting and treating errors in microdata

9. Recommended use of the method

1. Because of its expensive and time-consuming nature, it is best to apply manual editing only to that part of the data where expert judgement is really needed. In other words, one should always try to use this method as part of a strategy for selective editing or macro-editing (cf. “Statistical Data Editing – Main Module”). This usually means that manual editing is only applied to units that are either very large or complex, or for which the reported data are likely to contain many and/or influential errors.
2. A survey-processing system should be used that allows real-time interaction between manual adjustments and automated checks (i.e., manual editing should be interactive editing)
3. It is important to draw up editing instructions in advance, to guide the decisions made by the editors during manual editing. This lowers the risk of overediting or creative editing. It is also important to design efficient edit rules and informative error messages.

10. Possible disadvantages of the method

1. If recontacts are used as part of manual editing, the method places additional burden on survey units that are recontacted. Recontacts may also affect the timeliness of statistical production.

11. Variants of the method

1. n/a

12. Input data

1. A data set containing unedited microdata.

13. Logical preconditions

1. Missing values
 1. Allowed.
2. Erroneous values
 1. Allowed; in fact, the object of this method is to replace erroneous values with better values.
3. Other quality related preconditions
 1. n/a
4. Other types of preconditions
 1. n/a

14. Tuning parameters

1. A collection of edit rules for the microdata at hand.

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. A data set containing edited microdata.

17. Properties of the output data

1. If manual editing has been performed correctly, the records in the output data set are consistent with all hard edit rules. In addition, all remaining soft edit failures have been explained and accepted by a subject-matter expert.

18. Unit of input data suitable for the method

Incremental processing

19. User interaction - not tool specific

1. As the term ‘interactive editing’ suggests, user interaction is needed throughout. In fact, all changes made to the data during manual/interactive editing are initiated by a human editor.

20. Logging indicators

1. Comments made by the editors to explain the adjustments they made to the data, as well as the soft edit failures that they left in.
2. If recontacts are used: comments made by the editors regarding identified problems that caused respondents to report erroneous values in the original survey.
3. Process indicators for the efficiency and effectiveness of the edit rules used in manual editing include: failure rate, hit rate, missed error rate, estimated measurement bias. See also Section 3.4 of this module, EDIMBUS (2007), and Silva et al. (2008).

21. Quality indicators of the output data

1. It is not straightforward to assess the quality of manually edited data, because in many applications the results of manual editing are actually taken as the standard by which other forms of editing are to be measured. Nordbotten (1955) suggests a way to measure the quality of regular manual editing, i.e., as it occurs in everyday statistical practice. This method takes a random sample of the original data and subjects it to a very refined form of manual editing (under ideal conditions, with near-unlimited resources). The quality of the regular editing process may then be measured in terms of the similarity of the data edited under regular conditions to the data edited under ideal conditions.

22. Actual use of the method

1. Interactive editing is used at Statistics Netherlands in many production processes, including that of the structural business statistics. The survey-processing system Blaise is used as a tool.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Questionnaire Design – Editing During Data Collection
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Statistical Data Editing – Editing for Longitudinal Data

24. Related methods described in other modules

1. Statistical Data Editing – Automatic Editing

25. Mathematical techniques used by the method described in this module

1. n/a

26. GSBPM phases where the method described in this module is used

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1. Blaise
2. CSPro

Note: These tools support interactive editing, but – by its very nature – this method relies heavily on human interaction with the tool.

3. Snowdon-X

Note: This tool can be used to evaluate the efficiency of edit rules for manual editing.

28. Process step performed by the method

Statistical data editing

Administrative section

29. Module code

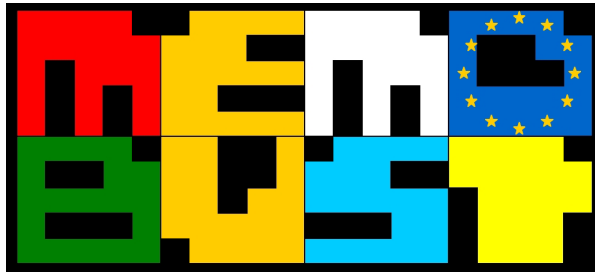
Statistical Data Editing-M-Manual Editing

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-06-2012	first version	Sander Scholtus	CBS (Netherlands)
0.2	01-03-2013	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	12-04-2013	improvements based on second Swedish review	Sander Scholtus	CBS (Netherlands)
0.4	11-11-2013	minor improvements based on final Swedish review	Sander Scholtus	CBS (Netherlands)
0.4.1	26-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:12



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Macro-Editing

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction to macro-editing.....	3
2.2 The aggregate method	4
2.3 The distribution method	6
3. Design issues	8
4. Available software tools.....	8
5. Decision tree of methods	8
6. Glossary.....	8
7. References	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

In most business surveys, it is reasonable to assume that a relatively small number of observations are affected by errors with a significant effect on the estimates to be published (so-called influential errors), while the other observations are either correct or contain only minor errors. For the purpose of statistical data editing, attention should be focused on treating the influential errors. *Macro-editing* (also known as *output editing* or *selection at the macro level*) is a general approach to identify the records in a data set that contain potentially influential errors. It can be used when all the data, or at least a substantial part thereof, have been collected.

Macro-editing has the same purpose as selective editing (see “Statistical Data Editing – Selective Editing”): to increase the efficiency and effectiveness of the data editing process. This is achieved by limiting the costly manual editing to those records for which interactive treatment is likely to have a significant effect on the quality of the estimates. The main difference between these two approaches is that selective editing selects units for manual follow-up on a record-by-record basis, whereas macro-editing selects units by considering all the data at once. It should be noted that in macro-editing all actual adjustments to the data take place at the *micro* level (i.e., for individual units), not the *macro* level. Methods that perform adjustments at the macro level are discussed in the topic “Macro-Integration”.

2. General description

2.1 Introduction to macro-editing

Macro-editing is a general approach to identify potentially influential errors in a data set for manual follow-up. It can be used when all the data, or at least a substantial part thereof, have been collected. In addition, the method is particularly effective when it is applied to data that contain only a limited number of large errors. Given these conditions, macro-editing is typically applied towards the end of a data editing process. At that stage, the errors that one expects to find in the data are either remaining errors that ‘slipped through’ previous editing efforts or errors that were actually introduced during data processing (processing errors). Possible sources of processing errors include automated data handling (e.g., loading the wrong data set, running an application with the wrong set of parameters, a bug in the software) as well as wrong decisions made by editors during manual editing. Macro-editing may succeed in finding these errors by examining the data from a macro rather than a micro level perspective – in other words, looking at the whole data set instead of one record at a time.

Macro-editing proceeds by computing aggregate values from a data set and systematically checking these aggregates for suspicious values and inconsistencies. The following types of checks are typically used:

- Internal consistency checks. In most business surveys, the definitions of the survey variables imply that the aggregated data should satisfy certain logical or mathematical restrictions. For instance, in each stratum, total net turnover (say X) should equal the sum of total net turnover from domestic sales (X_1) and total net turnover from foreign sales (X_2); i.e., it should hold that $X = X_1 + X_2$. In addition, based on subject-matter knowledge the fraction of total net

turnover from domestic sales may be expected to lie between certain bounds; i.e., $a < X_1 / X < b$ for certain constants a and b . These restrictions are the macro-level equivalents of edit rules that were used during micro-editing (see “Statistical Data Editing – Main Module”). Like edit rules, they may be either hard restrictions (identifying erroneous aggregates with certainty, such as the first example given above) or soft restrictions (identifying suspicious aggregates that may occasionally be correct, such as the second example).

- Comparisons with other statistics. It may be possible to compare aggregates to similar estimates from other data sources. If large differences occur, the corresponding aggregates are identified as suspicious. Such comparisons can be useful, if only to promote coherence between different statistical outputs. On the other hand, the comparability of aggregates from different sources is often affected in practice by conceptual and operational differences (e.g., different target populations, differences in variable definitions, different reference periods). It is important to be aware of these differences when they exist.
- Comparisons with previously published statistics. In repeated surveys, one can compare current aggregates to a time series of previously published values. If a sufficiently long time series is available, one may apply time series analysis to identify possible trend discontinuities and hence suspicious aggregates.
- Other quality information about the statistical process so far. For instance, a non-response analysis provides information on aggregates that have a high risk of being biased. If estimates of sampling errors are available, these may also be incorporated in the macro-editing procedure (see Section 2.2).

It should be noted that in macro-editing all actual adjustments to the data take place at the *micro* level, not the *macro* level. Therefore, after one has found suspicious aggregates by any of the above means, the next step is to identify individual units that contribute to these aggregates and may require further editing. The next two subsections describe two generic approaches to do this. The *aggregate method* (Section 2.2) proceeds by ‘drilling down’ from suspicious aggregate values to lower-level aggregates and, eventually, individual units. The *distribution method* (Section 2.3) examines the distribution of the microdata to identify outliers and other suspicious values. In practice, the two methods are often applied together.

2.2 The aggregate method

Given a data set that requires macro-editing, the aggregate method starts by calculating estimates of aggregates at the highest level of publication based on the current data (Granquist, 1994). These provisional publication figures are checked for plausibility and consistency, as discussed in Section 2.1. If an aggregate is identified as suspicious, the next step is to zoom in on the cause of the suspicious value by examining the lower-level aggregates that contribute to the suspicious aggregate. This procedure is sometimes called ‘drilling down’. In this way, macro-editing proceeds until the lowest level of aggregation is reached, i.e., the individual units. Finally, the units that have been identified as the most important contributors to a suspicious provisional publication figure are submitted to manual follow-up (see “Statistical Data Editing – Manual Editing”).

In practice, checking for suspicious aggregates is often implemented by means of score functions, similar to those that are used at the micro level in selective editing (see “Statistical Data Editing – Selective Editing”). In macro-editing, the score function is applied at the aggregate level (e.g., Farwell and Schubert, 2011). In practice, relatively simple score functions are often used, such as:

$$S_j = \frac{\hat{T}_{y_j} - \tilde{T}_{y_j}}{\tilde{T}_{y_j}}, \quad (1)$$

where \hat{T}_{y_j} is the estimated total of variable y_j based on the unedited data, and \tilde{T}_{y_j} is a corresponding anticipated (or predicted) total value. This score function measures the relative deviation from the anticipated value. Possible sources of anticipated values are: estimates from different data sources, such as a register or a different survey, or the value of the same total in a previous survey cycle – possibly corrected for development over time using a time series model (see also Section 2.1).

Comparisons based on ratios of aggregated values are also used, such as:

$$S_{jk} = \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right) / \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}}, \quad (2)$$

using notation similar to (1).

Since macro-editing is applied when all, or nearly all, data are available, there is no need to set a threshold value on the score function in advance. Instead, the aggregates can be put in order of suspicion by sorting on the absolute value of S_j or S_{jk} . In order to prevent the introduction of bias, it is important to treat large positive and large negative deviations from the anticipated values with equal care.

If the estimates are based on a sample of the population, as is often the case in business surveys, a natural amount of variation in the aggregates is expected due to sampling error. From a theoretical point of view, it is good to take this inaccuracy of the estimated aggregates into account in the score function. Thus, instead of (1), one could use

$$S'_j = \frac{\hat{T}_{y_j} - \tilde{T}_{y_j}}{se(\hat{T}_{y_j} - \tilde{T}_{y_j})},$$

and instead of (2), one could use

$$S'_{jk} = \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right) / se \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right),$$

where $se(.)$ indicates the standard error of an estimate. In these alternative score functions, deviations from the anticipated values are only seen as suspicious if they are large compared to the associated sampling error. This refinement is particularly important if there are large differences in accuracy between different aggregates.

For the final step in the aggregate method, the so-called ‘drilling down’ from suspicious aggregates to contributing individual units, the same score functions on the micro level can be used as in selective

editing (see “Statistical Data Editing – Selective Editing”). The main difference is that, again, there is no need to set a threshold value in advance here, because the score function can be computed for all records at the same time. This means that the records can be sorted on their score function value and treated in order of priority.

As an alternative to the aggregate method, one could also consider working directly with the sorted record-level score function values, by manually following up records in descending order of their absolute scores and continuing until all aggregates are deemed sufficiently plausible. This was called the *top-down method*¹ by Granquist (1994).

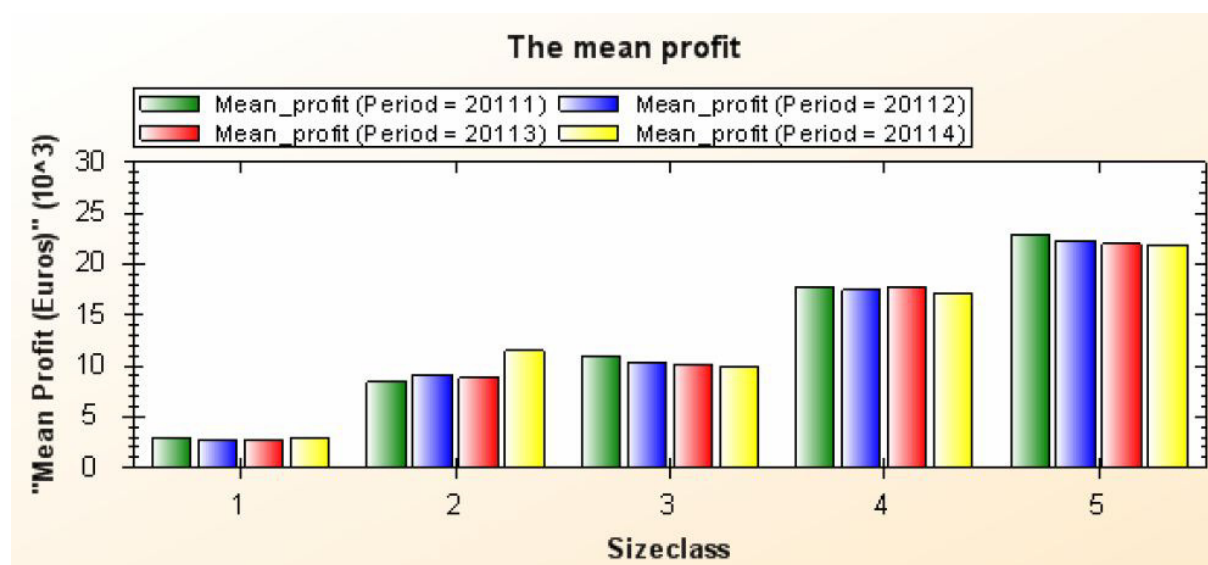


Figure 1. Example of a histogram for macro-editing (taken from Hacking and Ossen, 2012).

In addition to score functions, graphical aids can also be useful for identifying suspicious aggregates. As an example, Figure 1 shows a histogram that compares the mean value of profit across several reference periods and several size classes. It is seen that the mean profit in the last period for size class 2 is unusually high in comparison with previous periods and other size classes. This could be a reason to identify this aggregate as suspicious and drill down to the contributing units.

2.3 The distribution method

Another method for selecting individual units for manual editing, given all or most of the data, is known as the distribution method. This method tries to identify observations that require further treatment by applying techniques for detecting *outliers*, i.e., observations that deviate from the distribution of the bulk of the data. For the purpose of macro-editing, records are then prioritised for manual follow-up by ordering them on some measure of ‘outlyingness’. A discussion of outlier detection techniques in the context of statistical data editing can be found in EDIMBUS (2007).

¹ The name ‘top-down method’ is a potential source of confusion, because it is sometimes used as a synonym for the aggregate method (e.g., De Waal et al., 2011, p. 208). This probably derives from the fact that the aggregate method starts at ‘top level’ aggregates and ‘drills down’ to lower-level aggregates.

Theoretically speaking, there exists some overlap between this approach and the above approach based on score functions, because many common criteria for detecting outliers can be expressed as score functions; see, e.g., De Waal et al. (2011).

Graphical displays can also be useful for detecting observations that deviate from the distribution of the bulk of the data. Common examples include box plots, scatterplots, and other techniques from Exploratory Data Analysis (Tukey, 1977). Figure 2 gives an example of a scatterplot that could be used in this context. A graphical analysis can be particularly effective if the software allows an editor to interact with a display. In the plot of Figure 2, whenever a user moves his mouse to one of the points, information about the relevant unit is automatically displayed. This can be taken one step further by letting a user access a record for further editing by simply clicking on the point that represents the record in the graphical display. See, e.g., Bienias et al. (1997) and Weir et al. (1997) for examples of applications of graphical macro-editing. For some more recent innovations, see Tennekes et al. (2012).

In practice, the distribution method is often applied in conjunction with the aggregate method. Thus, the macro-analysis starts by identifying suspicious aggregates at the highest level and ‘drills down’ to suspicious aggregates at a lower level. Subsequently, the distribution method is applied to identify the records that are likely to contribute most to the total error in the identified low-level aggregates.

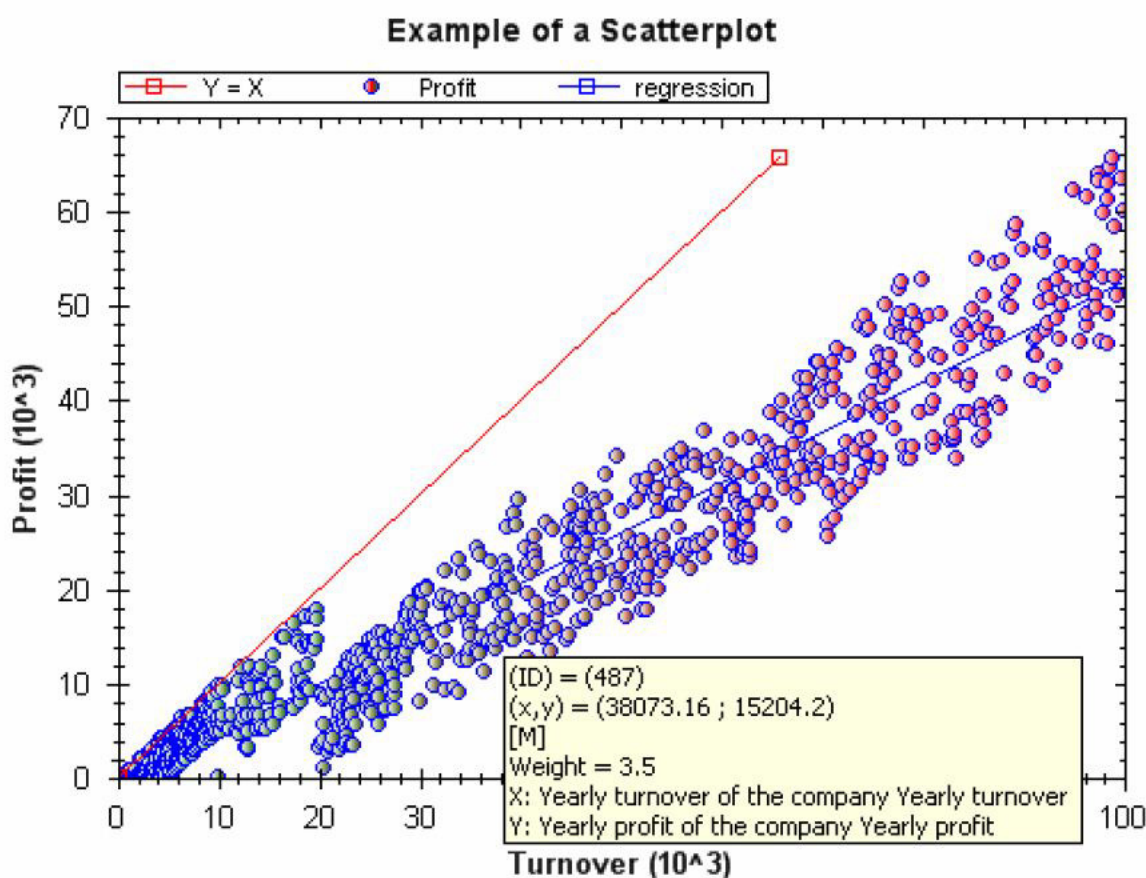


Figure 2. Example of a scatterplot for macro-editing (taken from Hacking and Ossen, 2012).

3. Design issues

4. Available software tools

Many statistical offices have developed macro-editing tools. Quite often, several such tools exist within one office, each one dedicated to a particular survey.

Statistics Netherlands has developed a generic macro-editing tool called *MacroView*; see Ossen et al. (2011) and Hacking and Ossen (2012). It is currently used for macro-editing in the production processes of the Dutch structural business statistics and the Dutch short-term statistics, as well as several smaller statistical processes. It is currently not made available to other statistical offices.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bienias, J. L., Lassman, D. M., Scheleur, S.A., and Hogan, H. (1997), Improving Outlier Detection in Two Establishment Surveys. In: *Statistical Data Editing, Volume 2: Methods and Techniques*, United Nations, Geneva, 76–83.
- De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Farwell, K. and Schubert, P. (2011), A Macro Significance Editing Framework to Detect and Prioritise Anomalous Estimates. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Granquist, L. (1994), Macro-Editing – a Review of Some Methods for Rationalizing the Editing of Survey Data. In: *Statistical Data Editing, Volume 1: Methods and Techniques*, United Nations, Geneva, 111–126.
- Hacking, W. and Ossen, S. (2012), User Manual MacroView. Report PMH-20121125-WHCG, Statistics Netherlands, Heerlen.
- Ossen, S., Hacking, W., Meijers, R., and Kruiskamp, P. (2011), MacroView: a generic software package for developing macro-editing tools. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.
- Tennekes, M., de Jonge, E., and Daas, P. (2012), Innovative Visual Tools for Data Editing. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.

Tukey, J. W. (1977), *Exploratory Data Analysis*. Addison-Wesley, London.

Weir, P., Emery, R., and Walker, J. (1997), The Graphical Editing Analysis Query System. In:
Statistical Data Editing, Volume 2: Methods and Techniques, United Nations, Geneva, 96–104.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Statistical Data Editing – Selective Editing
3. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

1. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

1. n/a

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

12. Tools explicitly referred to in this module

1. MacroView

13. Process steps explicitly referred to in this module

1. Statistical Data Editing

Administrative section

14. Module code

Statistical Data Editing-T-Macro-Editing

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	04-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.2	18-04-2013	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	19-07-2013	minor improvement based on second Swedish review	Sander Scholtus	CBS (Netherlands)
0.3.1	09-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:12



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing Administrative Data

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Statistical data editing of administrative data.....	4
2.2 Types of errors in administrative data	6
2.3 Data editing methods for administrative data.....	8
2.4 Information about data quality	11
3. Design issues	13
4. Available software tools.....	14
5. Decision tree of methods	14
6. Glossary.....	14
7. References	15
Interconnections with other modules.....	16
Administrative section.....	17

General section

1. Summary

The use of administrative data as a source for producing statistical information is becoming more and more important in Official Statistics. Several methodological aspects are still to be investigated. This module focuses on the editing and imputation phase of a statistical production process based on administrative data. The paper analyses how much the differences between survey and administrative data affect concepts and methods of traditional editing and imputation (E&I), a phase of the production of statistics that nowadays has reached a high level of maturity in the context of survey data. This analysis enables the researcher to better understand how and to which extent traditional E&I procedures can be used, and how to design the E&I phase when statistics are mainly based on administrative data.

2. General description

The use of external information in statistical production processes is increasing its importance in the National Statistical Institutes (NSIs).

External information generally refers to *secondary data*, i.e., data not collected directly by the user. An interesting discussion on the use of this kind of data can be found in Nordbotten (2012). In this paper, the focus is on administrative data, which is a subset of secondary data. They have the characteristic of being collected for non-statistical purposes and at the moment they are the mostly used external source of information in NSIs.

Administrative data are collected for administrative purposes, e.g., to administer, regulate or tax activities of businesses or individuals. Although not yet fully explored from a methodological point of view, the field of the statistical use of administrative data can be considered in an advanced state for a number of critical issues like accessibility, confidentiality and risk of misuse.

The usefulness of administrative data depends on their concepts, definitions and coverage (and the extent to which these factors stay constant), the quality with which the data are reported and processed, and the timeliness of their availability. These factors can vary widely depending on the administrative source and the type of information (Statistics Canada, 2010).

It is worthwhile to remark that, although this definition could be applied to survey data, in the context of administrative data it assumes a particular importance since most of the elements considered in the statement are not under the control of the NSIs, while on the contrary for survey data NSIs can, at least in principle, design opportunely all or most of them.

The main advantages deriving from the statistical use of administrative data include: the reduction of costs (in the long term) and of respondent burden, deriving from the reduction of information needs from direct surveys; the improvement of timeliness and accuracy of statistical outputs; the increased potentials for more detailed spatial-demographic and longitudinal analysis.

Main drawbacks are connected to the initial costs due to gain access to the new sources, matching classifications, harmonising concepts and definitions with respect to the target units and the statistics of interests, and assessing quality. Concerning the latter aspect, it is worthwhile noting that the quality of data collection, data capture, coding and data validation are under the control of the administrative

program and may focus on aspects that could be not relevant for the NSI's purposes. In general, these validation activities cannot be considered sufficient to ensure the statistical usability of the data, and extensive additional data editing activities need to be performed before incorporating external data into statistical processes. Methods and tools are to be developed to this aim taking into account the peculiarities of administrative data. In addition, the use of an administrative source generally implies the need of other sources (including surveys) to compensate for non-covered units/variables, thus editing strategies for multi-source data should be developed.

The impact of using administrative data in statistical production processes depends also on their supposed use. Two different scenarios can be distinguished:

- 1) administrative data support surveys: they are used to maintain frames, to improve the efficiency of sample surveys (calibration), to provide information which might be used to assist the E&I process, as an information source that might be used for quality assurance (for instance to compare results);
- 2) administrative data serve as a source for providing the statistical output required, in this case they can be used as a primary source or by integrating them with survey data.

In this paper the focus is on the use of administrative data under scenario 2.

The paper is structured as follows. In Section 2.1 the main objectives of data editing for survey data are discussed in the framework of administrative data. Section 2.2 is dedicated to the illustration of error characteristics in administrative data. The application of traditional methods used E&I is discussed in Section 2.3. How to provide information about data quality is illustrated in Section 2.4. General ideas about the design of E&I of administrative data are proposed in Section 3.

2.1 Statistical data editing of administrative data

The main objectives of statistical data editing are reported in the following list (cf. "Statistical Data Editing – Main Module"):

- OB1 To identify possible sources of errors so that the statistical process can be improved in the future;
- OB2 To provide information about the quality of the data collected and published;
- OB3 To detect and correct influential errors in the collected data;
- OB4 To provide complete and consistent data.

When discussing E&I for administrative data, the main question is how much the concepts developed so far for E&I of a single statistical survey (see EDIMBUS, 2007) can be translated into the administrative data framework. The question is translated in two main questions: 1) whether the above mentioned objectives are still valid, and 2) whether error characteristics and methods usually adopted for detection and treatment are the same. To give an answer to those questions, differences between administrative and survey data should be highlighted.

Two important distinctive characteristics are:

- i. the process of gathering information is not generally under the control of the entity (for instance the NSI) that will provide the final figures,
- ii. information is gathered for other purposes.

Other important differences are that:

- iii. generally the sizes of the data bases concerning administrative data are much larger than those concerning survey data,
- iv. administrative data are frequently used in a statistical production process where data sources are combined and integrated. The integration of data sources becomes a specific trait of the use of administrative data since, as they are gathered for other purposes, they generally do not observe all the variables of interest, and most of the times they refer to a population covering a part of the target population. In those cases, integration between administrative sources and surveys is required to fill the gaps.

Those peculiarities influence the objectives of statistical data editing procedures, a short discussion about interactions between main objectives of E&I and peculiarities of administrative data follows.

Objective OB1

The identification of source of errors becomes in this context particularly important. In fact, one of the main problems is that the definition of collected variables is not designed for the survey purposes, and even after a process of harmonisation, some differences may still remain. The process of editing can help to reveal unexpected differences and to find whether there is a systematic nature of the error suggesting that the definitions are still not completely harmonised. Unfortunately, the improvement of the statistical process is limited by the fact that the process is not completely under the control of the NSI. Most of the times it is not easy or even impossible to return to the administrative entity collecting data and to make the agency change the definition of the variables, the data collection and so on.

Objective OB2

As for E&I of survey data, the data quality assessment in terms of input and output data is a key aspect also for statistics based on administrative data. The fact that two separate entities influence the data and the data production process, i.e., data holder and statistics provider (NSI), implies that two different points of view can be used for quality evaluation: a data perspective and a perspective oriented to the production of statistics. The first one is useful to provide information to the data holder to improve data quality for other data collection occasions, while the second one is important to measure the quality of the statistics provided inside and outside the NSI.

Objective OB3

The generally large dimension of databases has an impact on the detection and correction of influential data (which especially characterise quantitative variables), since for their treatment an expensive data editing procedure based mainly on re-contacting units is generally adopted. On the other hand, the use of multiple data sources may lead to have multiple observed values for a single observation, this information can be used to improve the selective editing procedure in terms of both identification of influential errors and value correction when an influential observation is selected. The same considerations hold when longitudinal information is available on units covered by administrative sources. These aspects will be later discussed in the subsection on editing methods.

Objective OB4

In case of integration of several data sources, the data consistency becomes an essential aspect, because the integration will increase the possible conflicts into the available information. However, as

previously stated, the presence of multiple observations is an important aspect that can improve the E&I procedures, although at this time not many methods are developed to exploit as much as possible this richness of information. This issue will be discussed in the subsection on editing methods.

In the end, we can state that the general setting designed by the objectives of E&I of survey data remains still valid for administrative data. On the other hand, it is important to be aware of the impact of peculiarities of administrative data giving a different perspective to the objectives, those peculiarities will have an impact in the design and use of methods for E&I of administrative data

2.2 *Types of errors in administrative data*

As previously discussed, also in case of administrative data, one of the most important objectives of statistical data editing is to deal with errors, for this reason is important to discuss the characteristics of errors affecting administrative data. Before starting with the description of errors is useful to clarify a question: are administrative data affected by errors? It is difficult to imagine that data relating, for instance, to tax declaration can be affected by errors. It is nowadays accepted the idea that administrative data can be affected by errors (Groen, 2012), in fact also for this type of source errors may arise in many phases of the data production process, e.g., at the data transmission phase between data holder and NSI. Furthermore, there are also less controlled administrative data sources where the information is not so immediately sensible to make the data holder perform a check. A discussion about errors can be found later in this section.

Summarising, as well as survey data, administrative data are normally affected by different types of error: in the most recent literature, it is actually accepted that the non-sampling errors that normally emerge in surveys may also occur in registers (Bakker, 2011; Zhang, 2012). We start from the assumption that all the errors dealt with at the E&I phase in case of a single source survey are potentially present in a single administrative data source, hence the discussion is focused on the new additional aspects characterising errors in administrative data, with special attention to the case of statistics produced by integrating different data sources.

The E&I procedures are mainly designed to deal with measurement errors and missing values, the latter concerning usually item non-response. These sets of errors are analysed in the following.

Measurement errors are defined as differences between the recorded values of variables and the corresponding real values (*intended measure* of the variable). They mainly arise because of the fact that administrative sources are the result of processes which, being designed for purposes other than statistical, may use different concepts and/or definitions than those required for the specific statistical purposes. Important differences between the sources of measurement errors in survey data and in administrative data derive from the fact that the measurement process is very different in the two situations. In surveys using questionnaires, measurement errors derive from a cognitive process (comprehension of the question, retrieval of the information, judgment and estimation, reporting the answer) which also acts in case of administrative data but is not the most important one. A most important role in this case is played by administrative and legislation rules and accounting principles (Wallgren and Wallgren, 2007, p. 180). Typical measurement errors in administrative data are errors in accounting routines, or misunderstanding due to legally complicated questions, or errors deriving from the misspecification of rules used for deriving statistical variables from administrative variables. Furthermore, as some variables recorded for administrative purposes are more important than others,

their accuracy is expected to be superior, as it can be assumed that enterprises answer to less important questions with lower precision. It is worth mentioning that the cognitive process also acts in case of administrative data: measurement errors may derive from the fact that respondents may provide different data to the different government agencies depending on their specific purpose, they may understand administrative concepts and definitions incorrectly (thus introducing errors by deviating from definitions, e.g., including wrong elements in the reported variables), or they can make unintentional errors in providing information.

Among measurement errors, also in case of administrative data variable values may contain **systematic errors** (cf. “Statistical Data Editing – Main Module”), which in this case can be due, for example, to a misinterpretation of record descriptions, originated by changes in the record descriptions and/or variable names in the administrative data bases.

An important source of errors for statistics based on multi source administrative data is the process of data integration itself. When the statistical population is created, objects are adjoined and linked, variables are imported from different sources and derived variables are created. The most relevant types of errors associated to the integration process are *coverage errors*, *identification errors*, *consistency errors*, *aggregation errors*, *missing values* (Zhang, 2012; Wallgren and Wallgren, 2007, p. 177). While coverage errors are not usually treated through E&I, the others are dealt with by or have an impact on the E&I process, for this reason they are described in the following.

Identification errors. They may be originated by errors in identifying variables used to match the different sources. As a consequence, identification errors may give rise to doublets, mismatches (e.g., false hits), item and total non-response, data inconsistencies (as variables may be referred to not properly matched objects). Identification errors may also generate outliers, and influential errors.

Consistency errors. They may also originate from the integration of variables from many sources. This type of error is especially increased when using multi source data, on the contrary with a single statistical survey, the use of a unique questionnaire ensures a better consistency in the data. Consistency errors can be caused by errors in units and errors in variables. They may also have a longitudinal origin, e.g., due to identifying variables either in error or changing over time for a same unit, splits/fusions of a unit over time.

Incoherent variable values giving rise to consistency errors in microdata may occur in the situation where the integrated administrative sources are overlapping regarding (a subset of) variables.

Inconsistencies with information from other sources and outliers can be originated from modifications of the variables’ definitions adopted in a source (e.g., resulting from legislative changes), and from the fact that units may change their structural characteristics (e.g., fusions or splits). Outliers can also be determined by taxation measures that produce anomalous changes in variables values over time, and by integration errors (e.g., different units are linked in administrative sources). Outliers can either correspond or not to influential errors, depending on their impact on the target estimates.

Aggregation errors. They may occur when data from different administrative sources with different types of units are integrated in order to derive statistical variables (Wallgren and Wallgren, 2007), e.g., enterprise labour cost deriving from fiscal archives on enterprise employees. Aggregation errors may originate internal inconsistencies among variables referring to the same unit, outliers and longitudinal inconsistencies.

Missing values. As for statistical surveys, also in case of administrative data, missing values may correspond to two types of non-response : *unit non-response* (all the information for a statistical unit is unavailable) and *item non-response* (incompleteness of information, for some units, on topics which are of interest for statistical purposes). In case of administrative data, unit non-response corresponds to under-coverage, for example, when the integrated administrative sources relate to sub-populations which do not cover the overall target population. Item non-responses typically derive from the fact that the content of administrative sources is defined on the basis of administrative requirements, thus not all topics of interest may be covered by the administrative data. Possible sources of item non-response can arise for other different reasons: variable values can be missing for certain objects due to flaws of a source; mismatches at the integration phase due to missing objects in a source, giving rise to missing values for all the variables which are imported from that source; reported values which are “cancelled” as recognised invalid at the editing stage; values which fail to be reported, or are reported with a delay. Item non-response can also be associated to the fact that the content of a source is subject to modifications, resulting from legislative changes, like the drop-out of some information from the administrative forms; in a longitudinal perspective, non-responses can also appear as missing information on target variables for units considered over time: this can be due again to modifications of the units (fusions/splits, other structural changes) or to changes in legislation. Finally, as administrative sources may refer to either a point in time (i.e., they describe the units set at that point in time), or to a calendar year (in this case they contain all units that have existed at any point during the year), item non-responses may rise when sources with different time characteristics are integrated.

2.3 *Data editing methods for administrative data*

In this section we focus the attention on methods which can be used to detect and treat measurement errors and item non-response, that are in fact the errors dealt with by an E&I procedure..

Several classifications for the data editing techniques are available; we follow the one proposed in “Statistical Data Editing – Main Module”. The techniques can be classified as:

1. Deductive editing.
2. Selective editing.
3. Automatic editing.
4. Interactive editing.
5. Macro-editing.

The order follows the strategy that is generally adopted in an E&I process for a statistical survey (cf. “Statistical Data Editing – Main Module”).

In this section we discuss the impact of the peculiarities of administrative data on the features of each data editing technique.

Deductive editing is the phase where methods for detecting and treating errors with a structural cause that occurs frequently in responding units (systematic errors) are used (see “Statistical Data Editing – Deductive Editing”). In administrative data, especially when more sources are used, deductive editing has an important role in the production process. Variables collected in the administrative sources may have similar definitions but they may have structural gaps given to the convenience of declaring some

information in an item rather than in another one, for instance, declaring something either in a cost or in an investment item. The first step in an E&I process should be to look for systematic errors in the observed values, also in the case the definition of variables is almost the same with respect to the corresponding statistical target variable. Hence, deductive editing is substantially the same as the one carried out in a classical data editing process, in fact the detection of systematic errors implies the involvement of subject matter experts, and the error treatment, that is usually completely automated, is not affected by the large dimension of administrative databases.

The aim of *selective editing* is indeed the optimisation of the process of selection of units to be deeply revised (in most cases, re-contacted) by restricting the editing only to those affected by an important error, and this naturally stresses the importance of selective editing in this context where data sets have usually a large dimension. On the other hand the use of selective editing is actually limited by resources' constraints because even a small percentage of units to be analysed may be too large in a large data set. A further constraint for selective editing on administrative data derives from the difficulty of re-contacting units for this kind of data. This limitation is alleviated when multi-source data are used, in this case the availability of different values for the same observation is an important aspect that can help the statistician in understanding where the error is located and to recover a likely value. The previous considerations mainly illustrate the problems in applying selective editing to administrative data. However, some further remarks concerning positive aspects of selective editing with administrative data are worthwhile to be mentioned. In selective editing, observations are prioritised according to a score function measuring the impact on the target estimates of the expected error in the unit. The error is frequently measured by comparing the observed value with a suitable prediction. In the context of administrative data, there is frequently the possibility of using longitudinal data, and this can improve the efficiency of selective editing as better predictions can be obtained. Finally, it is worthwhile to note another specific difference characterising the application of selective editing in administrative data with respect to the survey data. In a survey, the error is generally weighted with sampling weights. Since the prioritisation of an observation should be based on the impact of the error on the estimates, the final sampling weights should be taken into account in this process. In practice, this can be rarely performed, as final weights are generally computed once the editing step is completed, so an approximation is generally used by considering initial sampling weights. In the case of administrative data this problem is naturally overcome because sampling weights are not an issue for these kinds of data and a more precise estimation of the impact of errors on estimates can be obtained.

Automatic editing refers to all E&I procedures that detect and treat errors and missing values in a data file in a fully automated manner, i.e., without human intervention (see "Statistical Data Editing – Automatic Editing"). In the last years, most of the methods for automatic editing are based on the Fellegi-Holt paradigm, which means that the smallest number of fields should be changed to a unit to be imputed consistently. The algorithms are based on edits that represent rules/constraints characterising the relationships among variables.

In principle, if the focus is just on one data source, we are in the same situation as the one we would have in an E&I process of statistical survey data. However, as already remarked, most of the times different data sources are integrated, and in this case some additional problems may arise. A first issue to take into account is whether the data sources should be treated simultaneously as a unique data set after the integration process. This could be an interesting option, because the amount of information

would increase, and an improvement in the E&I procedure is expected. In this case, edits simultaneously involving variables of the different data sources should be considered. A special but not infrequent case is when the same (at least in principle) variable is observed in the different data sources. For the sake of simplicity, let us suppose that there are only two data sets with the same variable. According to the Fellegi-Holt approach, we are assuming that with a high probability at least one of the two variables in turn is not affected by error. In the case that this assumption is not reliable, a different approach should be followed, for instance, a prediction conditionally on the observed values of the two variables can be obtained. Techniques developed to this aim are described in the module “Micro-Fusion – Reconciling Conflicting Microdata”.

Concerning *interactive editing* for administrative data, the most relevant aspect is that, as already remarked, it is frequently not possible to re-contact the observed units, so one of the main advantages motivating interactive editing declines. However, interactive editing can be considered effective in order to understand error sources and possibly resolve errors in the short term, while in the long term it can contribute to the increase of the subject-matter expertise for the staff working on administrative data, increasing their knowledge of the characteristics and the contents of administrative data and gaining understanding of how the data can be used in a more suitable way (Wallgren and Wallgren, 2007).

Macro-editing aims at looking for anomalous aggregates. The anomalies are identified based on the comparison of aggregates with some reference values that, for instance, may be obtained by previous published figures. Once anomalous aggregates are selected, a drill-down procedure is applied in order to find the units that mostly contribute to this behaviour (see “Statistical Data Editing – Macro-Editing”). This editing approach requires the computation of the final aggregates (e.g., domain estimates), and for this reason, in the usual E&I procedure it is generally performed at the end of the E&I process. In this context, one generally works on complete data sets, in fact administrative data are gathered for other purposes and they are usually provided to the NSIs at the end of their collection. This implies that in this context macro-editing methods can be used at the beginning of an E&I procedure in order to look for important errors.

Macro-editing can be a useful tool to reveal whether some important errors due to an incomparability of the sources in some estimation domain are still present in data. For instance, it can happen that the definition of a variable is the same in two data sources. Nevertheless, for a specific economic sector some particular businesses could not provide the complete amount of the value in one source because of fiscal benefits typically allowed only for that segment of units. Macro-editing can be useful to isolate those critical situations that the subject matter expert may study and interpret in order to fix the problem wherever it is possible. Macro-editing can also reveal errors due to data linking or to the incomplete delivery of some sources, as anomalous aggregates may result from not enough covered domains from one time period to the subsequent one.

As already mentioned, administrative data are subject to partial non-response as well. **Imputation** (see the topic “Imputation”) can be used to manage missing values in order to obtain a completed data set on which the usual statistical analysis can be applied. The methods usually adopted are based on the missing at random (MAR) assumption that is, roughly speaking, the probability of non-response on a given variable depends on the observed values and not on the unobserved ones of the variable itself. For instance, missing values in administrative data can be due to lack of timeliness, and it is generally

supposed that businesses answering in due time have the same behaviour as the not observed ones. Actually this situation could hide the presence of a problem in the business, and in this case the estimates could be biased because the observed and non-observed populations are actually different. A similar concept applies in the case of an integrated use of administrative data. It can happen that each administrative source covers only some specific part of the target population. Imputation can be used to complete the missing values, again under the assumption that the population not covered has the same behaviour of the observed one.

Finally, since the production process of administrative data is generally beyond the control of NSIs, a continuous assessment of the data quality should be planned. Edit rules and macro-editing based approaches could be used to this aim. An anomalous rate of edit failure and/or anomalous variation of statistical aggregates in two consecutive times could alert data producer that some important changes could have been introduced in the administrative data production process, which could be related to a change in the data collection, to a change in the legislation that impacts on the definition of measured variables, consequences of a different fiscal policy, and so on.

2.4 Information about data quality

One of the main goal of E&I is to provide information about the quality of the data collected and published.

Quality of statistical output has several dimensions, they are thoroughly discussed in Eurostat (2011) for the European Statistics Code of practice, Eurostat (2009) for a handbook (soon to be revised) on reporting quality of statistical data according to the European output quality components, and the handbook module “Quality Aspects – Quality of Statistics”.

In this section it is important to refer to the quality dimensions in the context of administrative data in order to describe on which of them the E&I is a useful tool for providing information. In the BLUE-ETS (2011) document, the quality dimensions of administrative sources and the related indicators are discussed. In that document the focus is on the quality dimensions of the administrative data sources in the input phase of a statistical production process, this point of view is adopted in this paper as well. As far as the quality dimension of the statistical output based on administrative data is concerned, we assume that at the end of the E&I process data are statistically transformed, and hence the general considerations made for statistical output based on survey data are still valid. This is a simplistic position, that is also motivated by the fact that at this time this issue is still under discussion, and further studies are needed in this context. For the use of E&I procedures as a useful tool for providing information on quality of statistical data, the reader may refer to EDIMBUS (2007).

A first interesting remark relates to the point of view chosen to look at the quality aspects. It reflects the peculiarity of statistics based on administrative data where generally two different main actors are involved: the data holder and the statistics provider (NSI). Two main points of view are introduced: a data archive perspective and a perspective oriented to the production of statistics. In the first one, the quality is independent of the specific statistical use of the administrative data that is supposed to be done, while in the second one the quality is related to the statistical use of the data planned at the NSI. Both these aspects are important for E&I, in fact the first one has to be assessed in order to foster data holder to improve the quality of the data, while the second one is related to the quality of published data.

In the BLUE-ETS document, the following quality dimensions are defined:

1. *Technical checks*, that is the technical usability of the file and data in the file.
2. *Accuracy*, that is the extent to which data are correct, reliable, and certified.
3. *Completeness*, that is the degree to which a data source includes data describing the corresponding set of real-world objects and variables.
4. *Time-related dimension*, in which timeliness, punctuality, and overall time lag applied to the delivery of the input data are taken into account.
5. *Integrability*, that is the extent to which the data source is capable of undergoing integration or of being integrated.

The *technical check* dimension is mainly related to IT aspects, e.g., data accessibility, correct conversion of the data, data complies with the metadata-definition. These aspects are not related to an E&I procedure as it is defined in “Statistical Data Editing – Main Module”.

E&I has certainly impact on *accuracy*, and it naturally provides information about some dimension indicators described in BLUE-ETS (2011) related to this aspect. Some of the dimension indicators for accuracy proposed in BLUE-ETS are supposed to measure:

- *Measurement error*: deviation of actual data value from ideal error-free measurement;
- *Inconsistent values*: extent of inconsistent combinations of variable values;
- *Dubious values*: presence of (or combinations of) implausible values for variables.

Those elements are treated and analysed during an E&I procedure, and indicators measuring them are developed and generally automatically provided by the usual procedures (see EDIMBUS, 2007).

E&I may be useful to gather information also for other quality dimensions, that apparently are less naturally related.

Completeness is a concept referred to units and variables, and for the latter the quality dimension indicators proposed in BLUE-ETS (2010) are: the amount of missing values and the amount of imputed values. As previously stated, the treatment of missing data (imputation) is one of the main activities carried out in an E&I process; hence, indicators on those aspects are easily obtained in this context.

As far as the *time related dimension* is concerned, a proposed indicator focuses on the stability of variables. To this aim, the comparison in different times of indicators generally provided by E&I may be useful: for instance, an anomalous variation of the failure rates of some edits may hide some changes in the administrative data production process or in the source contents, or in the use of a different definition for a variable, or in a different data collection mode. Also the comparison of the amount of imputed values and missing data can reveal some changes in the data source which have to be taken into account in order to avoid biasing effects on statistical results.

A summary of the editing undertaken and the results of the checks should be sent to the database owner to make him aware of the problems possibly existing in the data set, in order to reduce them as much as possible in the future and improve the overall quality of the data. As a consequence, managing and improving co-operation with administrative bodies plays a central role in this context:

NSIs need to increase co-operation and to determine appropriate incentives in order to improve the overall communication and interaction with data owners, to get them to set up better editing practices and conform to statistical classifications and definitions, and to provide feedback to the NSI in the data verification process (Shlomo and Luzi, 2004).

3. Design issues

In the design of an E&I process for administrative data the first important issue to take into account is whether the target statistics are based only on a single administrative source or on the use of multiple integrated administrative sources. Moreover, editing strategies must take into account the trade-off between the potential gain in accuracy deriving from the availability of detailed and extensive information, and the additional costs needed for validating it.

When only one source is used, as discussed in the previous sections, we are in a similar situation to that of E&I of a single survey, even if we remind that peculiarities of administrative data should be taken into account because of their impact in the E&I methods. The reference flow-chart introduced in the module “Statistical Data Editing – Main Module” can be applied to this case.

When more sources are integrated, different scenarios can be depicted.

A first scenario may consist of the following macro-phases:

1. check separately each single administrative source;
2. integrate the edited data sources;
3. edit the integrated sources in order to assess the consistency among variables’ values obtained from the different sources.

This is actually the flow-chart reported in Wallgren and Wallgren (2007, p. 101).

The drawback of this way of proceeding is that it is resource demanding since many different E&I procedures must be set and applied, and it is well known that the E&I is one of the most expensive parts of the statistical production process. Moreover, not all the amount of information is used at the same time, for instance, for the imputation of a variable in a data source it could be useful to exploit variables observed in the other data sources. Let us imagine the case when two data sources are integrated and in one source the income is observed, while in the other one information on consumption is gathered. The imputation of the two variables separately would disregard the strong relationship existing between them. An advantage of this way of proceeding is that certain typologies of errors (e.g., systematic errors like unity measure errors, balance errors, errors due to incomplete delivery of data for some administrative objects) can be removed from each single source before the integration phase, thus reducing the amount of consistency errors on the linked data deriving from these situations; longitudinal information could be used at this stage.

An alternative scenario corresponding to the opposite solution is:

1. integrate the sources;
2. apply an E&I procedure to the integrated data set.

In this case, less resources would be demanded since only one data verification process is required, but the complexity of such a process would increase. Furthermore, as the integrated data set is not

generally composed of all the variables observed in the different administrative sources, in this case some relations linking variables in each data source could be disregarded.

A third scenario is a compromise of the previous ones:

1. apply a 'light' E&I procedure to each single administrative source;
2. integrate the edited data sources;
3. edit the integrated data sources.

The question is when an E&I procedure can be defined as light. The idea is that the time and effort spent in editing sources should be minimised while maintaining an acceptable level of quality of the data sources. This general idea resembles what is done in selective editing, where the effort is focused on the most important errors having a high impact on the target aggregates. This situation is slightly different because there is no requirement on a sufficient level of quality of aggregates for each single data source, but the level of quality is required at micro level: in effect, the use of each single source will be in a micro perspective given that the integration process is generally performed at this level. A proposal could be that of applying only corrections of systematic errors in the first editing step.

It is clear that a general flow-chart is not available; however, at least three scenarios have been designed. The main point is to see the E&I process as a unique process possibly composed of two steps. The choice of the most appropriate strategy should be based on the trade-off between the expected quality of the final aggregates and the resources which are actually available to obtain the required level of quality. Concerning the latter, an element which can be considered as relevant to increase the effectiveness of editing and correction activities is the availability of subject-matter experts, who are familiar with the administrative systems that have generated the data and their specific contents, and who are in good relations with the data providers.

Finally, independently on the chosen scenario, indicators providing information about input and output data quality should be part of the E&I process. Moreover, since the process of gathering information is out of control of NSIs, it is important to establish a system of indicators alerting about some possible changes in the data production process of the data holder, in order to avoid important and non-measurable errors in the published statistics.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

- Bakker, B. F. M. (2011), Micro-Integration: State of the art. In: *Report WP1: State-of-the-art on Statistical Methodologies for Data Integration*, ESSNET on Data Integration, available at <http://www.cros-portal.eu/content/wp1-state-art>.
- BLUE-ETS Project (2011), *Deliverable 4.2: Report on methods preferred for the quality indicators of administrative data sources*. Available at <http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf>.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.
- Eurostat (2009), *ESS Handbook for Quality Reports*. Eurostat Methodologies and Working papers.
- Eurostat (2011), *European Statistics Code of Practice*. For the national and community statistical authorities. Adopted by the European Statistical System Committee 28th September 2011 (revised version).
- Groen, J. A. (2012), Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics* **28**, 173–198.
- Nordbotten, S. (2010), The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries. In: Carlson, Nyquist, and Villani (eds.), *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, 205–225. Available at officialstatistics.wordpress.com.
- Shlomo, N. and Luzzi, O. (2004), Editing by Respondents and Data Suppliers. In: *Federal Committee on Statistical Methodology, Statistical Policy Working Paper 38: Summary Report on the FCSM-GSS Workshop on Web-based Data Collection, April 2004*, 75–90.
- Statistics Canada (2010), *Survey Methods and Practices*. Catalogue no. 12-587-X. <http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf>.
- Wallgren, A. and Wallgren, B. (2007), *Register-based statistics – Administrative data for statistical purposes*. John Wiley and Sons, Chichester.
- Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* **66**, 41–63.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Main Module
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Imputation – Main Module
6. Weighting and Estimation – Estimation with Administrative Data
7. Quality Aspects – Quality of Statistics

9. Methods explicitly referred to in this module

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Statistical Data Editing – Deductive Editing
3. Statistical Data Editing – Automatic Editing
4. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

14. Module code

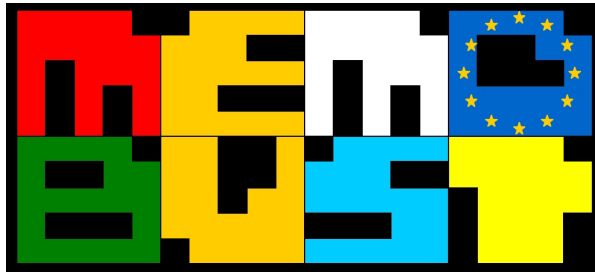
Statistical Data Editing-T-Administrative Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	13-03-2013	first version	M. Di Zio, O. Luzi	Istat
0.2	17-06-2013	introduction of a new section concerning quality indicators	M. Di Zio, O. Luzi	Istat
0.3	07-08-2013	minor revisions	M. Di Zio, O. Luzi	Istat
0.3.1	04-10-2013	preliminary release		
0.4	20-12-2013	revision based on EB comments	M. Di Zio, O. Luzi	Istat
0.4.1	09-01-2014	revision based on EB comments	M. Di Zio, O. Luzi	Istat
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:13



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing for Longitudinal Data

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Longitudinal data.....	3
2.2 Introduction to editing for longitudinal data.....	4
2.3 Editing scheme in a longitudinal context	4
2.4 Type of edits	5
2.5 Methods for longitudinal data	6
2.6 The case of categorical data	8
3. Design issues	9
4. Available software tools.....	9
5. Decision tree of methods	10
6. Glossary.....	10
7. References	10
Interconnections with other modules.....	11
Administrative section.....	12

General section

1. Summary

We refer to longitudinal data as repeated observations of the same variables on the same units over multiple time periods. They can be collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each unit from historical records. The process of Editing and Imputation can exploit the longitudinal characteristic of the data as auxiliary information, useful at both the editing and the imputation stages. This theme describes the editing process applied to longitudinal data, that could be performed for all aforementioned types of data, with special focus on Short Term Statistics context.

2. General description

2.1 Longitudinal data

Another term for longitudinal data is panel data. This definition focuses on the particular sample, which units are selected to be observed several times with some degree of regularity. The occurrence of those observations can be once along several years (every four years or biannual) or once a year (annually) or several times during the same year (quarterly or even monthly). Panel data are mostly used to describe patterns of change within and between the statistical units under observation, in other cases to highlight and to identify differences and changes over time of a specific parameter of the population under study. In general, for each unit $i = 1, \dots, n$ there are $t = 1, \dots, T$ different measurements, one for each wave of interview. The period t can be a month, a quarter or a year; the first two cases drive to infra-annual longitudinal data. As a consequence, given the period t , a vector of cross-sectional observations is available, while as regards the i -th observation a vector of longitudinal data is available and a strong correlation is expected among its values. According to the type of required estimates, different types of panel are considered, so it can always follow the same units or rotate some of them after a period (rotating panel). The different design will create different type of longitudinal data set.

In the context of business statistics, longitudinal data can be used both in structural and in short-term analysis. The difference between Structural Business Statistics (SBS) and Short Term Statistics (STS) actually depends on the combination of the survey occurrence and the type of final target parameter; see also the modules “General Observations – Different Types of Surveys” and “Repeated Surveys – Repeated Surveys”. In the SBS context, totals, means, levels are usually the object of the estimates; in the STS the main objective is usually to publish regular series of statistics on changes of totals for specific domains. These are frequently published in the form of index numbers, whose main purpose is to measure net changes between two periods. In these cases the rationale for a panel design is to improve the precision of estimates, because the minor variance of estimates is assured by the presence of historical correlation between data referred to the same units over the period in which the observations take place; see also the topics “Sample Selection” and “Weighting and Estimation”. On the other hand, also from an operational point of view, the use of a panel for an infra-annual survey can yield important cost savings. Indeed, to interview the same units is often less expensive than starting afresh, at each wave, the contacts on new units.

2.2 *Introduction to editing for longitudinal data*

In general, two main aspects are crucial in an editing process framework:

- 1) the rule to identify an acceptance region for a test variable;
- 2) the technique used to change a value detected as wrong during the process.

In a longitudinal context, these aspects have to be fitted to the specific target parameter, which is often given by the estimation of the change of a population parameter (mostly the mean) concerning a quantitative not-negative variable y . It is strongly recommended to use the available historical information of the observation units for two main reasons:

- 1) a strong correlation is expected among different measurements of the same variable on the same units, thus any detecting rule can rely on relevant information about the unit profile and can result in being more efficient;
- 2) since most of the time the target parameter is the change of a main parameter along time, any observed change between sequential periods on the observations can be used as a precious source of information with regards the final estimation.

In general, the editing process in a longitudinal context must take into account the characteristics of the change under investigation and the timeliness constraints. The control rules can be defined taking into account comparisons between values of the same variable on the same unit at different times, i.e., the two values y_t and y_{t-k} , where t is a month or a quarter, $t-k$ is a previous period and k varies according to the variable features and/or to the type of change under observation. Additional specifications are generally required, they are briefly described in the following.

2.3 *Editing scheme in a longitudinal context*

When the editing process is set on longitudinal data, there are some issues which assume a strategic meaning:

- 1) Longitudinal and cross-sectional checks can be carried out at the same time; this is because longitudinal surveys keep a statistical relevance for cross-sectional analyses as well. For instance, a certain variable x may have a direct connection with the target variable y and, as a consequence, a specific cross-sectional check is needed. In this case, a troublesome decision concerns the priority level among the cross-sectional and the longitudinal checks, even though the last ones should come first. Thus, it is important to coordinate them in order to avoid the risk to oversize the overall number of checks as well as the amount of changes carried out on the original micro-database (Granquist and Kovar, 1997). On the other hand only cross-sectional checks may be applicable in case of “new” units, for which no past data are available.
- 2) Given the target parameter and the characteristics of the variable under investigation, at each reference time t there is the need to specify which are the previous periods to be considered in the editing process. For example, for monthly data the periods $t-1$ and $t+1$ or $t-12$ and $t+12$, most of the times because of the presence of significant seasonal components.
- 3) Economic units may change their demographic features over time (such as change of their ownership, location, economic activities carried out, number of local units, employment and so on) as a result of events of different nature (i.e., mergers or splits). Statistical units interested by

these changes could lose their “longitudinal” identity and their data cannot be compared in a longitudinal data analysis process. As a consequence suspected changes may come up, which are not the results of real mistakes, but they are due to structural changes of the unit economic profile along time. In a longitudinal survey context – in particular, in a short-term survey framework – it is often difficult: a) to identify cases when there are anomalous increases or decreases due to demographic changes and not to real measurement errors (lack of updated information even from the business register); b) to apply a proper amendment to microdata able to overcome the non-comparability of data over time.

- 4) In a short-term survey framework, the required timeliness for the elaboration of the indicators becomes a hard constraint for the editing strategy, as it strongly reduces the available time to check all the microdata. It is a good solution to identify a sub-set of “critical” units, for which a deeper analysis can guarantee the required quality. This approach is generally defined as *selective editing*, which presumes the definition of a *score function* to rank the observations according to their impact on the target estimates; see the module “Statistical Data Editing – Selective Editing”. Several score functions are proposed in literature, the difference among them is mainly given by the way to measure the impact on the final estimates, that anyway usually depends on: i) the given sampling weights; ii) the size of the possible error; iii) the longitudinal behaviour of each respondent.

2.4 Type of edits

The error detection process usually consists of a set of integrated error detection methods dealing each with a specific type of error (EDIMBUS, 2007), which results are flags pointing to missing, erroneous or suspicious values. Error detection is often based on the use of edit rules, that are restrictions to the values of one or more data items that correspond to missing, invalid or inconsistent values potentially in error (cf. “Statistical Data Editing – Main Module”). In a longitudinal context, the coherence of individual historical data is the basic rationale to analyse the data, because the units are believed to be strongly characterised by their own longitudinal profile. According to this point of view, the data of each unit at the occasion t can be checked by comparison with other values observed on the same unit at other times, i.e., belonging to its profile, with regards to an expected value or range.

In the following, the typology of edits is described according the needs and the features of a longitudinal context:

- Consistency checks: their purpose is to detect whether the value of two or more variables on the same unit are in contradiction, hence, whether the values of two or more data items do not satisfy some predefined expected relationship. In this regard, comparisons with other sources which produce comparable microdata are included. Data items can refer also to measurement on the same unit in different periods, it is important that this reference data has been previously checked for errors¹. The reference data used and the way in which the comparison takes place depend on the target parameter.

¹ If the past value y_{t-k} refers to the previous year, past data can be supposed to have been fully checked on the basis of information available from sources external to the survey, so that normally suspect ratios y_t/y_{t-k} lead to change the actual value y_t (but not the past value). However, this rule is not rigid and past data may be changed as well (that is the case of wrong reporting by some units which can review past values even one year later).

- Balance edits: often the value of a variable at time t can be obtained by the sum of the values in the previous period and the registered flow in the reference period for that variable; e.g., the number of persons employed at the end of month $t-1$, plus the number of persons who started working between months $t-1$ and t , minus the number of persons who stopped working between months $t-1$ and t , must be equal to the number of persons employed at the beginning of month t .
- Check for unity measure errors: some errors are due to misunderstandings about the measure according to which a variable x is collected, e.g., thousand instead of billion and so on. In these cases, there is a thousand-error if one of the following relations is verified:

$$\text{abs}(x_t) > h \cdot [\text{abs}(x_{t-k})] \quad \text{for some } k \in \{1, \dots, P\} \quad (1a)$$

$$h \cdot [\text{abs}(x_t)] < \text{abs}(x_{t-k}) \quad \text{for some } k \in \{1, \dots, P\} \quad (1b)$$

where $x_{t-k} > 0$, $\text{abs}(x)$ is the absolute value of the variable x and h is a constant to be chosen properly by the expert.

- Ratio edits. These edit rules are bivariate restrictions taking the general form $a \leq x / y \leq b$, where x and y are numerical variables and a and b are constants. In a longitudinal context, the comparison is based on the two measurements y_t and y_{t-k} , k will vary according to case under study (type of data, characteristics of the variable, etc.).
- A further type of edit is related to a specific feature of longitudinal surveys, because it is possible to ask twice for the same data, with reference to the same variable for the same period. Normally, it happens when a certain value is asked in two consecutive waves at times $t-1$ and t . Let $y_{it(t-1)}$ be the value of the variable y on the unit i asked in the wave t even though referred to the $t-1$ period, then a frequent longitudinal check is given by:

$$y_{it(t)} = y_{it(t-1)} \quad (2)$$

This option may help both to check for the quality of supplied longitudinal information and to take under control changes of some accounting figures inside the unit; it is also very useful to achieve longitudinal data from units characterised by wave non response, e.g., those units which may be non-respondent in $t-1$ and respondent in t , or vice-versa. This solution has to be defined accurately, in order to be worth without increasing the statistical burden on the respondent units.

2.5 *Methods for longitudinal data*

In a longitudinal context, one of the most relevant test variables is the “individual trend” or “individual change”, defined as:

$$c_{it} = y_{it} / y_{it-k} \quad (3)$$

As a consequence most data controls are based on the study of (3) and on rules to check whether the individual trend is too large or too low. The main issue is to define a criterion to decide whether a given level satisfies or not the acceptance rules. The unit trend information can be used in different ways, a couple of them is shortly resumed as follows.

2.5.1 *The Hidioglou-Berthelot method for detecting outliers*

The empirical distribution of all the individual trends can supply useful information for the editing process, by comparing each c_{it} with some main indicators of such distribution. In this regards, the

Hidiroglou-Berthelot method (Hidiroglou and Berthelot, 1986) proposes a way to establish an acceptance interval for c_{it} , based on a function of its interquartile, in order to detect outliers.

Firstly, for each occasion t the median of all the c_{it} is elaborated, defined as $q_{0.5}(c_t)$. Afterwards, a transformation is applied to every c_{it} , to ensure more symmetry of the distribution tails:

$$s_{it} = \begin{cases} 1 - q_{0.5}(c_t)/c_{it}, & \text{if } 0 < c_{it} < q_{0.5}(c_t) \\ q_{0.5}(c_t)/c_{it} - 1, & \text{if } c_{it} \geq q_{0.5}(c_t) \end{cases} \quad (4)$$

Let also define:

$$E_{it} = s_{it} \cdot \{\max(y_{it}, y_{it-1})\}^U \quad (5)$$

which is the “effect” concerning unit i at time t ; it is based on the “individual trend” component s_{it} defined by (4) and the “size” component due to the y -levels of the same unit. The parameter $U \in [0,1]$ is a tuning parameter which should balance the magnitude of the size component with respect to the individual trend. Then, given the first and the third quartile, $q_{0.25}(E_t)$ and $q_{0.75}(E_t)$, the following values are defined:

$$D_1 = \max \{q_{0.5}(E_t) - q_{0.25}(E_t), A \cdot q_{0.5}(E_t)\} \quad (6)$$

$$D_3 = \max \{q_{0.75}(E_t) - q_{0.5}(E_t), A \cdot q_{0.5}(E_t)\} \quad (7)$$

where the constant A is chosen to avoid difficulties which can arise when the differences $q_{0.5}(E_t) - q_{0.25}(E_t)$, and $q_{0.75}(E_t) - q_{0.5}(E_t)$ are small (generally it is set to 0.05).

Hence, the acceptance region is defined as follows:

$$(q_{0.5}(E_t) - A \cdot D_1, q_{0.5}(E_t) + A \cdot D_3) \quad (8)$$

and each observation y_{it} which falls out of such interval is considered to be an outlier.

It is worthwhile to underline how the identification of anomalous ratios c_{it} due to errors (not necessarily outlier observations) may be carried out according to an analogous methodological scheme.

2.5.2 Score functions ranking

In case a selective editing scheme has to be defined, the basic rationale is the evaluation of the impact of the change of each unit on the overall trend, considering its size and its sampling weight. This kind of analysis can be carried out ranking the units on the basis of a score function, which takes into account the above mentioned dimensions. Thus, a simple score function to be applied to each unit depends on the three dimensions:

$$\text{Score} = (\text{longitudinal trend}) \times (\text{sampling weight}) \times (\text{size}).$$

In the following, a score function is described that takes these elements into account, for which a transformation of the individual trend c_{it} is defined in order to take into account different options of needs. A preliminary transformation is made to assign high priority to units characterised by either a very high or a very low change:

$$d_{ij} = \max(c_{it}, 1/c_{it}) = \max(y_{it}/y_{it-k}, y_{it-k}/y_{it}) \quad (9)$$

New units, for which no historical data are available, will be assigned $c_{it}=1$.

Then, the following conversion will be used to define the final score function:

$$r_{it} = |k_{1i}d_{it} - k_{2i}|$$

where k_{1i} and k_{2i} can be chosen according to any needs expressed by the given survey, a typical choice is to put both k_{1i} and k_{2i} equal to 1.

Thus, the score function for a generic unit i and a given time t can be built up as follows:

$$\Phi_{it} = r_{it}^{\alpha} w_{it}^{\beta} z_{it}^{\gamma} \quad (10)$$

where w is the sampling weight and z is a “size” variable (for instance, turnover, production, number of persons employed). Parameters α , β and γ should be used in order to balance the relative importance of each score component on the final score Φ . Normally it is recommended to use parameter values chosen from the interval [0,1] (Gismondi and Carone, 2008). After the calculation of the score (10) for each unit, scores can be ordered in a non-decreasing ranking: the units occupying the “first positions” in the ranking will be detected as influent suspicious units, to be checked with priority or even re-contacted. Some techniques for assessing the number of influent units have been proposed by McKenzie (2003), Philips (2003), Chen and Xie (2004).

2.6 The case of categorical data

There are particular kinds of business longitudinal surveys for which categorical variables play a fundamental role. That may happen when the main goal:

- a) is still the evaluation of the change of a quantitative variable, but a preliminary step consists in the assessment of the presence (or absence) of a certain phenomenon (binary variable: 1=present, 0=absent);
- b) consists in the evaluation of a set of opinions and their developments over time (qualitative variables).

An example of the kind a) is the survey on job vacancies. The main goal is the estimation of the number of job vacancies at the end of each quarter, but a preliminary step consists in assessing if an enterprise is searching for new personnel or not. There are the following possibilities:

- The firm declares an amount of job vacancies higher than zero, that implies the firm is searching for new staff. In this case no problem is encountered.
- The firm declares zero job vacancies. This value may be right, but it may be wrong as well, for instance, because the firm is not able to correctly count the number of job vacancies (and prefers to declare zero in order to tackle the question quickly). A signal in favor of a potential error may be given by a simple ex post longitudinal check: the comparison between the number of persons employed at times $(t+1)$ and (t) . If the former amount is higher than the latter, it is not possible that the number of job vacancies declared at time (t) was zero.
- The firm does not declare anything. Also in this case, longitudinal checks may be useful for making proper changes, but they may not be enough and the binary variable presence/absence of job vacancies will be object of estimation (for instance, using a logistic model where the explicative variables are often given by past responses provided by the same unit) or will be asked again to the firm (when it will be possible, according to budget and time constraints).

An example of the kind b) is given by tendency surveys. Tendency surveys concern enterprises and consumers and are aimed at asking a series of qualitative questions related to economic situation, household budget, purchases planning, employment, prices, etc. Questions ask for opinions concerning the development of each issue with respect to a previous period. Normally response modalities are: i) strong increase, ii) increase, iii) no change, iv) decrease, v) strong decrease. Macro figures are calculated as weighted differences between optimistic opinions i)+ii) and the pessimistic ones iv)+v). In tendency surveys main quality checks do not refer explicitly to past longitudinal data. This may be due to the use of rotated samples and/or to the weak correlation between responses provided by the same unit in two consecutive survey waves. The basic control is that for each unit and each question one and only one response must be provided.

3. Design issues

The design of the editing and imputation process should be part of the design of the whole survey process. In the frame of editing and imputation procedures three main logical phases are usually carried out, based on the following actions:

1. Identification and elimination of errors that are evident and easy to treat with sufficient reliability, that can involve both interactive and automatic methods;
2. Selection and treatment of influential errors through a careful inspection of influential observations; automatic treatment of the remaining non influential errors, through a selective editing procedure;
3. Check of the final output looking for influential errors that have been undetected in the previous phases or introduced by the procedure itself, that involves macro-editing procedures.

In a longitudinal context, the identification and the calculation of a set of indicators based on macrodata may be based on ratios between the same macrodata related to two different periods, where macrodata of the previous period are supposed to be good (already validated at previous occasions). If the macro indicator falls inside an acceptance range, then no other controls are needed, otherwise it is necessary to go back to microdata and to run again all or a part of controls already activated in the previous micro-editing phase a). Usually, acceptations intervals for macro indicators are determined according to subjective choices by survey experts.

Finally, in the last phase, provisional publication figures are elaborated and analysed using historical data or external sources. If the aggregate figures are implausible, the individual records are examined in order to check for further outliers or error affecting influential records; in these cases data can be modified if necessary. The errors detected at this stage may have been not individuated in the earlier phases of the editing process, or may have been introduced by the process itself. Anyway, also every treatment of these kinds of errors is always made at micro level. If the provisional figures are plausible, the detection of errors and their treatment process is concluded.

The edited file is used in the subsequent statistical process for aggregation purposes, for the estimation of totals and for further analyses.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Chen, S. and Xie, H. (2004), Collection Follow Up Score Function and Response Bias. *Proceedings of the SSC Annual Meeting – Survey Methods Section*, Statistics Canada, 69–76.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.

http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Gismondi, R. and Carone, A. (2008), Statistical criteria to manage non-respondents’ intensive follow up in surveys repeated along time. *Rivista di Statistica Ufficiale*, 1/2008, 5–29.

Granquist, L. and Kovar, J. (1997), Editing of Survey Data: How Much Is Enough? In: L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, 415–435.

Hidirolou, M. A. and Berthelot, J. M. (1986), Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology* **12**, 73–83.

McKenzie, R. (2003), *A Framework for Priority Contact of Non Respondents*. Available at: www.oecd.org/dataoecd.

Philips, R. (2003), The Theory and Application of the Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Proceedings of the SSC Annual Meeting – Survey Methods Section*, Statistics Canada, 121–126.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Different Types of Surveys
2. Repeated Surveys – Repeated Surveys
3. Sample Selection – Main Module
4. Statistical Data Editing – Main Module
5. Statistical Data Editing – Selective Editing
6. Weighting and Estimation – Main Module

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

1. 2.5 Design statistical processing methodology
2. 5.3 Review, validate and edit

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

1. Data validation

Administrative section

14. Module code

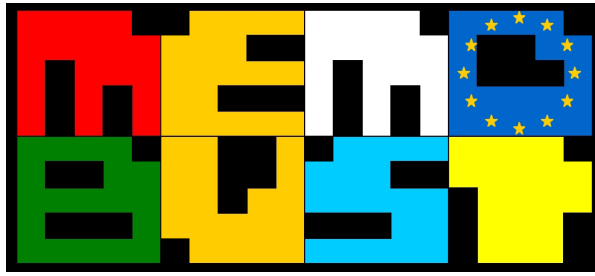
Statistical Data Editing-T-Longitudinal Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-02-2013	first version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.2	30-05-2013	second version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.3	20-08-2013	third version (accepted corrections)	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4	15-11-2013	fourth version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4.1	20-12-2013	preliminary release		
0.4.2	08-01-2014	final release	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:13



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Derivation of Statistical Units

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction	3
2.2 New definitions	5
2.3 Statistical units	5
2.4 The derivation of the statistical units.....	7
3. Design issues	13
4. Available software tools.....	13
5. Decision tree of methods	13
6. Glossary.....	13
7. References	14
Interconnections with other modules.....	15
Administrative section.....	16

General section

1. Summary

This module describes the derivation of the main statistical units which should be made available in a Statistical Business Register (SBR) in order to be used in the production of statistics.

The following units are described: Enterprise Group, Enterprise and local unit. Enterprise groups are a combination of one or more enterprises which operate in certain location (local units). Where the enterprise group is the unit for making statistics concerning financing, the enterprise is aimed at production and the local unit divides the information on enterprise level into geographical parts.

The enterprise group is often first determined based on the result of finding the largest combination of legal units under common control. These enterprise groups have one or more market oriented activities which they carry out. Often these activities will result in enterprises. The enterprises carry out their activities on specific locations. Based on the different locations of the enterprises where the actual activities are carried out, it is possible to derive local units.

2. General description

2.1 Introduction

Statistical units are entities about which information is sought and about which statistics are ultimately compiled. Statistical units are at the basis of statistical aggregates. The different regulations concerning (the use of) statistical units are aimed at (international) comparable statistics, which cannot be realised unless standardisation is applied to both definitions and classifications. One prerequisite to be able to compare two or more statistical collections, which cover the same economic activity over time, is that the comparison applies to the same units. The statistical unit serves as a tool to measure in an unduplicated and exhaustive manner several aspects of the economy.

The following statistical units for the production system are defined (European Parliament, 1993):

- the enterprise;
- the institutional unit;
- the enterprise group;
- the kind-of-activity unit (KAU);
- the unit of homogeneous production (UHP);
- the local unit;
- the local kind-of-activity unit (local KAU);
- the local unit of homogeneous production (local UHP).

For a detailed discussion of statistical units, the reader is referred to the module “Statistical Registers and Frames – The Statistical Units and the Business Register”. Part of this discussion is repeated here to make this module as self-contained as possible.

NOTE: Each (economic) statistic is created with a certain target population which consist of a certain statistical unit.

Example:

- *Structural Business Statistics (SBS) cover the 'business economy' (NACE Rev. 2 Sections B to N and Division 95) which includes industry, construction, and distributive trades and services. SBS are based on the Enterprise.*
- *National Accounts describe the economic activity of a nation. NA describe the production process according to statistical units which are defined according to their economic behaviour, economic function and economic objectives. NA are based on the institutional unit.*

There is an administrative way and there is the statistical way of looking at the different units.

Figure 1 depicts several important relationships:

- the relationship of administrative units;
- the relationships between the different statistical units;
- the relationship between the administrative and the statistical world.

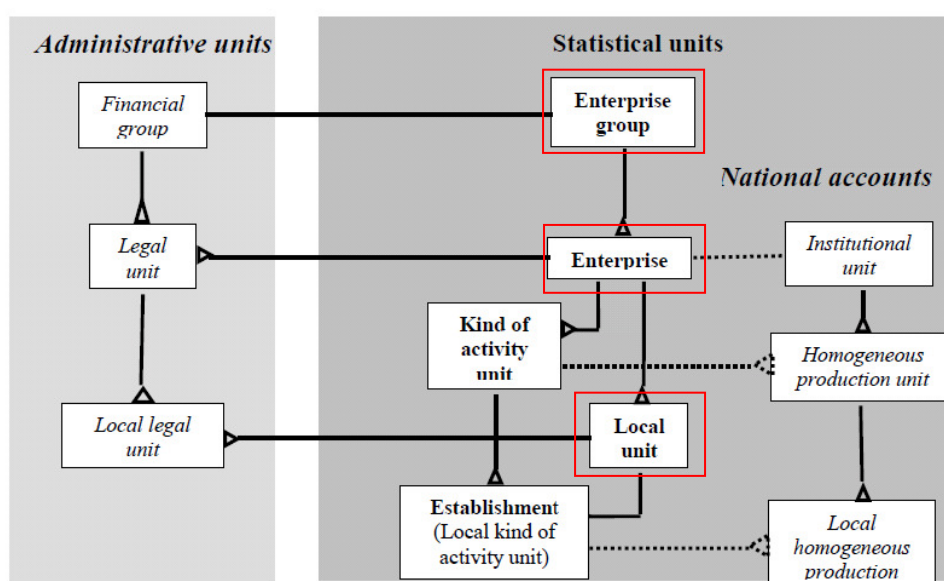


Figure 1. Relationship between the different statistical units

The SBR should hold the following statistical units: the Enterprise, the Enterprise Group and the local unit (European Union, 2008). Only these must be included in the SBR¹. The statistical units defined in the SBR regulation concerns only those statistical units which are needed for statistics. The other statistical units (e.g., KAU, UHP) are defined in different regulations. The Legal unit is also defined in

¹ The legal unit is not an actual statistical unit, but it is often an important building block for deriving the statistical units.

this regulation as the legal unit (and the local legal unit) is in many cases the starting point for creating the statistical units.

2.2 *New definitions*

In 2012 Eurostat started with an investigation for the revision of the statistical unit regulation. The new definitions are still being defined and it will take some time before the new regulation will be accepted. As a result of this, this module uses the ‘old’ statistical unit definitions as they are defined in council regulation 696/93.

2.3 *Statistical units*

One of the most important goals of economic statistics is to describe the economic transactions (and their developments) not on a micro level but in an aggregated way on a macro level. The aggregated information provides insight of a certain group over a certain amount of time. Economic activities are performed by individuals and organisations of individuals. When describing the economic process in a correct and real way it is needed to identify units which act in reality as the actual actors in the economic process.

Actual actors can be read as autonomous units which are part of the process. Autonomy should be treated from an economic aspect (free to decide on production factors, etc.). Autonomy can also be present on different levels: global, national, regional.

There are two main statistical units, the Enterprise Group and the Enterprise. The enterprises have local activities which are carried out in local units. The aims of both statistical units is completely different, but they are closely related. It is assumed that correct information concerning financing, profit, accounts, etc. can be obtained at the level of the Enterprise Group, whereas relevant information concerning the production process (e.g., turnover, value added, persons employed, etc.) can be obtained at the level of the Enterprise.

This section describes these statistical units.

2.3.1 *Enterprise*

The key statistical unit is the enterprise. This unit describes the actual active actors in the market oriented production process (of services and goods).

Since legal units are a construct of law and administration and thus do not always reflect economic reality, there is a need for creating Enterprises. There may be legal or fiscal advantages of separating production factors into two or more different legal units (see the box below for an example). In the economic view, these individual legal units cannot act without the others and should be seen as one unit.

The Enterprise is an economic entity which can correspond to a grouping of several legal units. Some legal units, in fact, perform (ancillary) activities exclusively for other legal units and their existence can only be explained by administrative factors (e.g., tax reasons), without them being of any economic significance (United Nations, 2007).

Example: Consider an enterprise group in NACE 3512 Transmission of electricity. Within this group certain enterprises can be operational. One example of an enterprise can be trade of electricity.

Example: There are two legal units which are part of the same Enterprise Group, one is a production legal unit and the other is a transport legal unit. This transport legal unit exclusively exports goods for one other company which is part of the same Enterprise Group.

In this example the transporting legal unit is not market oriented and can be seen as an ancillary activity within this enterprise. As a result this legal unit should be included into the same enterprise as the actual production unit is.

2.3.2 Enterprise Group

In some cases enterprises are grouped together under the control of the same (ultimate) owner. This is done to achieve economic advantages (such as economies of scale, control of a wider market, etc.). The integration of enterprises into one group can be vertical or horizontal. The enterprise group as a unit is useful for financial analyses and for studying company strategies. Often the enterprise groups are too varied in nature to serve as a unit for statistical surveys and analysis. For this reason the enterprise should be used.

An enterprise group is a set of enterprises controlled by the group head. The group head is a parent legal unit which is not controlled either directly or indirectly by any other legal unit. All the subsidiary enterprises of the enterprise group are considered to be (indirect) subsidiaries of the parent enterprise. It is useful to recognise all (majority and minority) links between the group head and the controlled enterprise via the network of subsidiaries and sub-subsidiaries. This allows the group's entire organisation to be depicted (United Nations, 2007).

Enterprise groups take decisions which might have an impact on the production process and might affect the whole group. This is dependent on the aspect of autonomy. Enterprises have a certain degree of autonomy for which they are responsible for taking decisions separate of the whole of the group. It seems logical to take the enterprise group as a starting point when finding the enterprises belonging to the enterprise group.

Enterprise groups are not bounded by geographical borders. Enterprise groups often divide their activities over different countries depending on their special needs. This module describes the 'national' part of the enterprise groups².

2.3.3 Local Unit

An enterprise is often active at more than one location, and for some statistical purposes³ it is useful to see this geographical distinction.

² As stated before in Section 2.2, the definitions of the statistical unit are currently being revised. One of the modifications will be the international aspect which will be included in the new definition or in the implementation of the new definition.

³ Statistics on regions enables identification of more detailed geographic patterns and trends concerning production (factors) than national data.

A geographically identified place must be interpreted on a strict basis: two units belonging to the same enterprise at different locations (e.g., two different addresses) must be seen as two local units.

There is a direct relationship between the enterprise and the local unit. Logically this requires that the enterprises are derived first before the local units can be derived.

Since the enterprise can have more than one activity which may result in different Kind of Activity Units within the enterprise, it is also possible that local units carry out more than one activity. These different activities can be allocated to different local kind of activity units.

2.4 The derivation of the statistical units

This section describes the delineation of enterprise groups, enterprises and the local units. This description poses no requirements for the organisational structure within a statistical office or on the implementation of the (business register) systems where the statistical units are stored.

The following description assumes that for the delineation of the statistical units all information is available. With all information is meant not only all information available in administrative sources, but also statistical information, international trade information, possible information as a result of direct contact with the enterprise group etc.

As stated before the enterprise group is the best starting point for delineating the enterprises. Therefore the construction of the enterprise groups is the first step.

2.4.1 Legal Unit

Before the derivation of the statistical units can be started, the legal units must be identified and described. Without legal units it is difficult to define statistical units and almost impossible to identify the link between statistical units and administrative information. Legal units are the building blocks of the statistical units.

Legal units include:

- legal persons whose existence is recognised by law independently of the individuals or institutions which may own them or are members of them;
- natural persons who are engaged in an economic activity in their own right.

The legal unit (or part of it) forms, either by itself or sometimes in combination with other legal units, the legal basis for the statistical unit known as the 'enterprise'.

2.4.1.1 Economic/statistical relevance

Administrative sources most often have different goals for registering their administrative units. An important aspect of the SBR is to filter out those units which are not relevant to describe the national economic statistical figures.

Economic relevance can roughly be divided in two parts, financial (Enterprise Group) or production (Enterprise)⁴. This is often done for the business demography, but legal units can exist without

⁴ Different statistics might demand different definitions concerning inactivity.

carrying out any economic activity. They are legally alive and have a legal personality but are economically 'inactive'. A few examples of inactive/dormant legal units are:

- businesses set up to facilitate inward or outward overseas investment or for other international trade purposes;
- businesses that are not yet trading but have registered with the intention of starting trading in the future;
- businesses that have ceased trading but not yet de-registered as legal units;
- businesses that are only active during a specific period in the year;
- etc.

These legal units might be inactive or dormant (used when periodically inactive) but could play an important role when creating the cluster of control which results in the set of legal units which defines the enterprise group.

In order to define if a legal unit is economically relevant a few characteristics can be used. Examples are: the persons employed, the activity of the legal unit, turnover. Two clear indications that a legal unit is economically active and should be part of an enterprise:

- if persons employed > 0;
- if persons employed = 0 and turnover > 0.

This indicates that this legal unit is statistically relevant and should lead or be part of an enterprise.

Also information on ancillary activities, indications of bankruptcy might be used as input in the decision whether a legal unit is economic active.

NOTE: The above does not state that the legal units which are not economically relevant will not or cannot be part of enterprises. It only states that these legal units are not economically relevant by themselves and therefore will not be the cause of the derivation of an enterprise.

Often these units have activities which they perform solely for other units within the enterprise group.

2.4.2 Enterprise Group

2.4.2.1 Structure of the Enterprise Group

Enterprise groups are an association of enterprises bound together by legal and/or financial links. As an operational definition, it can be stated that the enterprise group is the largest collection of legal units which are under the same control. This set/combination of legal units is derived based on the legal and/or financial control links which these legal units have among each other. Control is exercised on whole legal units.

Based on the control aspect, build the cluster of control which will result in the largest set of legal units which are controlled by the same ultimate unit. See Figure 2 which provides an example how to delineate this largest set of units under common control.

In this figure the 'structure of ownership' shows all available ownership relationships between the legal units. Some of the relationships are control relationships, some will result in 'indirect' control, some are minority ownership relationships which will not be part of a control chain. The structure of control is the result of deriving the complete cluster of control relationships between the legal units and defining which legal units are under common control. Some conclusions based on Figure 2 are:

- Legal unit A controls legal unit B and C and, as a result, legal unit A indirectly controls legal unit D for 70%.
- Legal unit E is controlled by legal unit A and legal unit F both for 50%. As a result no legal unit has the absolute control of legal unit E. Since legal unit E controls no other legal unit this legal unit is an Enterprise Group by itself.

NOTE: Figure 2 is a simplified model depicting an example of deriving the units which are under common control.

Since control is a very complex concept which has different meanings and exceptions, the model described is just to provide an idea.

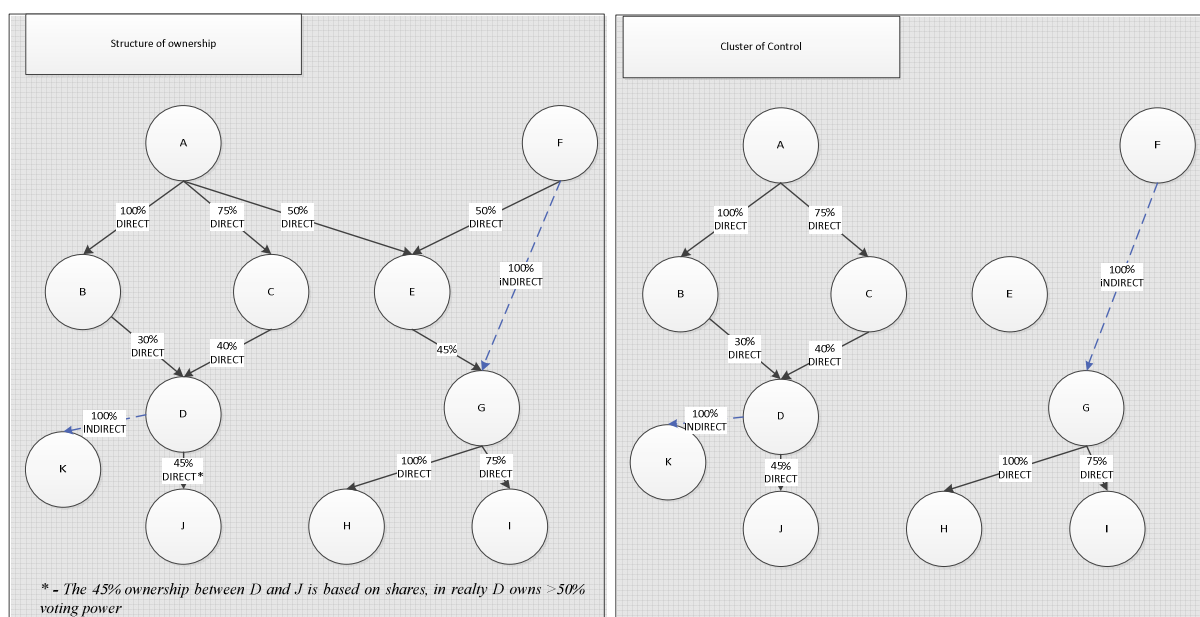


Figure 2. Structure of ownership and the Cluster of Control

Within the set of legal units at least one of the legal units has to be statistically relevant (from a financial point of view). If none of the legal units involved is statistically relevant, no active Enterprise Group can be created, since the whole group is not economically active (the group can be seen as dormant)⁵.

⁵ These non-active groups can be included in the business register as dormant.

2.4.3 Enterprise

Delineating the Enterprises starts with the complete set of legal units which are defined in the Enterprise Group. Within this set of legal units, subsets of legal units have to be identified which will result in Enterprises.

2.4.3.1 Structure of the Enterprise

Within each enterprise group at least one activity is exercised, but it is possible that more than one activity is performed. This section describes the delineation process of enterprises within this set of legal units (which result in an enterprise group).

The structure of an enterprise is delineated in the following steps.

1. Investigate the operational structure (activities) of the enterprise group

The aim of this step is to describe the organisation of the enterprise group from an operational point of view. With the operational structure is meant which are the most important activities of the enterprise group for their environment.

The operational structure of a group is often different from the administrative structure of the group. The administrative structure is designed to have financial or legal benefits. In this step it is important to discover those parts of the group which play a vital role. These parts will result in the activities of the group. The organisation chart of a group provides these insights.

For each activity the following information has to be defined:

- Internal/external flows have to be quantified. Is this activity for the majority intra group oriented?
- Degree of autonomy. Possible indications are the liberty to choose their own suppliers, decide on their own marketing, in what way they can determine the deployment of production factors. If possible determine the intra group flows, which will provide information on the dependencies between the different operations.
- In what way is this activity described? Has the company organised their bookkeeping around these operational structures?

2. Identify the activities which are externally oriented

Enterprises are externally oriented. Their products/services are meant for outside the Enterprise Group. Based on the information which was extracted in the previous step it is possible to identify the externally oriented activities.

3. Combine the activities which are externally oriented into autonomous clusters

Autonomy (in its different forms) is a key attribute of an enterprise. Autonomy only exists when to a certain degree the enterprise can make its own decisions and is not dependent on intra group entities for their survival.

For each operation it is known whether it depends on another operation. This information was gathered in a previous step. Based on this information externally oriented clusters will be combined with not externally oriented operations into autonomous clusters. If externally oriented clusters depend on (larger) not externally oriented clusters, the externally oriented operation

carries out the last part in the production process (e.g., the selling). This step ensures that the not externally oriented activities are part of one of the enterprises. This combining of activities is done when they comprise a production process. Criteria for this could be:

- a. Shared use of production factors
- b. Having the same management

The combining into one autonomous cluster is useful only if the company is able to provide relevant figures for this entity.

4. Extend these autonomous clusters with operation entities (not externally oriented) that serve only them. This concerns the allocation of operational entities of the EG that are not externally oriented and that provide products to exactly one other operational entity belonging to the autonomous cluster. Such operational entities may carry out an ancillary activity, but this is by no means necessary: they may provide an intermediate product which is processed further by the autonomous cluster or which may form part of the product of the cluster (so-called partial activities).
5. Allocate all other not externally oriented activities to the autonomous clusters. It is possible that there are still operational entities left. These are obviously not externally oriented and are by necessity oriented towards more than one autonomous cluster. The allocation of these operational entities would have to be based, in principle, on the proportion of the flows of goods and services.

The next step is not essential in the derivation of the enterprise, but is a vital step in enabling the use of administrative data. Administrative data can be aggregated to the enterprise level which results in new relevant information.

6. Link enterprises to the legal units which are part of an Enterprise Group. The result of this step is that data from administrative sources for describing enterprises can be used, by linking enterprises to legal units which consist of administrative units.

NOTE: Enterprise Groups vary in complexity. The majority of the Enterprise Groups are relatively simple groups which can be derived in an automated way. Large complex Enterprise Groups on the other hand can be difficult to derive. These groups often have very complex legal/administrative/financial constructions which provide them the best position in the real world. The challenge is to derive the correct statistical units from this complex construction.

The complex Enterprise Groups require a manual approach (profiling) since automated rules which treat easy and complex constructions lead to very complex algorithms.

The profiling process defines the statistical units in collaboration with the Enterprise Group. This collaboration ensures that the Enterprise Group can also provide meaningful information on the statistical units which have been manually created.

NOTE: The derivation described above assumed that all possible information is available. In many cases this is not the case (much information is not accessible or referring to the correct unit or period) and when it is the case it is a costly process using all those data.

Many countries also feel the pressure of making more use of administrative data in order to lessen the administrative burden on the enterprise and also to lessen the burden on the statistical office in deriving the statistical unit.

A challenge with administrative data is that they provide insight in a legal/administrative world. This legal/administrative world has different purposes than describing the actual economic situation. The challenge is defining a way to interpret the legal/administrative world into a statistical world.

For a more detailed discussion on the challenge of using administrative data the reader is referred to the module “Data Collection – Collection and Use of Secondary Data”.

2.4.4 Local Unit

Enterprises carry out their activities at certain geographical places. The geographical information of the enterprises should be used to derive the local units.

2.4.4.1 Structure of the Local Unit

Collect all information concerning the actual locations where the enterprise carries out their activities. Keep in mind that:

- Local units should have a unique location. This location is characterised by the address, a street or a region where it is located. A rule which can be used is that at the location of the local unit products, materials could be stored.
- Local units can, as a general rule, only exist if at least one person is working at that certain location. But sometimes it consists of only a PO-Box where nobody is working.
- An enterprise can have one or more locations where activities are carried out. Therefore the enterprise can consist of more than one local unit.
- Activity of the local unit. There must be consistency between the local units figures compared to the enterprise figures. For this to be achieved it is essential that the activity code of the local unit is equal to the activity code of the enterprise to which it belongs. This can be achieved by using the activity code of the enterprise on local unit level. Each local unit has its own economic activity (code), but in order to obtain consistency between national and regional accounts, one may use the economic activity (code) of the enterprise to compile regional economic indicators.

Example on deriving local units using legal local units:

One of the possible sources for geographical information on the enterprises can be derived from the legal local units. A legal local unit is a part of a legal unit that is located at a certain address. In other countries where there are no local legal units other types of information on geographical locations can be used. A legal local unit can operate in several different industries. In practice, a legal local unit is in most cases equal to the local unit. By linking these local legal units to the legal units and the enterprises the geographical activity is available for these enterprises. Local units are created within Enterprises. One important rule is that regional figures can be added up to the higher level.

Figure 3 shows an example how to derive these local units.

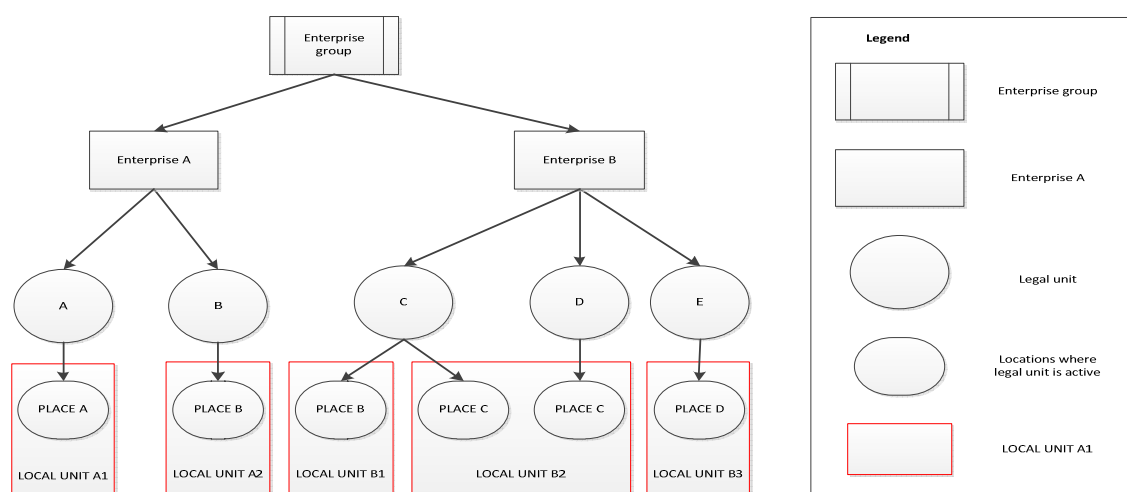


Figure 3. Local Units

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

European Union (1993), *European Parliament and the Council of the European Union [1993]: Council regulation (EEC) no 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the community.*⁶

European Union (2008), *Regulation (EC) no 177/2008 of the European Parliament and of the Council of 20 February 2008 establishing a common framework for business registers for statistical purposes.*⁷

Eurostat (2010), *Business registers, recommendations manual.* Luxembourg.⁸

United Nations (2007), *Statistical Units.* New York.⁹

⁶ <http://eur-lex.europa.eu/lexuriserv/lexuriserv.do?uri=celex:31993r0696:en:html>

⁷ <http://eur-lex.europa.eu/lexuriserv/lexuriserv.do?uri=oj:l:2008:061:0006:01:en:html>

⁸ <http://ec.europa.eu/eurostat/ramon/statmanuals/files/ks-32-10-216-en-c-en.pdf>

⁹ <http://unstats.un.org/unsd/isdts/docs/statisticalunits.pdf>

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Registers and Frames – Main Module
2. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
3. Statistical Registers and Frames – Building and Maintaining Statistical Registers to Support Business Surveys
4. Statistical Registers and Frames – Survey Frames for Business Surveys
5. Statistical Registers and Frames – The Design of Statistical Registers and Survey Frames
6. Statistical Registers and Frames – The Statistical Units and the Business Register
7. Statistical Registers and Frames – Quality of Statistical Registers and Frames
8. Data Collection – Collection and Use of Secondary Data

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

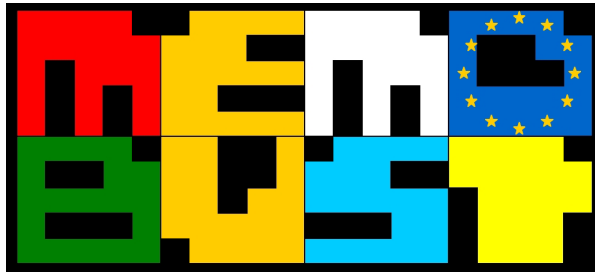
Derivation of Statistical Units-T-Derivation of Statistical Units

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	22-04-2013	first version	Barry Coenen	CBS (Netherlands)
0.2	18-09-2013	second version	Barry Coenen	CBS (Netherlands)
0.3	04-11-2013	updated based on review	Barry Coenen	CBS (Netherlands)
0.4	06-01-2014	updated based on review	Barry Coenen	CBS (Netherlands)
0.5	04-02-2014	updated based on review	Barry Coenen	CBS (Netherlands)
0.5.1	06-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:40



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Business Demography

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Introduction	3
2.2 Main Components of Business Demography	4
2.3 New Enterprises (Births)	4
2.4 Ceased Enterprises (Deaths).....	11
2.5 Surviving Enterprises	12
2.6 Business Demography Indicators	13
2.7 Background on Entrepreneurship	14
2.8 Employer Enterprise Birth.....	15
2.9 Employer Enterprise Death	17
2.10 Employer Surviving Enterprise Definition.....	18
2.11 High Growth Enterprise Definition	18
2.12 Gazelle Enterprise Definition	19
2.13 Entrepreneurship Indicators.....	20
3. Design issues	20
4. Available software tools.....	20
5. Decision tree of methods	20
6. Glossary.....	20
7. References	20
Interconnections with other modules.....	22
Administrative section.....	23

General section

1. Summary

The term business demography is used to cover a set of variables which explain the characteristics and demography of the business population. The creation of new enterprises and the closure of unproductive ones are considered important indicators of the business dynamics.

There is a large demand for information on business demography both at national and international level. At European level, demands are for coherent and comparable data across the members of the European Statistical System (ESS). The European Commission, as key customer, has assured its commitment to a policy that promotes entrepreneurship as an essential instrument for improving competitiveness and generating economic growth and job opportunities since its communication to the Council on 'Promoting Entrepreneurship and Competitiveness'. The support of entrepreneurship and entrepreneurial dynamics, the presence of which can be revealed by the analysis of business demography statistics over time. As a consequence, there is high demand for comparable data on business demography for the purposes of monitoring and policy formulation.

This module aims to provide theoretical guidance in the production of data on business demography. It has been developed taking Eurostat-OECD Manual on Business Demography Statistics into account, used in the EU business demography harmonised data collection.

After a short introduction on the importance to collect business demography data, in the second section we give a general description of two main topics: first we describe the methodology and indicators for business demography statistics such as: Enterprise Births, Enterprise Deaths and Surviving Enterprises; second we describe methodology and indicators for entrepreneurship such as: Employer Enterprise Births, Employer Enterprise Deaths, Employer Surviving Enterprises, High-Growth Enterprises and Gazelle Enterprises.

2. General description

2.1 Introduction

The term business demography is used to cover a set of variables which explain the characteristics and demography of the business population. The creation of new enterprises and the closure of unproductive ones are considered important indicators of the business dynamics.

Business demography data are currently produced to fulfil the European regulation, for the European Union (EU) and European Free Trade Association (EFTA) members; it is used to satisfy the requirements for producing the Structural Indicators used for monitoring progress of the Lisbon process, regarding business births, deaths and survival. It also provides key data for the joint OECD-Eurostat "Entrepreneurship Indicators Programme".

A new methodology has been developed for the production of data on enterprise births (and deaths), that is, enterprise creations (cessations) that amount to the creation (dissolution) of a combination of production factors and where no other enterprises are involved. This methodology aims to harmonise how business demography is computed within the ESS. The present module is an overview of this methodology fully described in the [Eurostat-OECD Manual on Business Demography Statistics](#). The

reader who is interested in more details and background can read this manual that is available from the Eurostat website.

The methodology and definitions adopted in this module are also based on the Business Registers Recommendations Manual and indications, as the Business Registers is the source for the Business Demography data. In addition we focus on business demography output on a yearly basis.

2.2 Main Components of Business Demography

The real enterprise birth and death definitions must take into consideration the enterprise definition. A fundamental requirement in measuring business entries (creation) and exits (destruction) concerns the definition of a **business** itself. Statistical System distinguishes a number of unit type such as: establishment, enterprises, legal units, local kind of activity units, etc. (See the module “Statistical Registers and Frames – The Statistical Units and the Business Register”.) In order to harmonise the business demography data collections, the statistical unit to be used is the enterprise. According to the statistical units Regulation (Council Regulation (EEC) No 696/93 of 15 March 1993) “*The enterprise is the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit.*”

In the next subsections we give an overview of the main Business Demography components: Enterprise Births, Enterprise Deaths and Surviving Enterprises. In addition we give an overview of the main Entrepreneurship components: Employer Enterprise Births, Employer Enterprise Deaths, Employer Surviving Enterprises, High-Growth Enterprises and Gazelle Enterprises.

For each of these indicators we give the definition, a description of identification process and, where it was possible, a case study.

2.3 New Enterprises (Births)

The number of Enterprise Births (*RB*) is a key variable in the analysis of Business Demography; other variables such as the survival and growth of newly born enterprises are relevant and related to this concept. The production of statistics on births must be based on a clear and acceptable definition and interpretation.

2.3.1 Concept

According to the Commission Regulation No 2700/98, we define a new Enterprise (Birth) as:

A birth amounts to the creation of a combination of production factors with the restriction that no other enterprises are involved in the event. Births do not include entries into the population due to: mergers, break-ups, split-off or restructuring of a set of enterprises. It does not include entries into a sub-population resulting only from a change of activity. An enterprise creation can be considered an enterprise birth if new production factors, new jobs in particular, are created.

Inclusions

Enterprises started by a person who previously performed the same activity, but as an employee should be included in the statistics on enterprise births.

Exclusions

Events leading to a creation of a new enterprise, but which should be excluded from the statistics on enterprise births are:

1. Enterprises that are created by merging production factors or by splitting them into two (or more) enterprises (breakups, mergers, split-offs, restructuring)¹;
2. Newly created enterprises that simply take over the activity of a previously created enterprise (take-over)²;
3. Any creations of additional legal units/enterprises solely for the purpose of providing a single production factor (e.g., the real estate or personnel) or an ancillary activity (see note below) for an existing enterprise.
4. An enterprise that is registered when an existing enterprise changes legal form. E.g., a successful sole proprietor moves operations from his home to another location and at the same time changes the legal form of the enterprise to a limited liability company.
5. Reactivated enterprises if they restart activity within two calendar years³.

2.3.2 The identification process

Users that want to know how many new enterprises have been created in a specific year usually compare populations of active enterprises referred to two adjacent periods (t , $t-1$).

According to the [Eurostat-OECD Manual on Business Demography Statistics](#) the Business register serves as primary and preferred source of information for business demography statistics. The main reasons that we choose this source are: -there is a degree of harmonisation of statistical business registers in EU Member States following the adoption of the business statistical Regulation (Council Regulation (EEC) No 2186/93); - under the EU Regulation, Member States are required to hold data on the enterprise, a harmonised statistical unit that removes the impact of different legal and organisational infrastructures; - using data from business registers minimised the burden on businesses. The BR contains the population of active enterprises. This population consists of all enterprises that had either turnover or employment at any time during the reference period. A unit present into the BR is tested according to a methodology determining whether it is active or not active on the basis of these signals of activity (presence of turnover or employment); on the basis of this information entries can be identified in a population of active units.

The identification of real enterprise births is based on the application of a complex and expensive procedure that is build up on a set of automated and manual steps aiming to eliminate (i.e., to identify) the “non-real” components from the set of new enterprises (entries) units into the BR for a reference period t .

A formalisation of the demographic flow is given by the following equation:

$$N_t = N_{(t-1)} + E_t - U_{(t-1)} = A_{t-1,t} + E_t \quad (1)$$

¹ For more details see the Glossary of this module.

² For more details see the Glossary of this module

³ For more details see you Eurostat-OECD Manual of Business Demography Statistics pp.34-35

where:

N_t = population of active enterprises in a reference period t

$N_{(t-1)}$ = population of active enterprises in a reference period $(t-1)$

E_t = Entries in a reference period t . We define Entries in year t as the subset of the population of active enterprises in year t , which have taken up economic activity between 01.01 and 31.12. These new enterprises are identified as enterprises that are only present in year t and not in year $(t-1)$.

$U_{(t-1)}$ = Exits from the reference period $(t-1)$. We define Exits as the subset of the population of active enterprises in year $(t-1)$, which have ceased their economic activity in year $(t-1)$. These enterprises are identified as enterprises that are only present in year $(t-1)$ and not in year t .

$A_{t-1,t}$ = active enterprises both in the period t and $(t-1)$

Entries (and exits) are affected only by the structure of signals of activity (presence/absence of turnover or employment) or by other national methodology establishing if the enterprise is active or not.

Regarding the new enterprises (entries), some steps are followed to reach the final result, i.e., the real enterprise births (data on enterprise deaths are produced with a “mirror” process). These steps are performed for the whole population of active enterprises.

We summarise all the process into three steps each determining a relative subset of data.

Step 1 – The first step consists in comparing three subsequent populations of active enterprises (N_t , N_{t-1} and N_{t-2}). This operation allows to eliminate reactivated units.

A merge by identification code (or fiscal code) of the three subsequent populations of active enterprises determines the following pattern:

$t-2$	$t-1$	t	Output	
N_{t-2}	Missing	N_t	E_r	Reactivations
Missing	Missing	N_t	E_1	Entries

In detail, it is necessary to split the Entries (E_t) into two components: 1) E_r the subset of reactivations; 2) E_1 the subset of entries cleaned from reactivations.

The resulting equation (1) becomes:

$$N_t = A_{t-1,t} + [E_1 + E_r] \quad (2)$$

Step 2 – Mergers and other events of structural changes involving enterprises’ re-organisation cause creations of new units into the register. The identification of births is carried out by eliminating creations due to these events (break-ups, split-offs, mergers and take-overs). It possible to have information about these events from different sources (pilot studies, administrative sources, statistical sources, survey etc...).

In the previous equation E_1 is split into two components according to the following formula:

$$N_t = A_{t-1,t} + [(E_2 + E_{ev}) + E_r] \quad (3)$$

where:

E_2 = entries not due to events of structural changes

E_{ev} = entries due to events of structural changes

Step 3 – The further step consists in identifying and excluding those false entries because continuing the activity of some other units. Continuity rules are assigned through a general matching process, that matches units according to economic activity and location, name and location, economic activity and name, and the use of other nationally available information, such as telephone number, date of registration/deregistration at the administrative source, employer/employee links etc. This step contains also the results of manual controls for large enterprise births. E_2 component is split as follows:

$$N_t = A_{t-1,t} + [(RB_t + E_{cont}) + E_{ev} + E_r] \quad (4)$$

where:

RB_t = Real births in the year t

E_{cont} = entries due to the continuity rules application

The RB_t component identifies the subset of real enterprise births.

2.3.2.1 The notion of continuity

Within the study of enterprises population development it is relevant to identify whenever a change happens whether it produces a discontinuity of the unit: an enterprise is considered to be continued if it modifies without any significant change in its identity, in terms of its production factors. The production factors include the set of means (employment, machines, raw material, capital management, buildings) that the enterprise uses in its production process and leading to the output of goods and services. It is clear that measuring the continuity of all production factors and weighting them can be quite difficult and costly. For those reasons Eurostat suggests, as a practical criterion, to use precise variables available in the register that are correlated to the most important production factors for identifying the enterprise: the basic hypothesis is that a change in such variables would stand for a change in the production factors. The variables (characteristics) considered are:

- ⇒ The controlling legal unit of the enterprise (**N**) - The continuity of the management of the enterprise may be assumed to be positively correlated with the continuity of the controlling legal unit. The same may be assumed for some immaterial assets.
- ⇒ The economic activities carried out (**S**) - Continuity of the four-digit NACE Rev.2 code of the principal activity may be assumed to be positively correlated with the continuity of the production factors, especially employment, machines and equipment, land and buildings.

⇒ The locations where the activities are carried out (**L**) - The continuity of the locations where the activities are carried out is of course closely linked to the continuity of the land and buildings used by the enterprise.

The empirical rule suggested is that an enterprise is considered to be discontinued if at least two over three modifications in the previous factors occur. Continuity rules reflect a notion of identity based on the consideration that the enterprise is a set of specific resources, procedures and relationships with the environment. In the suggested rules an element of discontinuity is introduced when changes are “of great extent” and quick. The concrete applicability of such rules must be evaluated according to the economic structure in which they have to perform, because of the peculiarity of each country. For instance, for some domains of study, as for demography of very small enterprises, it does not make sense to separate the legal subject (the entrepreneur) from the statistical subject (the enterprise); for such cases a new controlling legal unit becomes a factor producing discontinuity even if it is the only one to change.

For the continuity rules, several software and standard matching systems can be applied according to the national sources and experiences.

2.3.3 *Case study: the matching process in the production of Italian Business Demography data*

In Italy, the identification of the “real” components is based on the application of Record Linkage (RL) techniques. The matching process matches on name, economic activity and location of enterprise.

A reminder of the main elements for a record linkage is needed, a detailed description can be found in the module “Micro-Fusion – Object Matching (Record Linkage)”. Let A and B be two files respectively containing records na and nb . The record linkage procedures compare each unit in A with each unit in B . The object of interest of the record linkage problem is the pair of units (a,b) of a set $A \times B$. This set can be partitioned in a set M of pairs representing the same business entity and a set U of pairs representing different entities. Record linkage methods aim at determining which pairs belong to the set M . We can distinguish two approaches: deterministic and probabilistic. For both these two approaches some preliminary steps are necessary.

- 1) Identification of populations (file A and file B) and units.
- 2) Selection of match characteristics (components). This selection should be based on the quality of the available data, their discriminating power and the purpose of the study.
- 3) Standardisation, parsing and string comparison of match variables.
- 4) Blocking (comparison reduction). The size of the files usually considered does not often allow explicit consideration of comparison of all pairs of records, and usually only pairs with some common characteristics are actually compared, by using blocking criteria.
- 5) Agreement/disagreement rules to evaluate the similarity of records and weighting system to take into account that some information is more important than other; for example, an agreement on economic activity can contribute less than an agreement on enterprise names.
- 6) Determination of thresholds. A match is accepted only if its level of agreement is higher than a designated “threshold” level.

For Business Demography purpose, a record linkage technique is applied to find pairs of records across files that correspond to the same entities and to perform matching to identify “real” births and deaths according to the continuity rules.

2.3.3.1 Identification of populations (file A and file B) and units

If the purpose is the identification of **Real Births** the populations compared are:

File A: Stock of enterprises in the year t

File B: Entries of enterprises in the year t

and

File A: Exits of enterprises in the year $(t-1)$

File B: Entries of enterprises in the year $(t-1)$

If the purpose is the identification of **Real Deaths** the populations compared are:

File A: Stock of enterprises in the year t

File B: Exits of enterprises in the year $(t-1)$

and

File A: Exits of enterprises in the year $(t-1)$

File B: Entries of enterprises in the year t

2.3.3.2 Selection of match characteristics (variables)

After having identified the populations, it is necessary to select the matching variables (components). According to **continuity rules**, for Business Demography the **matching variables** are: 1) Enterprise name; 2) Address; 3) Economic activity code (Nace Rev.2 at 4 digits). Other variables are used as support for a better specification of the enterprise name: the Fiscal code and the legal form. When matched variables are identified, their standardisation, parsing and string comparison functions follow.

2.3.3.3 Standardisation, parsing and string comparison of match variables

For Business Demography appropriate standardisation and parsing of enterprise name and address components is crucial for computerised probabilistic or deterministic record linkage.

Standardisation of **enterprise name** consists in three main activities:

- 1) Parsing: to divide the name in sub-components;
- 2) Editing of each sub-component of the name;
- 3) Use of the dictionary to attribute to each sub-component its meaning.

The main difficulty with enterprise name is that even when they are properly parsed, the identifying information may be indeterminate.

Standardisation of **enterprise address** consists in two main activities:

- 1) Parsing: to divide the free form address in three sub-components toponymic (T) (address type as: square, road, avenue etc...), street name (sNa) and street number (sNb);

- 2) Editing of each sub-component of the address.

2.3.3.4 Blocking (comparison reduction)

According to blocking criteria, two alternative blocks are used: **municipality** and **postal code +economic activity code (NACE Rev.2 - 3 digits)**.

After blocking criteria, agreement-disagreement rules on enterprise name and address are applied. These rules came from a set of specific values.

2.3.3.5 Agreement rules

Agreement rules for enterprise name

According to the Italian enterprise structure, enterprise name takes different formats according to its legal form. We aggregate the legal form (J) into 3 classes: (**I**= Sole proprietorship; **Sp**= Partnership; **Sc**= Limited liability company). The adopted rules are:

When comparing $J=(I, Sp)$ or $J=(Sp, Sp) \Rightarrow$ Surname and Name are compared.

If $J=(Sp, Sc)$ or $J=(Sc, Sc) \Rightarrow$ Significant Sub-components (words) are compared.

If $J=(I, I)$ or $J=(I, Sc) \Rightarrow$ We have a disagreement by default.

Agreement rules for enterprise address

We have an outcome for each component of address. The possible outcomes on the toponymic (**T**) component are **Equal** or **Differ** or **Missing**. To compare street name (**sNa**), a simple string comparator is used. This string comparator divides each sNa in consecutive sub-strings of 3 characters, finds the number of common sub-strings and computes the percentage of common sub-strings on the length of the shorter sNa. Finally the possible outcomes on street number (**sNb**) component are **Equal** or **Differ** or **Missing**. Enterprise address matching rules are delineated according to different combinations of components outcomes. For example when you compare two addresses of two different enterprises we have an agreement if toponymic components are equal or missing for both addresses, the percentage of matched address name sub-strings is high ($\geq 80\%$) and the street numbers are equal. We have a disagreement when the toponymic components are different.

Agreement rules for enterprise economic activity

NACE Rev.2 codes at 4 digits are compared to produce outcomes. We have an agreement if NACE codes of two enterprises are equal, we have a partial agreement if NACE codes are compatible and we have a disagreement if NACE codes are different.

Outcomes defined by each *variable* rule constitute the components of the comparison vector γ .

2.3.3.6 Determination of thresholds

For Business Demography purpose, ad hoc matching decision rule is applied to divide the pairs in link and non-link. In presence of a comparison vector having two over three agreements the pair is a match (according to the continuity). There is an exception: if the comparison vector γ is composed by a disagreement on the enterprise name, an agreement on the enterprise address, an agreement on

enterprise economic activity ($\gamma=(D,A,A)$) and the NACE code belong to a list of economic activity “at risk” (construction, hotel..) then the pair is declared as non-matched.

Once you identify the two sets of pairs M and U it is possible to determine the set of Entries due to the continuity rules application (E_{cont}).

$$RB_t = E_t - E_r - E_{ev} - E_{cont}$$

where:

E_t = Entries in a reference period t

E_r = reactivations: enterprises active both in year $(t-2)$ and t , but not active in $(t-1)$

E_{ev} = entries due to events of structural changes

E_{cont} = entries due to the continuity rules application (Record-Linkage).

For the sake of consistency, and in line with user needs, the method of comparing populations of active enterprises used for the production of data for enterprise births should also be followed for enterprise deaths. This will also help to gain from synergies in processing.

2.4 Ceased Enterprises (Deaths)

2.4.1 Concepts

According to the Commission Regulation No 2700/98, we define a Ceased Enterprise (Death) as:

A death amounts to the dissolution of a combination of production factors with the restriction that no other enterprises are involved in the event. Deaths do not include exits from the population due to mergers, take-overs, break-ups and restructuring of a set of enterprises. It does not include exits from a sub-population resulting only from a change of activity.

2.4.2 The identification process

Like the real enterprise births, the identification of real enterprise deaths is based on the same procedure that is built up on a set of automated and manual steps aiming to identify the “non-real” components from the set of exit units into the BR. The cessations in year t are a subset of the population of active enterprises in year t , which have ceased their economic activity between 01.01 and 31.12. They can be identified by comparing the population of active enterprises in year t with the population of active enterprises in year $(t+1)$. Exits are identified as enterprises that are only present in year t . Like enterprise births, exits should be checked for reactivation in the following two calendar years, because enterprises dormant for less than two years are considered reactivations and therefore not deaths. An enterprise death occurs only if the unit has been inactive for at least two years. In order to find the events that were not real enterprise deaths, but rather cessations due to events like break-ups, mergers or take-overs, a matching criteria (as for enterprise births) should be carried out. Finally the continuity rule is applied to identify the cases where another unit is involved in the cessation of the enterprise. As for enterprise births, the matching should consider name, location and economic activity.

According to this procedure for the identification of real enterprise deaths in year t , populations in $(t+1)$ and $(t+2)$ are needed.

In practice, at the time t , it is possible identify only provisional real deaths $(t-1)$ (due to the unknown reactivations at year $t+1$).

For the above reasons, the following procedure for the identification of enterprise real deaths $(t-1)$ determines a one year lag difference compared to the real births (t) . To improve timeliness of real deaths, it is necessary to develop further estimation techniques.

2.4.3 Case study. Real Deaths estimation method: the Italian methodology

To estimate Real Deaths at year t , a time series of enterprise deaths rates of the previous five years is analysed. In addition, nationally available administrative source, the Social Security data SS with information at $(t+1)$ is used to identify employer enterprise deaths.

The estimation method consists in the building of strata made up by three variables: Economic activity, Legal form and Size class. In total we have about 5,600 strata. For each j -th stratum, death rate (t) is given by a average over the period $[t-5, t-1]$. Then it is reweighted using the ratio=number of SS enterprises/ number of BR enterprises if strata presents employees.

Formally:

$$death_rate(t)_j = \left[\frac{1}{5} \sum_{i=1}^5 death_rate(t-i)_j \right] w_j \quad \text{for } j=1, \dots, n \text{ and } i=1, \dots, 5.$$

where:

$$w_j = \begin{cases} 1 & \text{if } j \text{ is with employees } = 0 \\ \left(1 - (n^{\circ}enterprise sSS)_{(t+1)} / (n^{\circ}enterprise sBR)_{(t)} \right) & \text{if } j \text{ is with size classes } > 0^4 \end{cases}$$

If the number of SS enterprises in $(t+1)$ is greater than the number of BR enterprises in year (t) then $w_j < 0$ and consequently the number of deaths in year t are 0.

The number of deaths year t in stratum j is given by the weighted average rate over period multiplied by the number of active units in year t . This technique allows to estimate only provisional real deaths in year (t) (due to the unknown reactivations at year $t+2$).

Actually we are studying a new methodology that takes into account the timeless information about the number of deregistrations in the Chamber of Commerce (administrative data). These types of information is useful to forecast sudden change in death rate due to social or economic events.

2.5 Surviving Enterprises

2.5.1 Concepts

According to the Commission Regulation No 2700/98, the definition of a Surviving Enterprise is:

⁴ If the number of SS enterprises in year $(t+1)$ is equal to the number of BR enterprises in the previous year (t) then $w_j = 0$ and consequently the number of deaths in year t are 0.

an enterprise born in year $(t-1)$ or having survived to year $(t-1)$ from a previous year is considered to have survived in year t if it is active in terms of turnover and/or employment in any part of year t (= survival without changes).

If the enterprise is not active in year t it has survived if its activity is taken over by a new enterprise set up specifically to take over the factors of production of that enterprise in year t (= survival by take-over).

This definition of survival excludes cases where enterprises merge, or are taken over by an existing enterprise in year $(t-1)$. In these cases the continuation of the enterprise involves an enterprise established before year t and therefore the enterprise is not considered to have survived.

To ensure consistency between data on births and survivals, it is important that the identification of cases where an enterprise is taken over by a new enterprise is based on the use of the same information as when evaluating whether a new enterprise is a birth or not. Therefore an enterprise is only counted as survived, if the enterprise that takes over the factors of production is a new enterprise.

2.5.2 *The identification process*

The production of statistics on survival can be based on three populations, which are all part of the production of the statistics on births: a) Births in year $(t-1)$, or enterprises having survived to t from a previous year ($RB(t-1)$); b) Active enterprises in year t ($A(t)$); c) Enterprises that have commenced activity in year t with the purpose of taking over the factors of production of an enterprise that commenced activity before t ($TO(t)$). By matching these populations by identification codes (like the BR code, or a fiscal code) it is possible to have the following outcomes:

- $RB(t-1)$ is present neither in $A(t)$ nor in $TO(t)$. $RB(t-1)$ is not survived in year t .
- $RB(t-1)$ is present only in $A(t)$. $RB(t-1)$ is survived in year t without any change.
- $RB(t-1)$ is present only in $TO(t)$. $RB(t-1)$ is survived in year t with change (take-over took place in $t-1$).

2.6 *Business Demography Indicators*

The business demography data will be used to produce additional indicators related to enterprise births, deaths and survival such as the following:

- Births/Deaths as a percentage of the population of active enterprises (birth/death rates).
- Births/Deaths by size class.

Additional indicators will be produced to demonstrate the impact of the newly born/death enterprises on the economy:

- Persons employed in newly born/death enterprises in year t as a proportion of the total number of persons employed in the population of active enterprises in year t (both in head counts).
- Employees in newly born/death enterprises in year t as a proportion of number of persons employed in newly born/death enterprises in year t (both in head counts).

The first of these indicators reflects the employment creation/destruction potential of newly born/death enterprises. The second reflects the potential employment creation/destruction going beyond the entrepreneurs themselves. Other possible indicators are:

- Gross Job Turnover (**GJT**)= the sum of the number of jobs by creation and destruction during a year, it is a measure of job reallocation.
- Net Job Turnover (**NJT**) = the difference between the number of jobs by creation and the number of job of destruction.

Possible indicators for surviving:

- Survival rate in years t , as ratio between the number of enterprises born in year $(t-i)$ ($i=1$ to n) and survived in year t and the number of enterprise births in year $(t-i)$.
- The number of persons employed in surviving enterprises in their i -th year of activity divided by the numbers of persons employed in real births in the initial year
- The number of persons employed in t divided by the number of survival enterprises in t .
- Although it is not an indicator, the study of the enterprises cohorts surviving from the year of birth $(t-i)$ (usually $i=1$ to 5) to year t is very interesting. The analysis of the characteristics of such enterprises could describe the behaviour of younger enterprises in terms of employment growth or in terms of economic variables growth such as turnover and value-added over time. Further analysis focuses on the study of efficiency and productivity of such enterprises with respect to the older enterprises.

2.7 *Background on Entrepreneurship*

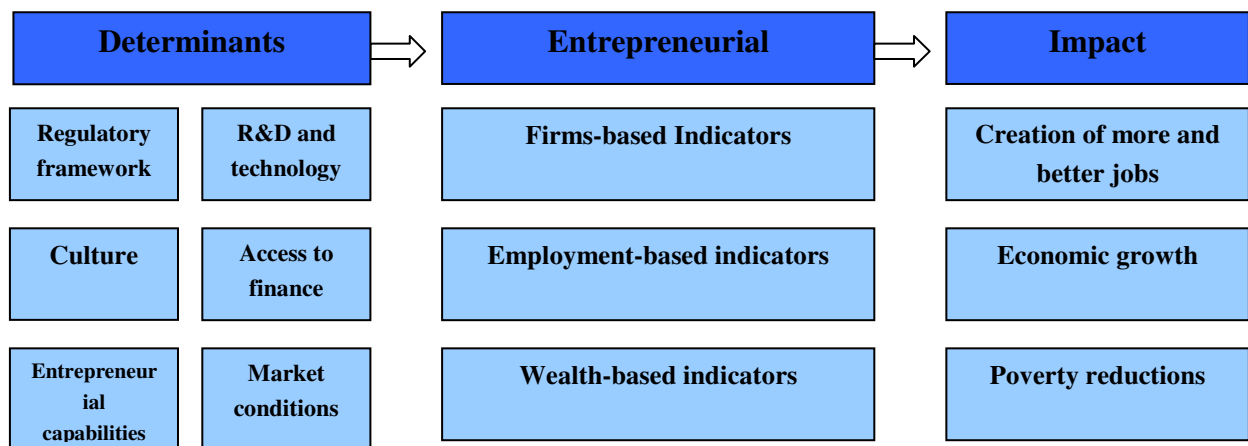
In recent years, the political and academic interest in entrepreneurship and its determinants has grown. Policy makers give more importance to the development of high growth enterprises and to the conditions that foster this growth. The Entrepreneurship Indicators Programme (EIP), launched by OECD in September 2006, has as main goal to build internationally comparable statistics on entrepreneurship and its determinants. In 2007, Eurostat joined forces with the OECD to create a joint OECD-Eurostat EIP, and work began with the development of standard definitions and concepts as a basis for the collection of empirical data.

The OECD-Eurostat approach has tried to combine the more conceptual definitions of entrepreneurship with (available) empirical indicators. The following definitions were established:

- ENTREPRENEURS are those persons (business owners) who seek to generate value through the creation or expansion of economic activity, by identifying and exploiting new products, processes or markets.
- ENTREPRENEURIAL ACTIVITY is enterprising human action in pursuit of the generation of value through the creation or expansion of economic activity, by identifying and exploiting new products, processes or markets.
- ENTREPRENEURSHIP is the phenomenon associated with entrepreneurial activity.

Given the multifaceted nature of entrepreneurship and the myriad factors that may affect, a simple entrepreneurship model was proposed as a first step towards establishing a framework for the

development of empirical indicators that are both relevant and available. These indicators are grouped together into three themes: *Determinant* (factors that impede or motivate entrepreneurship); *Entrepreneurial performance* (measures that provide information of the state of entrepreneurship); and, *Impacts* (outcomes of that performance on the economy as a whole).



Focusing the analysis on entrepreneurial performance, it is possible to identify three sets of indicators: the first set is relating to firms-based indicators such as employer firm birth rate, employer firm death rate; the second and third set is relating to the entrepreneurship effects in terms of employments and wealth such as high-growth firm rate by employment, gazelle rate by employment, high-growth firm rate by turnover, gazelle rate by turnover.

2.8 Employer Enterprise Birth

2.8.1 Concepts

With reference to entrepreneurship we are interested in the subpopulation of enterprises with one or more employees. While the “standard” BD on enterprise birth covers all units (without any threshold concerning very small units) the employer enterprise birth focuses on the enterprises with at least one employee. Therefore this new definition of employer enterprise births (EEB) is added to complement the enterprise birth.

By definition there are two conditions which qualify an enterprise as an employer birth: it was an enterprise birth in year t (real birth), and had at least one employee in the year of birth, or it existed before year t , was not an employer for the two previous years and had at least one employee in year t (entry by growth). Results on take-overs should be available from the methodology used to identify enterprise deaths. Where possible, the information on units that took over other units (which ceased to exist but were not deaths) should be used to identify enterprises that reached the one employee threshold by taking over another one. These should be removed from the population of births by growth.

2.8.2 The identification process

To identify the Employer Enterprise Births it is necessary to have the following sets of population:

- N_t population of active enterprises (with zero and >zero employees) in year t
 $N(1)_t$ population of active enterprises (with >zero employees) in year t
 $N(0)_{t-1}$ population of active enterprises (with zero employees) in year $t-1$
 $N(1)_{t-1}$ population of active enterprises (with >zero employees) in year $t-1$
 $N(0)_{t-2}$ population of active enterprises (with zero employees) in year $t-2$
 $N(1)_{t-2}$ population of active enterprises (with >zero employees) in year $t-2$
 $RB(1)_t$ real births (with >zero employees) in year t

A merge by identification code of the three years populations determines the following patterns:

$t-2$	$t-1$	t	Output
$N(0)_{t-2}$	$N(0)_{t-1}$	$N(1)_t$	Births by Growth
$N(0)_{t-2}$	$N(1)_{t-1}$	$N(1)_t$	-
$N(0)_{t-2}$	missing	$N(1)_t$	Births by Growth
$N(1)_{t-2}$	$N(0)_{t-1}$	$N(1)_t$	-
$N(1)_{t-2}$	$N(1)_{t-1}$	$N(1)_t$	-
$N(1)_{t-2}$	missing	$N(1)_t$	-
missing	$N(0)_{t-1}$	$N(1)_t$	Births by Growth
missing	$N(1)_{t-1}$	$N(1)_t$	-
missing	missing	$RB(1)_t$	Real Births with at least one employee

In summary, the Employer Enterprise Births in year t (EEB_t) are the Real Births with at least one employee $RB(1)_t$ and the active enterprises in year t with at least one employee ($N(1)_t$) which are in population $N(0)_{t-2}$ or $N(0)_{t-1}$ or both and which are neither in population $N(1)_{t-2}$ nor in $N(1)_{t-1}$ (Births by Growth). In order to remove from births by growth some active units that grow because events of takeover the following links for t have been identified: a) Employer Enterprise Births (EEB_t) linked to $Exits_{t-1}$ that cease for events; b) Employer Enterprise Births (EEB_t) linked to Active units ($A_{t-1,t}$) that shrink for events; c) Employer Enterprise Births (EEB_t) linked by continuity rules to $Exits_{t-1}$.

2.9 Employer Enterprise Death

2.9.1 Concepts

Like employer enterprise births there are two conditions which qualify an enterprise as an employer death (EED): it was an enterprise death in year t (real death), and had at least one employee in the year of death, or it had at least one employee in year t and continued to exist in years $t+1$ and $t+2$ without employees (death by decline). Results on split-offs should be available from the methodology used to identify enterprise births. Where possible, the information on new enterprises that were split-offs should be used to identify original enterprises that moved below the one employee threshold because a new unit emerged from a split-off. These original enterprises should be removed from the population of deaths by decline.

2.9.2 The identification process

To identify the Employer Enterprise Deaths in a reference year t , it is necessary to have the following sets of population:

- N_t population of active enterprises (with zero and >zero employees) in year t
- $N(1)_t$ population of active enterprises (with >zero employees) in year t
- $N(0)_{t+1}$ population of active enterprises (with zero employees) in year $t+1$
- $N(1)_{t+1}$ population of active enterprises (with >zero employees) in year $t+1$
- $N(0)_{t+2}$ population of active enterprises (with zero employees) in year $t+2$
- $N(1)_{t+2}$ population of active enterprises (with >zero employees) in year $t+2$

A merge by identification code of the three years populations determines the following patterns:

t	$t+1$	$t+2$	Output
$N(1)_t$	$N(0)_{t+1}$	$N(0)_{t+2}$	Deaths by Decline
$N(1)_t$	$N(1)_{t+1}$	$N(0)_{t+2}$	-
$N(1)_t$	$N(0)_{t+1}$	$N(1)_{t+2}$	-
$N(1)_t$	$N(0)_{t+1}$	missing	Deaths by Decline
$RD(1)_t$	missing	missing	Real Deaths with at least one employee

In summary, the Employer Enterprise Deaths in year t (EED_t) are the Real Deaths with at least one employee $RD(1)_t$ and the active enterprises in year t with at least one employee ($N(1)_t$) which are in population $N(0)_{t+1}$ or $N(0)_{t+2}$ or both and which are neither in population $N(1)_{t+1}$ nor in $N(1)_{t+2}$ (Deaths by Decline). Results on split-offs should be available from the methodology used to identify enterprise births. Where possible, the information on new enterprises that were split-offs should be

used to identify original enterprises that moved below the one employee threshold because a new unit emerged from a split-off. These original enterprises should be removed from the population of Deaths by decline.

2.10 Employer Surviving Enterprise Definition

An employer enterprise born in year $(t-1)$ or having survived to year $(t-1)$ from a previous year is considered to have survived in year t if it is active in terms of turnover and/or employment in any part of year t and if it has at least one employee in year t (= survival without changes).

If the enterprise is not active in year t it has survived if its activity is taken over by a new enterprise (with at least one employee) set up specifically to take over the factors of production of that enterprise in year t (= survival by take-over).

2.11 High Growth Enterprise Definition

A variety of approaches can be considered as proving the basis for defining high-growth enterprises. According to Eurostat-OECD Manual on Business Demography, the enterprise's growth should be measured both in terms of employment (number of employees) and in terms of turnover.

2.11.1 Concepts

The definition of high-growth enterprises recommended is as follows:

All enterprises with average annualised growth in employees (or in turnover) greater than 20% per annum, over three year period and with 10 or more employees at the beginning of the observation period should be considered as high-growth enterprises.

Because the percentage of growth can be too high and to avoid excluding too many enterprises, another definition of Medium-Growth enterprises is proposed:

All enterprises with average annualised growth in employees (or in turnover) between 10% and 20% per annum, over three year period and with 10 or more employees at the beginning of the observation period should be considered as medium-growth enterprises.

From the High-Growth enterprises (HG) and the Medium-Growth enterprises (MG) are excluded the enterprises whose growth was due to demographic events such as mergers, take-overs and break-ups.

2.11.2 The identification process

When trying to identify high-growth enterprises, previously it is necessary to define the potential population of high-growth as all enterprises that were active in three consecutive years, excluding the enterprises born at the beginning of the observation period. It is necessary because the newly born enterprises with at least one employees in the year $(t-3)$ could be born at different periods in time during the year $(t-3)$. Consequently their average turnover in the birth year is significantly lower than in following years simply because of the shorter average period of activity in the birth year. The same problem would not occur if only employment are measured, because it is measured as an annual average over the operating period and does not accumulate over the year. But, because high-growth enterprises are always identified from the same population, the real births in the year $(t-3)$ are removed both to identify the high-growth measured in terms of employment and the high-growth measured in terms of turnover.

In practice:

Let N_t be the population of active enterprises in year t :

Step1: a merge by identification code of the population N_t and N_{t-1} to N_{t-3}

Step2: we exclude the real births in $(t-3)$ from the Potential High-growth:

$$Potential_HG_t = (N_{t-3} \cap N_{t-2} \cap N_{t-1} \cap N_t) \setminus RB_{t-3}$$

Step3: Size threshold of 10 or more employees at the beginning of the period $(t-3)$.

$$Employees_{t-3} \geq 10$$

To identify the HG_t (or MG_t) enterprises only the enterprises with 10 or more employees are taken into account. This threshold of 10 employees is a convention and it is applied to avoid the introduction of biases that overstress the importance of small enterprises.

Step4: Growth threshold : 20% per annum for the HG

Growth threshold 10%-20% per annum for the MG

For example the HG are obtained applying to the population of reference (step 3) the following rules to employees or to turnover:

$$\sqrt[3]{\frac{employees_t}{employees_{t-3}}} - 1 \geq 0.2$$

or

$$\sqrt[3]{\frac{turnover_t}{turnover_{t-3}}} - 1 \geq 0.2$$

Step5: From HG_t (or MG_t) should be exclude the enterprises that grow because events of mergers or takeovers from units that cease or from units that transfer activity. By excluding such events we obtain the “pure” High Growth or “pure” Medium Growth enterprises.

2.12 Gazelle Enterprise Definition

2.12.1 Concepts

Gazelles are the subset of high-growth enterprises that are born (real births) at most five years ago.

We define Gazelles as “all enterprises up to 5 years old with average annualised growth greater than 20 percent per annum, over a three year period.

2.12.2 The identification process

In a given reference year t , gazelles may be in different cohorts of newly born enterprises RB_{t-3} , RB_{t-4} and RB_{t-5} , i.e., enterprises in their third, fourth or fifth year of survival. But, to be consistent with the definition of High-Growths, survivals from population RB_{t-3} are not considered.

In practice:

Let :

HG_t the High-Growth enterprises in the year t ;

RB_{t-4} the Real Births in the year $(t-4)$

RB_{t-5} the Real Births in the year $(t-5)$

Merging by identification code the HG_t with RB_{t-4} and HG_t with RB_{t-5} , we obtain the Gazelles in year t ($Gazelles_t$).

$$Gazelles_t = (HG_t \cap RB_{t-4}) \cup (HG_t \cap RB_{t-5})$$

2.13 Entrepreneurship Indicators

Data on High-Growth (Medium-Growth) enterprises and Gazelles can be used to build indicators of entrepreneurship performance such as:

Rate of HG (MG): number of HG (MG) as percentage of total potential population of HG with at least 10 employees.

Rate of Gazelles among newly born enterprises: number of gazelles as a percentage of potential population of HG with 10 employees were born 4 or 5 years ago.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Commission Regulation (EC) No 2700/98 of 17 December 1998 concerning the definitions of characteristics for structural business statistics.

Council Regulation (EEC) No 2186/93 establishing a common framework for business registers for statistical purposes and repealing.

EUROSTAT-OECD *Manual on Business Demography Statistics*. Eurostat Methodologies and working papers, 2007 Edition.

Garofalo, G. and Viviano, C. (1999), Continuity rules: re-delineation in the Italian context. In: *13th International Roundtable on Business Survey Frames*, Paris.

- Garofalo, G. and Viviano, C. (2000), The problem of links between legal units: statistical techniques for the enterprise identification and the analysis of continuity. In: *Quaderni di Ricerca*, Istat N° 1/2000.
- Garofalo, G., Paggiaro, A., Torelli, N., and Viviano, C. (2001), A record linkage procedure for the management and the analysis of the Italian statistical business register. In: *ICES-II, Proceedings of the Second International Conference on Establishment Surveys (Survey Methods for Businesses, Farms, and Institutions, June 17-21, 2000, Buffalo, New York)*, American Statistical Association, Alexandria, Virginia, 1612–1617.
- OECD. *Entrepreneurship at a Glance 2012*. OECD Publishing <http://dx.doi.org>.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Registers and Frames – The Statistical Units and the Business Register
2. Micro-Fusion – Data Fusion at Micro Level
3. Micro-Fusion – Object Matching (Record Linkage)

9. Methods explicitly referred to in this module

- 1.

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

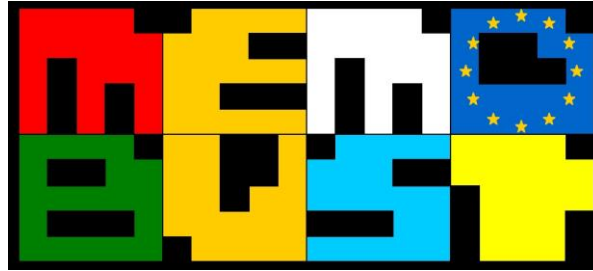
Dynamics of the Business Population-T-Business Demography

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-02-2013	first version	Patrizia Cella	Istat
0.2	29-05-2013	revision	Patrizia Cella	Istat
0.3	06-08-2013	revision	Patrizia Cella	Istat
0.4	01-10-2013	revision	Patrizia Cella	Istat
0.4.1	17-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:41



This glossary is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Memobust Glossary

Edited by Rob van de Laar, Statistics Netherlands

The glossary contains a list of words and concepts with a description of their meaning. It also contains several terms outside the SDMX and GSBPM standards. The work on the glossary started at the beginning of the Memobust project (2011), and as the project progressed, new words and concepts were added. In the final stage of the project (2014) a lot of work has been done on the glossary to integrate and harmonise statistical terms and definitions originating from different modules and different topics in the Handbook. Definitions that are not part of a standard glossary and were formulated by the authors of the Memobust modules are the so-called *Memobust definitions*. For these the “ISO/IEC 11179-4 Part 4: Formulation of data definitions” standard was applied. These concepts were used by the authors in the indicated modules. Intentionally different definitions for the same term from different standards have been kept as separate definitions in the glossary, so that in future work differences in standard definitions can be removed or definitions can be combined to arrive at even better definitions. In some cases where two or more definitions for the same term exist with one definition from a standard source and another definition a Memobust definition, the author was not able to use the standard definition and provided a definition for the purpose of his module. Accidental differences in Memobust definitions have been removed by the editor of the Memobust glossary. This work of harmonising and integrating terms and definitions could be continued after the end of the Memobust project if resources exist for this task. At the moment we integrated and harmonised the Memobust glossary for 764 definitions and 695 terms. Some of these definitions are intended to be different, so homonyms have not been prevented completely. They are indicated with a light background colour. Web addresses of sources for standard definitions are provided at the end of the document.

This Memobust glossary was used during the writing of the Handbook in order to facilitate the use of harmonised vocabulary right from the start. From the beginning this glossary was based on the SDMX glossary, and contains all concepts relevant to the Memobust handbook. For internal reviews this glossary was used as it helped reviewers to check the specific vocabulary of a module. It is intended for readers of the modules in the Memobust handbook as an easy reference, but it can also be used to find quickly modulus of the Handbook with relevant information from key terms. For each term references are provided to the relevant modules. Definitions are not repeated as part of the modules, so maintenance of the glossary is limited to this ‘global’ Memobust glossary.

Term	Definition	Source of definition	Synonyms (optional)	Module
(n,k) rule	A cell is regarded as confidential, if the n largest units contribute more than k % to the cell total, e.g. n=2 and k=85 means that a cell is defined as risky if the two largest units contribute more than 85 % to the cell total. The n and k are given by the statistical authority. In some NSIs the values of n and k are confidential.	Glossary on Statistical Disclosure Control (2014)	Dominance rule	Theme: Statistical disclosure control methods for quantitative tables
(p,q) rule	It is assumed that prior to publication of tabular data the contribution of one individual to a cell total can be estimated to within q per cent (a priori relative error in estimating the individual contribution). If after publication of the statistic the value can be estimated to within p per cent (a posteriori relative error in estimating the individual contribution), the cell is declared as confidential. The parameters p and q are determined by the statistical authority. In some NSIs the values of p and q are confidential.	Glossary on Statistical Disclosure Control (2014)	Ambiguity rule; prior posterior rule	Theme: Statistical disclosure control methods for quantitative tables
μ-ARGUS	Software that creates safe micro-data files.	Argus (2013)		Theme: Logging
Acceptance region	A component of an edit rule that defines, for a given edit group, for which values of the test variable the edit is satisfied.	Norberg (2011)		Method: Manual Editing
Accepted burden	An allowable level of response burden created e.g. by increasing nonresponse rates, which has a positive effect on response burden. To avoid such undesirable "rewards" and, consequently, a less alert attitude towards declining response rates, survey managers should be confronted with burden figures which include hypothetical non response burden as well	Willeboordse <i>et al.</i> (1997)		Theme: Response Burden
Accessibility	The ease and conditions under which statistical information can be obtained.	Eurostat's Concepts and Definitions Database (2013)		(1) Theme: Quality of Statistics; (2) Theme: Overall Design
Accessibility of a log	The ease and conditions under which logs can be obtained.	Memobust definition (2014)		Theme: Logging
Accuracy	The closeness of estimates to the unknown true values.	ESS Handbook for Quality Reports (2009) (2009)		(1) Theme: Overall Design; (2) Theme: Repeated Surveys
Accuracy	Closeness between the estimated value and the true value measured by the statistic (usually unknown)	OECD (2006)		Theme: Revisions of Economic Official Statistics
Accuracy	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) Method: RAS; (3) Method: Stone; (4) Theme: Macro Integration.

Accuracy (of estimates)	The closeness of estimates to the true values.	ESS Handbook for Quality Reports (2009)		Theme: Quality of Statistics
Accuracy (of estimates)	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure. Context: The accuracy of statistical information is the degree to which the information correctly describes the phenomena. It is usually characterized in terms of error in statistical estimates and is often decomposed into bias (systematic error) and variance (random error) components. Accuracy is associated with the "reliability" of the data, which is defined as the closeness of the initial estimated value to the subsequent estimated value.	SDMX (2009)		(1) Theme: Methods and Quality; (2) Theme: Quality and Risk Management Models
Active enterprise	Within the Business Demography context, activity is defined as any turnover and/or employment in the period from 1st January to 31st December in a given year.	Eurostat-OECD Manual on Business Demography Statistics (chapter 6)		Theme: Business Demography
Activity	An activity can be said to take place when resources such as equipment, labour, manufacturing techniques, information networks or products are combined, leading to the creation of specific goods or services. An activity is characterised by an input of products (goods and services), a production process and an output of products. Activities can be determined by reference to a specific level of NACE Rev. 2.	CODED		Theme: Derivation of Statistical Units
Activity	The combination of actions that result in a certain set of products. An activity can be said to take place when resources such as equipment, labour, manufacturing techniques or products are combined, leading to specific goods or services. Thus, an activity is characterised by an input of resources, a production process and an output of products. Context: In practice the majority of units carry on activities of a mixed character. One can distinguish between three types of economic activity: - Principal activity: The principal activity is identified by the top-down method as the activity which contributes most to the total value added of the entity under consideration. The principal activity so identified does not necessarily account for 50% or more of the entity's total value added. - Secondary activity: A secondary activity is any other activity of the entity that produces goods or services.	RAMON, Eurostat's metadata server		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys

Activity	An activity can be said to take place when resources such as equipment, labour, manufacturing techniques, information networks or products are combined, leading to the creation of specific goods or services. An activity is characterised by an input of products (goods and services), a production process and an output of products. Activities can be determined by reference to a specific level of NACE Rev. 2. If a unit carries out more than one activity, all the activities, which are not ancillary activities are ranked according to the gross value added. On the basis of the preponderant gross value added generated, a distinction can then be made between principal activity and secondary activities. Ancillary activities are not isolated to form distinct entities or separated from the principal or secondary activities of entities they serve.	RAMON, Eurostat's metadata server		Theme: Statistical Registers and Frames – The statistical units and the business register
Actual burden	The burden based on a realistic level of difference between signals on non-response and response. More precisely, it is a reasonably allowable level of non-response.	Hedlin <i>et al.</i> (2005)		Theme: Response Burden
Adjacency matrix	0-1 matrix that indicates which nodes in a graph (or a digraph) are connected by an edge (or an arrow).	Hacking & Willenborg (2012)		Method: Automatic coding based on semantic networks
Administrative data	The data derived from an administrative source, before any processing or validation by the NSIs.	Essnet Admin Data Glossary 1.1		(1) Theme: Collection and Use of Secondary Data; (2) Theme: Editing Administrative Data; (3) Theme: Estimation with administrative data
Administrative data holder	The organisational unit holding an administrative source	Essnet Admin Data Glossary 1.1		Theme: Collection and Use of Secondary Data
Administrative data provider	The administrative data holder who is bound to provide their data to the NSI, by law or by virtue of a specific agreement	Essnet Admin Data Glossary 1.1		Theme: Collection and Use of Secondary Data
Administrative population	The set of units that an administrative source is meant to cover, as defined by the relevant administrative regulation. This population may or may not correspond exactly to a given target	Essnet Admin Data Glossary 1.1		Theme: Collection and Use of Secondary Data
Administrative register	Administrative registers come from administrative sources and become statistical registers after passing through statistical processing in order to make it fit for statistical purposes (production of register based statistics, frame creation, etc.).	UN/ECE Glossary of Terms on Statistical Data Editing (2007)		Theme: Collection and Use of Secondary Data
Administrative regulation	A set of detailed directions having force of law, developed to put a policy into practice (such as decrees, ordinances, and other similar provisions). It is normally addressed to a designated population of natural and/or juridical persons, which are bound to observe it.	Essnet Admin Data Glossary 1.1		Theme: Collection and Use of Secondary Data
Administrative source	A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations.	Essnet Admin Data Glossary 1.1 (first part) & SDMX, 2009		Theme: Collection and Use of Secondary Data

Administrative source	A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations. In a wider sense, any data source containing information that is not primarily collected for statistical purposes.	Essnet Admin Data Glossary 1.1		Theme: Editing Administrative Data
Administrative source	A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations. Context: A wider definition of administrative sources, is used in the Eurostat Business Registers Recommendations Manual: a data holding containing information which is not primarily collected for statistical purposes. The organisational unit responsible for maintaining one or more administrative sources is known as an administrative organisation.	SDMX (2009)		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (4) Theme: Statistical Registers and Frames – The statistical units and the business register
Administrative units	With reference to the use of administrative data for statistical purposes, the units for which administrative data are recorded. These units may or may not be the same as those required for the statistical output (which are referred to as statistical units).	Essnet Admin Data Glossary 1.1		Theme: Editing Administrative Data
Aggregation	Aggregation in a system of time series is commonly referred in a literature as benchmarking to contemporaneous constraints.	Stuckey et.al. (2004)		(1) Theme: Issues on Seasonal Adjustment; (2) Theme: Seasonal adjustment – introduction and general description.
AIC	Measure of the relative goodness of fit of a statistical model $AIC = 2k - 2\log(lik)$, where k is the number of parameters in the model and lik is maximum value assumed by the likelihood function.	Memobust definition (2014)		(1) Method: EBLUP Unit level for Small Area Estimation; (2) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)
Allocating (sample elements to interviewers)	The allocation consists of associating each telephone number (belonging to a sample element) with an interviewer. So the allocation of interviewers to sample elements is via their telephone numbers.	Memobust definition (2014)		Theme: CATI Allocation
Ambiguity rule	See: (p,q) rule.	Glossary on Statistical Disclosure Control (2014)	(p,q) rule; prior posterior rule	Theme: Statistical disclosure control methods for quantitative tables
Annual Alignment	The constraint that annual data has to be consistent with sub annual data. Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.	Memobust definition (2014)		Method: Denton for Benchmarking
Anticipated value	Anticipated values are used in score functions and are predictions for the values which are expected in the actual survey.	EDIMBUS Manual	Predicted values	Theme: Selective Editing

ARGUS	Two software packages for Statistical Disclosure Control are called Argus. μ -Argus is a specialized software tool for the protection of microdata. The two main techniques used for this are global recoding and local suppression. In the case of global recoding several categories of a variable are collapsed into a single one. The effect of local suppression is that one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. Both global recoding and local suppression lead to a loss of information, because either less detailed information is provided or some information is not given at all. τ -Argus is a specialized software tool for the protection of tabular data. τ -Argus is used to produce safe tables. τ -Argus uses the same two main techniques as μ -Argus: global recoding and local suppression. For τ -Argus the latter consists of suppression of cells in a table.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
ARIMA models	These are a versatile family of models for modelling and forecasting time series data. Seasonal ARIMA models have a special form for efficiently modelling many kinds of seasonal time series and are heavily used in seasonal adjustment. ARIMA is an acronym for AutoRegressive Integrated Moving Average	US Census Bureau		Method: Seasonal adjustment of economic time series
Assisted coding	Coding of textual variable performed during the interview	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools
Attribute	A quality of feature, especially one that is considered to be good or useful. Examples: availability, accuracy, integrity, confidentiality, effectiveness.	Longman (2010)		(1) Theme: Methods and Quality; (2) Theme: Quality and Risk Management Models
Attribute disclosure	Attribute disclosure is attribution independent of identification. This form of disclosure is of primary concern to NSIs involved in tabular data release and arises from the presence of empty cells either in a released table or linkable set of tables after any subtraction has taken place. Minimally, the presence of an empty cell within a table means that an intruder may infer from mere knowledge that a population unit is represented in the table and that the intruder does not possess the combination of attributes within the empty cell.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables

Attribute of register unit	Attribute of a register unit is a regularly updated characteristic of a register unit. Remark: Attributes of statistical register units can be arranged in groups. Accordingly, attributes referring to identification, contact, classification, demographic characteristics, relation to other register units, attributes supporting register maintenance and statistical processes (for example organization of data collection, sampling, etc.) can be defined. In respect of maintainability and changes of attributes over time, administrative and statistical attributes are distinguished	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys
Automatic coding	Coding (in batch) using a program. The program takes all of the decisions.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Automatic coding based on semantic networks; (3) Theme: Coding
Automatic coding	The computer assigns codes to the verbal responses working in “batch” processing	Macchia S. and Murgia M (2002)	AUC	Theme: Different Coding Strategies
Automatic coding precision	The percentage of correctly coded descriptions ((Number of correctly coded description/Number coded descriptions)	Memobust definition (2014)		Method: Automatic coding based on pre-coded datasets
Automatic coding rate	The percentage of coded descriptions(Number of coded description/Number descriptions to be coded)	Memobust definition (2014)		Method: Automatic coding based on pre-coded datasets
Automatic editing	An umbrella term for editing methods in which the data are checked and adjusted by a computer.	Memobust definition (2014)		(1) Method: Automatic Editing; (2) Method: Deductive Editing; (3) Theme: Editing Administrative Data; (4) Theme: Statistical Data Editing
Autoregressive model	A representation of a type of random process; as such, it describes certain time-varying processes. The autoregressive model specifies that the output variable depends linearly on its own previous values.	Memobust definition (2014)		Method: Chow-Lin Method for Temporal Disaggregation
Autoregressive model	An econometric model-based upon the autoregressive process but also containing lagged versions of some or all of the endogenous variables considered in the model specification.	Memobust definition (2014)		Method: Preliminary estimates with model-based methods
Auxiliary variable	A variable that correlates with the target variable and is observed for all units.	CBS Methods Series Glossary		(1) Theme: Donor Imputation; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Model-based Imputation; (5) Theme: Sample selection; (6) Theme: Design of Estimation – Some Practical Issues; (7) Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered

Bag-of-words assumption	The assumption that, for a description, only the separate words that occur play a role, and not the order and the combinations of these words in the description.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Theme: Coding
Barnardisation	A method of disclosure control for tables of counts that involves randomly adding or subtracting 1 from some cells in the table.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Base register	Registers kept as a basic resource for public administration. The function of base registers is typically to keep stock of the population at any given time. In addition, they have to maintain identification information to be used by other sources.	UN/ECE Glossary of Terms on Statistical Data Editing (2007)		Theme: Collection and Use of Secondary Data
Benchmarking	Achieving consistency between data that are published at different frequencies (for instance quarterly data that has to comply with annual data).	Memobust definition (2014)		Method: Denton for Benchmarking
Benchmarking	Achieving consistency between data that are published at different frequencies (for instance quarterly data that has to comply with annual data).	Memobust definition (2014)		Theme: Macro Integration
Benchmarking	Achieving consistency between data that are published at different level of aggregation.	SDMX (2009)		(1) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot); (2) Method: EBLUP Unit level for Small Area Estimation
Benchmarking	Benchmarking (to temporal constraints) involves enforcing consistency across time with respect to another time series.	Stuckey et.al. (2004)		(1) Theme: Issues on Seasonal Adjustment; (2) Theme: Seasonal adjustment – introduction and general description.
Bias	An effect which deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average.	Eurostat's Concepts and Definitions Database (2013)	Systematic error.	Theme: Quality of Statistics
Bias	The bias of an estimator is the difference between its mathematical expectation and the true value of the parameter. In case it is zero, the estimator is said to be unbiased. Expectation is usually calculated on the set of all possible samples (Randomization approach). Otherwise is calculated with respect to the assumed model (model-based approach).	Memobust definition (2014)		(1) Theme: Weighting and Estimation; (2) Theme: Estimation with administrative data; (3) Method: EBLUP Unit level for Small Area Estimation
Bias	The bias of an estimator is the difference between its mathematical expectation and the true value it estimates. If this difference is zero, the estimator is said to be unbiased. Expectation is usually calculated on the set of all possible samples.	SDMX (2009)		(1) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot); (2) Method: Preliminary estimates with design-based methods; (3) Method: Deductive Editing; (4) Theme: Statistical Data Editing

Bias	The bias of an estimator is the difference between its mathematical expectation and target parameter. In the case it is zero, the estimator is said to be unbiased. Expectation is usually calculated on the set of all possible samples.	Statistical Data and Metadata Exchange (SDMX)		Method: Generalised regression estimator
Bias (of an estimator)	An effect which deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average.	SDMX (2009)		(1) Theme: Sample co-ordination; (2) Method: Denton for Benchmarking; (3) Theme: Macro Integration
BIC	This is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function.	Memobust definition (2014)		Method: EBLUP Unit level for Small Area Estimation
BIC	Measure of the relative goodness of fit of a statistical model $BIC = k \log(n) - 2\log(\text{lik})$ where k is the number of parameters in the model, n is the number of observation and lik is the maximum value of the likelihood function.	Memobust definition (2014)		Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)
Binding constraint	See hard constraint	Memobust definition (2014)		Method: Denton for Benchmarking
Birth rate	The birth rate of a given reference period is the number of births as a percentage of the population of active enterprises.	Memobust definition (2014)		Theme: Business Demography
Blocking variable	A variable that is used to partition matching data sets, that is, divide in a number of subfiles, with the intention of reducing the search space.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching; (5) Method: Fellegi-Sunter and Jaro Approach to Record Linkage
BLUE (Best Linear Unbiased Estimator)	Estimator minimizing the square loss in the class of linear unbiased estimators (unbiasedness is referred to the model distribution).	Memobust definition (2014)		Method: Small area estimation methods for time series data
BLUP (Best Linear Unbiased Predictor)	Predictor which minimizes the square loss in the class of linear unbiased predictors (unbiasedness is referred to the model distribution).	Memobust definition (2014)		Method: Small area estimation methods for time series data
Bounds	The range of possible values of a cell in a table of frequency counts where the cell value has been perturbed or suppressed. Where only margins of tables are released it is possible to infer bounds for the unreleased joint distribution. One method for inferring the bounds across a table is known as the Shuttle algorithm.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Break of time series	Break occurring when there is a change in the standards for defining and observing a variable over time.	SDMX (2009)	Time series break	Theme: Repeated Surveys

Break-up	This event involves a splitting of the production factors of an enterprise into two or more new enterprises, in such a way that the previous enterprise is no longer recognisable. There is no continuity or survival, but the closure of the previous enterprise is not considered to be a death. Similarly the new enterprise are not considered to be births.	Eurostat-OECD Manual on Business Demography Statistics (chapter 4).		Theme: Business Demography
BSDG	Bussiness Statistics Directors Group	Eurostat website/CROS portal		Theme: The European Statistical System
Business register for statistical purposes	Regulation (EC) No 177/2008 of the European Parliament and of the Council establishes a common framework for business registers for statistical purposes in the Community. Member States shall set up one or more harmonised registers for statistical purposes, as a tool for the preparation and coordination of surveys, as a source of information for the statistical analysis of the business population and its demography, for the use of administrative data, and for the identification and construction of statistical units. The registers shall be compiled of: All enterprises carrying on economic activities contributing to the gross domestic product (GDP), and their local units; The legal units of which those enterprises consist; Truncated enterprise groups and multinational enterprise groups; and All-resident enterprise groups.	Business Register Regulation (EC) No 177/2008, Articles 1 and 3 (1)	Statistical business register	Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames
CAI	Computer Assisted Interviewing. The use of computer during interviewing.	Economic Commission for Europe of the United Nations (UNECE), "Glossary of Terms on Statistical Data Editing", Conference of European Statisticians Methodological material, Geneva (2000)		(1) Theme: Electronic Questionnaire Design; (2) Theme: Editing During Data Collection; (3) Theme: Testing the Questionnaire; (4) Theme: Questionnaire Design; (5) Theme: Data Collection: Techniques and Tools; (6) Theme: CATI Allocation
cAIC	As model selection measure, cAIC is well -suited for small area estimation. It is relevant to inferences regarding the clusters, or areas, in the context of linear mixed models. inferences regarding the clusters, or areas, in the context of linear mixed models. The criterion is based on the conditional likelihood for fixed and random effects vectors evaluated at their estimated values, and y is the data. The effective number of degrees of freedom is essentially given by the trace of the hat matrix H	Memobust definition (2014)		Method: EBLUP Unit level for Small Area Estimation

cAIC	As model selection measure, cAIC is well-suited for small area estimation. It is relevant for inferences regarding clusters, or domains, in the context of linear mixed models. The criterion is $cAIC = 2peff - 2\log(lik)$, where lik is maximum values assumed by the conditional likelihood, that is the likelihood function when fixed and random effects vectors evaluated at their estimated values. The effective number of degrees of freedom $peff$ is essentially given by the trace of the hat matrix H .	Memobust definition (2014)		Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)
CAII	Computer Assisted Internet Interview	Willeboordse <i>et al.</i> (1997)		Theme: Response Burden
Calculated interval	The interval containing possible values for a suppressed cell in a table, given the table structure and the values published.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Calendar adjustment	Calendar adjustment refers to the correction for calendar variations. Such calendar adjustments include working day adjustments or the incidence of moving holidays (such as Easter and Chinese New Year)	OECD (2006)		Method: Seasonal adjustment of economic time series
Calendar effects	Influences deriving from differences in the number of working days or the dates of particular days which can be statistically proven and quantified	Eurostat (2009)		(1) Method: Seasonal adjustment of economic time series; (2) Theme: Seasonal adjustment – introduction and general description
Calibration	One of the most important methods of weighting commonly used by many statistical agencies in survey sampling, whose main aim is to compute weights to be used in estimation, given an input of auxiliary information.	Memobust definition (2014)		Method: Calibration
Calibration equation	In the calibration procedure for totals, equations in which calibration weights applied to all auxiliary variables in the sample exactly reproduce the known population totals of the auxiliary variables.	Memobust definition (2014)		Method: Calibration
Calibration estimator	An estimator which is a weighted sum of sample observation, whose weights are obtained in order to minimize a distance with the design weights subject to the constraint that the weighted sum of an auxiliary variables reproduce the known amount. See Module XIX 2.c for further details	Memobust definition (2014)		Method: Generalised regression estimator
Calibration estimator	Estimator which takes into account calibration weights which satisfy calibration equations.	Memobust definition (2014)		Method: Calibration
Calibration weights	Weights which replace the original initial design weights and satisfy calibration equations.	Memobust definition (2014)		Method: Calibration
Call scheduler	Software that runs in the scheduling system according to the values of the scheduling parameters set by survey responsible	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools

Capacity of call room	The maximum number of interviewers that can work simultaneously in the call room for CATI survey work.	Memobust definition (2014)		Theme: CATI Allocation
CAP	Computer Assisted Personal Interviewing. A method of data collection in which an interviewer uses a computer to display questions and accept responses during a face-to-face interview.	United States Bureau of Census, Glossary of Selected Abbreviations and Acronyms.		(1) Theme: Data Collection; (2) Theme: CATI Allocation; (3) Theme: Data Collection: Techniques and Tools; (4) Theme: Response Burden; (5) Method: Computer-assisted coding; (6) Theme: Questionnaire Design; (7) Theme: Mixed Mode Data Collection – design issues; (8) Theme: Electronic Questionnaire Design; (9) Theme: Editing During Data Collection; (10) Theme: Coding; (11) Theme: Quality of Statistics
CASI	Computer Assisted Self-Interviewing. The technique whereby respondents independently complete electronic questionnaires, assisted only by specially-designed computer programs.	Glossary, Adapting new technologies to census operations (2001)		(1) Theme: Electronic Questionnaire Design; (2) Theme: Editing During Data Collection; (3) Theme: Testing the Questionnaire; (4) Theme: Questionnaire Design
CASI	Computer Assisted Self Interviewing is a method of data collection in which the respondent operates the computer: questions are read from the computer screen and responses are entered directly in the computer. A well-known form of CASI is the web survey.	Memobust definition (2014)		Theme: Mixed Mode Data Collection – design issues
CATI	Computer Assisted Telephone Interviewing. A method of data collection by telephone with questions displayed on a computer and responses entered directly into a computer.	United States Bureau of Census, Glossary of Selected Abbreviations and Acronyms.		(1) Theme: Questionnaire Design; (2) Theme: Mixed Mode Data Collection – design issues; (3) Theme: Data Collection; (4) Theme: CATI Allocation; (5) Theme: Data Collection: Techniques and Tools; (6) Theme: Response Burden; (7) Method: Computer-assisted coding; (8) Theme: Electronic Questionnaire Design; (9) Theme: Editing During Data Collection; (10) Theme: Testing the Questionnaire; (11) Theme: Coding; (12) Theme: Quality of Statistics
CATI Interviewer	A person who on behalf of a statistical office carries out interviews by telephone. In this module we assume that these people work from a call room.	Memobust definition (2014)		Theme: CATI Allocation

CAWI	Computer Assisted Web Interviewing. A method of data collection based on web questionnaire. The respondent accesses the questionnaire via a web connection and fills it in.	Memobust definition (2014)	Web Survey	(1) Theme: Coding; (2) Theme: Quality of Statistics; (3) Theme: Data Collection: Techniques and Tools; (4) Method: Computer-assisted coding; (5) Theme: Mixed Mode Data Collection – design issues
CBA	Cost Benefit Analysis – a model enabling identification of which cost and benefits to include to evaluate effects of participating in the survey, discounting future benefits and costs over time to obtain a present day value and identification of relevant constraints.	Haraldsen <i>et al.</i> (2013), Pres and Turvey (1965)		Theme: Response Burden
Cell suppression	In tabular data the cell suppression SDC method consists of primary and complementary (secondary) suppression. Primary suppression can be characterised as withholding the values of all risky cells from publication, which means that their value is not shown in the table but replaced by a symbol such as 'x' to indicate the suppression. According to the definition of risky cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or presenting a case of dominance have to be primary suppressed. To reach the desired protection for risky cells, it is necessary to suppress additional non- risky cells, which is called complementary (secondary) suppression. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the risky cells with the least amount of suppressed information.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Chain-linking	Joining together two indices that overlap in one period by rescaling one of them to make its value equal to that of the other in the same period, thus combining them into single time series. More complex methods may be used to link together indices that overlap by more than period	OECD (2006)		Theme: Issues on Seasonal Adjustment
Changes in inventories	Changes in inventories are measured by the value of the entries into inventories less the value of withdrawals and the value of any recurrent losses of goods held in inventories.	ESA (2010)		Theme: Manual Integration
Characteristic	See: Attribute.	Memobust definition (2014)		Theme: Methods and Quality
Checking rule	see Edit	Memobust definition (2014)	Edit	Theme: Statistical Data Editing
Clarity	The ease with which users can understand the statistics.	ESS Handbook for Quality Reports (2009)		Theme: Overall Design

Clarity	The extent to which easily comprehensible metadata are available (for the user), where these metadata are necessary to give a full understanding of statistical data.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Clarity of log information	The degree to which the log information can be read, understood and interpreted.	Memobust definition (2014)	Readability, interpretability	Theme: Logging
Classification scheme	A hierarchical arrangement of kinds of things (classes) or groups of kinds of thing	Wikipedia, English edition		(1) Method: Manual coding; (2) Theme: Coding
Cluster sampling	A sampling technique used when 'natural' groupings are evident in a statistical population	Wikipedia Cluster Sampling		Theme: Sample selection
Coder	A specialist trained to interpret and classify descriptions (in a certain area) in the light of a classification used for that purpose.	Hacking & Willenborg (2012)		(1) Method: Manual coding; (2) Method: Automatic coding based on pre-coded datasets; (3) Method: Computer-assisted coding; (3) Theme: Coding
Coding	The activity in the statistical process in which it is determined whether a code from a classification can be assigned to a description, and, if so, which code this could be.	Hacking & Willenborg (2012)		(1) Method: Manual coding; (2) Method: Automatic coding based on pre-coded datasets; (3) Method: Computer-assisted coding; (4) Method: Automatic coding based on semantic networks; (5) Theme: Coding
Coding	The process of converting verbal or textual information into codes representing classes within a classification scheme, to facilitate data processing, storage or dissemination	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools
Coding	The process of converting verbal or textual information into codes representing classes within a classification scheme, to facilitate data processing, storage or dissemination.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Coding error	The assignment of an incorrect code to a data item.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Coding precision	The percentage of <u>correctly</u> coded descriptions ((Number of correctly coded description/Number coded descriptions)	Memobust definition (2014)	Automatic coding precision	Theme: Coding
Coding rate	Percentage of coded texts on the total of texts to be coded	D'Orazio M. and Macchia S (ROS) (2002)	Efficacy, Automatic coding rate	(1) Theme: Coding; (2) Theme: Measuring Coding Quality
Coefficient of variation	The ratio of the square root of the variance of the estimator to its expected value.	ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys		(1) Method: Generalised regression estimator; (2) Theme: Quality of Statistics
Coherence	The degree to which the statistical processes by which statistics were generated used the same concepts – classifications, definitions and target populations – and harmonised methods.	ESS Handbook for Quality Reports (2009)		Theme: Quality of Statistics

Coherence	Adequacy of statistics to be reliably combined in different ways and for various uses.	ESS Handbook for Quality Reports (2009) (2009)		(1) Theme: Overall Design; (2) Theme: Repeated Surveys
Coherence	Adequacy of statistics to be combined in different ways and for various uses.	SDMX (2009)		Theme: Weighting and Estimation
Coherence	Adequacy of statistics to be combined in different ways and for various uses.	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) Method: RAS; (3) Method: Stone; (4) Theme: Macro Integration.
Cold deck imputation	A form of donor imputation in which the donor record comes from a different data set than the recipient record.	Memobust definition (2014)		Theme: Donor Imputation
Collection unit	Collection unit is the unit from which data are obtained and by which questionnaire survey forms are completed. Data supplier and data provider are collection units.	United Nations, DEPARTMENT of Economic and Social Affairs, Statistics Division [2007]: Statistical Units. United Nations, New York		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design.
Commodity	Goods and services produced and used in an economy.	Memobust definition (2014)		Theme: Manual Integration
Communication mode	A channel used in a survey to contact businesses, seek survey cooperation, communicate information, instructions, procedures and non-response follow-up, and to support businesses.	Snijkers & Jones (2013)		Theme: Mixed Mode Data Collection – design issues
Communication strategy	How businesses are contacted and followed-up in case of non-response, aimed at receiving timely, accurate and complete responses.	Memobust definition (2014)		Theme: Mixed Mode Data Collection – design issues
Comparability	The degree to which the same data items can be compared but for different reference periods or different sub populations (regions or domains).	ESS Handbook for Quality Reports (2009)		Theme: Quality of Statistics
Comparability	Adequacy of statistics to be reliably compared; measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared.	ESS Handbook for Quality Reports, 2009; modified and expanded.		(1) Theme: Overall Design; (2) Theme: Repeated Surveys.
Comparison functions	Functions that compute the distance between records compared on the chosen matching variables.	Memobust definition (2014)		Theme: Probabilistic Record Linkage
Complementary suppression	See: Secondary suppression	Glossary on Statistical Disclosure Control (2014)	Secondary suppression	Theme: Statistical disclosure control methods for quantitative tables
Completeness of log information	The degree to which log information meets all current and potential needs of the user of the log information.	Memobust definition (2014)		Theme: Logging

Composite estimator	A weighted sum of two component estimators defined to reduce the mean-squared-error (MSE) of the resulting estimator.	Memobust definition (2014)		(1) Method: Preliminary estimates with design-based methods; (2) Method: Composite Estimators for Small Area Estimation
Composite unit	A unit that is composed of units from a lower order. A household is an example of a composite unit; 'persons' are the simple units from which 'households' are composed.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Computer assisted coding (CAC)	The operator assigns codes working interactively with the computer, that gives him a support in "navigating" inside the dictionary to search for codes to be assigned to the input descriptions.	Macchia S. and Murgia M (2002)	Interactive coding	Theme: Different Coding Strategies
Computer assisted survey information collection	Computer assisted survey information collection (CASIC) encompasses computer assisted data collection and data capture. CASIC may be more broadly defined to include the use of computer assisted, automated, or advanced computing methods for data editing and imputation, data analysis and tabulation, data dissemination, or other steps in the survey or census process.	UN Statistical Commission, UNECE, 2000. Glossary of Terms on Statistical Data Editing.		Theme: Testing the Questionnaire
Computer supported coding	See Computer-assisted coding	Memobust definition (2014)		Method: Computer-assisted coding
Computer-assisted coding	A form of coding in which a coder makes all the coding decisions, possibly while using an electronic file or index.	Hacking & Willenborg (2012)	Interactive coding	(1) Method: Computer-assisted coding; (2) Theme: Coding
Concentration rule	Rule to assess whether a cell is a risky cell, based on comparing the size of the individual contributions to the cell. Examples are the dominance rule and the p% rule.	Hundepool et al. (2012)	Dominance rule; P% rule	Theme: Statistical disclosure control methods for quantitative tables
Conditional mean matching	Model based imputation method: imputes the missing value with its expectation given the observed variables	Memobust definition (2014)		Method: Statistical Matching Methods
Confidentiality of log information	The degree to which log information cannot be made available to users of log information.	Memobust definition (2014)		Theme: Logging
Connected component	A maximal connected subgraph of a graph.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Consistency	Sum of sub-annual values of a time series is equal to the annual values; in case of aggregation, the total values are equal to the aggregated values	Dagum and Cholette (2006)		Theme: Issues on Seasonal Adjustment

Consistency	Logical and numerical coherence.	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Method: RAS; (4) Method: Stone; (5) Theme: Macro Integration; (6) Theme: Data fusion at micro level; (7) Theme: Quality of Statistics; (8) Theme: Weighting and Estimation
Consistency	Data values are said to be consistent if they conform to specified edit rules.	SDMX (2009)		(1) Method: Generalised Ratio Adjustments; (2) Method: Minimum Adjustment Methods; (3) Method: Prorating; (4) Method: Reconciling Conflicting Micro-Data
Consistency	An estimator is called consistent if it converges in probability to its estimand as sample increases	The International Statistical Institute, "The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press (2003).		Theme: Small area estimation
Constrained distance hot deck	The donor can be chosen just once and the subset of the donors is selected in order to minimize the overall matching distance.	Memobust definition (2014)		Method: Statistical Matching Methods
Constraint	Specification of what may be contained in a data or metadata set in terms of content or, for data only, in terms of the set of key combinations to which specific attributes (defined by the data structure) may be attached.	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Method: RAS; (4) Method: Stone; (5) Theme: Macro Integration
Consumption of government	Final consumption expenditure consists of expenditure incurred by resident institutional units on goods or services that are used for the direct satisfaction of the collective needs of members of the community.	ESA (2010)		Theme: Manual Integration
Consumption of households	Final consumption expenditure consists of expenditure incurred by resident institutional units on goods or services that are used for the direct satisfaction of individual needs or wants.	ESA (2010)		Theme: Manual Integration
Contact strategy	when and how respondents are contacted, and what material (questionnaire, cover letter, instructions et cetera) is used in each contact.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 2) – Contact strategies
Contempeous constraints	Constraints within one period, between different time-series	Memobust definition (2014)		Method: Denton for Benchmarking

Control	Control is the ability to determine general corporate policy by choosing appropriate directors. Control is when owning more than half of the voting shares or otherwise controlling half of the shareholders' voting power (e.g. by controlling the shareholder or by a contract of control). This type of control can be registered as it has a legal basis. Control can be direct but can also be indirect .	European System of Accounts (ESA 1995), paragraph 2.26		Theme: Derivation of Statistical Units
Controlled rounding	Controlled rounding: To solve the additivity problem, a procedure called controlled rounding was developed. It is a form of random rounding, but it is constrained to have the sum of the published entries in each row and column equal to the appropriate published marginal totals. Linear programming methods are used to identify a controlled rounding pattern for a table.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Controlled Tabular Adjustment	A method to protect tabular data based on the selective adjustment of cell values. Sensitive cell values are replaced by either of their closest safe values and small adjustments are made to other cells to restore the table additivity. Controlled tabular adjustment has been developed as an alternative to cell suppression.	Glossary on Statistical Disclosure Control (2014)	CTA	Theme: Statistical disclosure control methods for quantitative tables
Conventional Rounding	A disclosure control method for tables of counts. When using conventional rounding, each count is rounded to the nearest multiple of a fixed base. For example, using a base of 5, counts ending in 1 or 2 are rounded down and replaced by counts ending in 0 and counts ending in 3 or 4 are rounded up and replaced by counts ending in 5. Counts ending between 6 and 9 are treated similarly. Counts with a last digit of 0 or 5 are kept unchanged. When rounding to base 10, a count ending in 5 may always be rounded up, or it may be rounded up or down based on a rounding convention.	Glossary on Statistical Disclosure Control (2014)	Deterministic rounding	Theme: Statistical disclosure control methods for quantitative tables
Co-ordination of samples	Increasing the sample overlap for some surveys rather than drawing the samples independently is known as positive co-ordination. Reducing the overlap between samples for different surveys is known as negative co-ordination.	SDMX (2009)		(1) Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered; (2) Method: Sample co-ordination using simple random sampling with permanent random numbers; (3) Theme: Sample co-ordination; (4) Theme: Design of Estimation – Some Practical Issues; (5) Theme: Sample selection

CoP	The European Code of Practice provides 15 principles covering the institutional environment, the statistical production processes and the output of statistics. A set of indicators of good practice for each of the principles provides a reference for reviewing the implementation of the Code.	European Code of Practice (2011)		(1) Theme: Specification of User Needs for Business Statistics; (2) Theme: Dissemination of Business Statistics; (3) Theme: Methods and Quality
Corpus	Coded set of descriptions.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Computer-assisted coding; (3) Theme: Coding
Correction rule	An if-then rule that is used to treat a particular error in a deterministic manner.	loosely based on UN/ECE Glossary of Terms on Statistical Data Editing		Method: Deductive Editing
Correctness of log information	The degree to which log information reflects reality.	Memobust definition (2014)		Theme: Logging
Covariance matrix	A mathematic measure of reliability.	Memobust definition (2014)		Method: Stone
Coverage	The definition of the population that statistics aim to cover.	SDMX (2009)		Theme: Sample selection
Coverage error	Error caused by a failure to cover adequately all components of the population being studied, which results in differences between the target population and the sampling frame.	Eurostat's Concepts and Definitions Database, SDMX Metadata Common Vocabulary (http://sdmx.org), 2009		(1) Method: Denton for Benchmarking; (2) Theme: Quality of Statistics
Creative editing	A process whereby manual editors invent editing procedures to avoid reviewing another error message from subsequent machine editing.	UN/ECE Glossary of Terms on Statistical Data Editing (2007)		Method: Manual Editing
Cross validation	CV methods allow to test the robustness of the models, quantifying their predictive power by leaving out one or more observations when fitting the models, and subsequently assessing the model predictions for the left-out observation(s). It can be quantified in alternative ways, for instance averaging the prediction errors	Memobust definition (2014)		(1) Method: EBLUP Unit level for Small Area Estimation; (2) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)
Cross-section	This involves some observations of all population, or a representative subset at one specific point in time.	Memobust definition (2014)		Method: Little and Su Method
CTA	See: Controlled Tabular Adjustment	Memobust definition (2014)	Controlled Tabular Adjustment	Theme: Statistical disclosure control methods for quantitative tables
Cut-off sampling	A sampling procedure in which a predetermined threshold is established with all units in the universe at or above the threshold being included in the sample and all units below the threshold being excluded. The threshold is usually specified in terms of the size of some known relevant variable. In the case of establishments, size is usually defined in terms of employment or output.	Memobust definition (2014)		Method: Subsampling for Preliminary Estimates

Cut-off survey	A survey in which all the entities falling above or below a threshold determined according to one or more characteristics of those entities are either included or excluded	SDMX (2009)		Theme: Sample selection
Cut-off threshold	A threshold used, mainly for cost or burden reasons, to exclude from the target population (hence from the frame) units contributing very little to the requested statistics, small businesses for instance.	SDMX (2009)		Theme: Sample selection
Cut-off value	A value to limit the matching weights (upwards or downwards).	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Damerau-Levenshtein distance	A metric defined to measure the distance between strings. It measures the minimum number of elementary steps to transform one string into another.	Memobust definition (2014)	Levenshtein distance	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Data	Characteristics or information, usually numerical, that are collected through observation.	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Method: RAS; (4) Method: Stone; (5) Theme: Macro Integration
Data cleaning	see Editing	Memobust definition (2014)	Editing	Theme: Statistical Data Editing
Data collection mode	The technical set-up for presenting and answering survey questions to respondents, and the collection of the survey data to the central administration.	Memobust definition (2014)		Theme: Mixed Mode Data Collection – design issues
Data Integration	The process of combining <u>data</u> from two or more sources to produce statistical outputs.	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) Method: RAS; (3) Method: Stone; (4) Theme: Macro Integration.
Data provider	The unit that actually reports the data about the reporting unit in the name of the data supplier. This could be a representative, e.g. an accounting firm.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys. (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Collection and Use of Secondary Data

Data reconciliation	The process of adjusting <u>data</u> derived from two different sources to remove, or at least reduce, the impact of differences identified.	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) Method: RAS; (3) Method: Stone; (4) Theme: Macro Integration; (5) Method: Reconciling Conflicting Micro-Data; (6) Method: Generalised Ratio Adjustments; (7) Method: Minimum Adjustment Methods; (8) Method: Prorating; (9) Theme: Data fusion at micro level
Data set	Any organised collection of <u>data</u>	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) RAS; (3) Stone; (4) Macro Integration; (5) Method: Chow-Lin Method for Temporal Disaggregation
Data supplier	The unit which is formally responsible to provide data about its reporting unit(s). The survey organization has a legal relationship with the data supplier.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design
Data validation	see Editing	Memobust definition (2014)	Editing	Theme: Statistical Data Editing
DBMS	Database Management System.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Death rate	The death rate of a given reference period is the number of deaths as a percentage of the population of active enterprises	Memobust definition (2014)		Theme: Business Demography
Deductive editing	An umbrella term for editing methods that use logical reasoning to derive adjustments from the unedited data.	Memobust definition (2014)		(1) Method: Automatic Editing; (2) Method: Deductive Editing; (3) Theme: Statistical Data Editing
Deductive imputation	An umbrella term for imputation methods that use logical reasoning to derive imputed values in a deterministic manner.	CBS Methods Series Glossary	Logical imputation	(1) Method: Deductive Imputation; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Imputation under Edit Constraints
Deduplication	Taking the duplicate records out of a file, one by one, that occur multiple times, and that all relate to the same unit (in a certain period).	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching

Definition	Step 1 in the OQRM model, where the object and the focus area is defined.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Definitive interruption rate	The proportion of observation units for which the reporting unit has been successfully contacted, but has interrupted in cooperation before the very end of the questionnaire	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools
Degree	The degree of a point in a graph is the number of edges in the graph connected to this point.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Delphi method	A research tool in which opinions are solicited from many experts about a topic for which there is no consensus. The answers of other experts are fed back anonymously in several rounds until consensus is reached. The method is named after the Oracle of Delphi.	Daas and Arends-Toth (2012)		Theme: Collection and Use of Secondary Data
Dependencies	Step 10 in the OQRM model, where dependencies of a focus area with other focus areas are determined. Example: The soundness of methodology contributes to the accuracy of estimates.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Design burden	The burden which includes all aspects of the survey environment that are not directly associated with the respondent e.g. method of data collection, mode of collection and the contents of the survey, errors in sampling frame, incorrect sampling, etc.	Hedlin <i>et al.</i> (2005)		Theme: Response Burden
Design weight	For a sampling unit, it is the inverse probability of selection.	ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys, EUROSTAT 2013		(1) Theme: Weighting and Estimation; (2) Method: Generalised regression estimator; (3) Method: Preliminary estimates with design-based methods
Design weight	Weight which is the inverse of the inclusion probability.	Memobust definition (2014)		Method: Calibration
Design-consistency	Convergence in probability as the sample size increased.	Memobust definition (2014)		Theme: Weighting and Estimation
Deterministic imputation	A deterministic imputation method determines one unique value for the imputation of a missing or inconsistent data item. This means that when the imputation process is repeated, the same values will be imputed.	EDIMBUS Manual		(1) Theme: Donor Imputation; (2) Theme: Imputation; (3) Theme: Model-based Imputation
Deterministic record linkage	Linkage method that detects links if and only if there is a full agreement of unique identifiers or a set of common identifiers, the matching variables.	Memobust definition (2014)	Object identifiers	Theme: Probabilistic Record Linkage
Deterministic rounding	See: conventional rounding	Glossary on Statistical Disclosure Control (2014)	Conventional rounding	Theme: Statistical disclosure control methods for quantitative tables
DIME	Directors of Methodology	Eurostat website/CROS portal		Theme: The European Statistical System

DIMESA	Direcors' Meetings of Environmental Statistics and Accounts	Eurostat website/CROS portal		Theme: The European Statistical System
Direct estimator	An estimator of the target parameter for a given sub-population (domain) is said to be a direct estimator when it is based only on sample information from the sub-population itself. The more common direct estimators used in large scale business surveys are Calibration estimators.	Memobust definition (2014)		Theme: Weighting and Estimation
Direct estimator	An estimator which takes into account only domain-specific data. In many cases this estimator gives unacceptable results due to the fact that small areas are not represented in the sample by many units.	Memobust definition (2014)		(1) Method: Composite Estimators for Small Area Estimation; (2) Method: Synthetic Estimators for Small Area Estimation
Disaggregation	The breakdown of observations, usually within a common branch of a hierarchy, to a more detailed level to that at which detailed observations are taken.	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) Theme: Macro Integration
Disclosive cells	See: risky cells.	Glossary on Statistical Disclosure Control (2014)	Risky cells	Theme: Statistical disclosure control methods for quantitative tables
Disclosure	Disclosure relates to the inappropriate attribution of information to a data subject, whether an individual or an organisation. Disclosure has two components: identification and attribution.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
Disclosure risk	A disclosure risk occurs if an unacceptably narrow estimation of a respondent's confidential information is possible or if exact disclosure is possible with a high level of confidence.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
Dissemination	Supply of data in any form whatever: publications, access to databases, microfiches, telephone communications, etc.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
Dissemination	Dissemination is the release to users of information obtained through a statistical activity.	OECD Glossary of Statistical Terms		Theme: Dissemination of Business Statistics
Dissimilarity measure	A measure to express the differences between two objects or entities. Somewhat similar to a metric.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Distance function	See Metric	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Distance function	In the calibration procedure, a function which measures the distance between initial design weights and calibration weights.	Memobust definition (2014)		Method: Calibration
DMES	Directors of Macro-Economic Statistics	Eurostat website/CROS portal		Theme: The European Statistical System

Dominance rule	See: (n,k) rule.	Memobust definition (2014)	(n,k) rule; concentration rule	Theme: Statistical disclosure control methods for quantitative tables
Donor file	File where one variable (say Z) has been observed and that will be used for imputation purposes on a file where Z is missing (recipient file)	Memobust definition (2014)		Theme: Statistical Matching
Donor imputation	An imputation method for which the imputed value is copied from a donor record that closely matches the recipient record on many features.	CBS Methods Series Glossary		(1) Theme: Donor Imputation; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Imputation under Edit Constraints
Doubt category	A category that can be used if a description cannot be classified with sufficient certainty. The same or other coders can review the descriptions designated as such at a later stage in the process.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Automatic coding based on semantic networks; (3) Theme: Coding
DPoD	Day - Part of Day combination. The basic time unit for allocating CATI interviewers. For the sake of concreteness we have assumed three DPoD's in this module: morning, afternoon, evening. Other choices are possible and allowed, however.	Memobust definition (2014)		Theme: CATI Allocation
DSS	Directors of Social Statistics	Eurostat website/CROS portal		Theme: The European Statistical System
EBLUP	Empirical Best Linear Unbiased Predictor – estimator obtained by plugging in the estimation of variance components in a BLUP estimator, that is the estimator that in the class of all linear unbiased estimator minimize square loss. Unbiasedness is referred to model distribution.	Memobust definition (2014)		(1) Method: EBLUP Unit level for Small Area Estimation; (2) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)
EBLUP (Empirical Best Linear Unbiased Predictor)	Predictor obtained by plugging in the estimates of the variance components in the BLUP.	Memobust definition (2014)		Method: Small area estimation methods for time series data
ECSC	European Coal and Steel Community	ECSC		Theme: The European Statistical System
EDI	Electronic Data Interchange	Willeboordse <i>et al.</i> (1997)		Theme: Response Burden
Edit	A logical condition or a restriction to the value of a data item or a data group which must be met if the data is to be considered correct.	EDIMBUS Manual		(1) Method: Generalised Ratio Adjustments; (2) Method: Minimum Adjustment Methods; (3) Method: Prorating; (4) Method: Reconciling Conflicting Micro-Data; (5) Theme: Data fusion at micro level; (6) Theme: Imputation for Longitudinal Data; (7) Theme: Imputation under Edit Constraints; (8) Theme: Editing for Longitudinal Data

Edit	A check (logical condition or a restriction to the value of a data item or a group of data items) that identifies missing, invalid or inconsistent values or that points to data records that are potentially in error.	EDIMBUS Manual	Edit rule, Checking rule	(1) Method: Automatic Editing; (2) Theme: Statistical Data Editing
Edit	A logical condition or a restriction to the value of a data item or a data group which must be met.	EDIMBUS Manual		Theme: Selective Editing
Edit constraints	see Edit	Memobust definition (2014)		
Edit distance	Distance that returns the minimum cost in terms of insertion, deletions and substitutions needed to transform a string of one record into the corresponding string of the compared record	Memobust definition (2014)		Theme: Probabilistic Record Linkage
Edit group	A component of an edit rule that identifies a (homogeneous) subset of the units for which the acceptance region is applicable for the test variable.	Norberg (2011)		Method: Manual Editing
Edit rule	see Edit	Memobust definition (2014)	Edit	
Editing	The application of checks that identify missing, invalid or inconsistent entries or that point to data records that are potentially in error.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Editing	An activity that aims to detect, understand, and correct missing values and erroneous values in data.	Memobust definition (2014)	Data cleaning, Data validation	Theme: Statistical Data Editing
Editor	A person who performs interactive or manual editing.	Memobust definition (2014)		Method: Manual Editing
EEA	European Economic Area	EEA		Theme: The European Statistical System
EFQM	European Foundation for Quality Management	Memobust definition (2014)		Theme: Quality and Risk Management Models
EFTA	European Free Trade Association.	EFTA		Theme: The European Statistical System
Employer enterprise birth	Birth of an enterprise with at least one employee. This population consists of enterprise births that have at least one employee in the birth year and of enterprises that existed before the year in consideration, but were below the threshold of one employee	Eurostat-OECD Manual on Business Demography Statistics (chapter 5).		Theme: Business Demography
Employer enterprise death	An Employee Enterprise death occurs either as an enterprise death with at least one employee in the year of death or as an exit by decline, moving below the threshold of one employee.	Eurostat-OECD Manual on Business Demography Statistics (chapter 7).		Theme: Business Demography

Enterprise	The enterprise is the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit. Note: The definition does not limit enterprise to one country. However, by convention this is generally done in the European statistical context. Enterprise may thus be used elsewhere in the meaning of enterprise group, in America also in the meaning of truncated enterprise group	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III A		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Statistical Registers and Frames – Survey frames for business surveys; (4) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (5) Theme: Statistical Registers and Frames – The statistical units and the business register.
Enterprise Birth	A birth amounts to the creation of a combination of production factors with the restriction that no other enterprises are involved in the event. Births do not include entries into the population due to mergers, break-ups, split-off or restructuring of a set of enterprises. It does not include entries into a sub-population resulting only from a change of activity.	Definition of SBS Regulation variables, Eurostat-OECD Manual on Business Demography Statistics (chapter 5).	Real Birth Enterprise	Theme: Business Demography
Enterprise Death	A death amounts to the dissolution of a combination of production factors with the restriction that no other enterprises are involved in the event. Deaths do not include exits from the population due to mergers, take-overs, break-ups or restructuring of a set of enterprises. It does not include exits from a sub-population resulting only from a change of activity.	Definition of SBS Regulation variables, Eurostat-OECD Manual on Business Demography Statistics (chapter 7).	Real Death Enterprise	Theme: Business Demography
Enterprise group	An enterprise group is an association of enterprises bound together by legal and/or financial links. A group of enterprises can have more than one decision-making centre, especially for policy on production, sales and profit. It may centralise certain aspects of financial management and taxation. It constitutes an economic entity, which is empowered to make choices, particularly concerning the units that it comprises.	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III C		(1) Theme: Derivation of Statistical Units; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (4) Theme: Statistical Registers and Frames – The statistical units and the business register.
Error	In general, a mistake or error in the colloquial sense.	Eurostat's Concepts and Definitions Database (2013)	Mistake	Theme: Quality of Statistics
Error message	For electronic questionnaire: a window containing a text that described what sort of inconsistency happened and the list of variables involved in it	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools

ESAC	European Statistical Advisory Committee	Regulation N° 99/2013		Theme: The European Statistical System
ESCB	European System of Central Banks	Regulation N° 223/2009		Theme: The European Statistical System
ESGAB	European Statistical Governance Advisory Board	Decision N° 235 (2008)		Theme: The European Statistical System
ESS	European Statistical System. The ESS is the partnership comprising Eurostat, National Statistical Institutes (NSIs) and other national statistical bodies responsible in each Member State (MS) for producing and disseminating European statistics.	ESS Regulation No 223 (2009)		(1) Theme: Different types of surveys; (2) Theme: The European Statistical System
ESSC	European Statistical System Committee	Regulation N° 223/2009		Theme: The European Statistical System
ESS-VIP	ESS Vision Implementation Project	Eurostat website/CROS portal		Theme: The European Statistical System
Establishment	An establishment is defined by the System of National Accounts (SNA) as an enterprise, or part of an enterprise, that is situated in a single location and in which only a single (non-ancillary) productive activity is carried out or in which the principal productive activity accounts for most of the value added. According to the Regulation on statistical units the local kind-of-activity unit (local KAU) corresponds to the operational definition of the establishment. According to the European System of Accounts (ESA) the local KAU is called the establishment in the SNA and ISIC Rev. 3.	System of National Accounts (SNA) 1993, (5.21), P. 116, European System of Accounts (ESA) 1995, [2.106] footnote 15 and Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III G (2)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – The statistical units and the business register; (3) Theme: Derivation of Statistical Units; (4) Theme: Derivation of Statistical Units
Estimate	The particular value yielded by an estimator in a given set of circumstances.	SDMX (2009)	Estimated value.	(1) Method: Preliminary estimates with design-based methods; (2) Theme: Design of Estimation – Some Practical Issues; (3) Theme: Quality of Statistics; (4) Method: Balanced Sampling for Multi-Way Stratification; (5) Method: Subsampling for Preliminary Estimates; (6) Theme: Methods and Quality; (7) Theme: Quality and Risk Management Models
Estimator	A rule or method of estimating a parameter of a population.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Method: Little and Su Method; (3) Method: Subsampling for Preliminary Estimates; (4) Method: Preliminary estimates with design-based methods; (5) Theme: Design of Estimation – Some Practical Issues; (6) Theme: Quality of Statistics

Estimator effect	Ratio between variance of the estimator and variance of the HT estimator for the same sampling design.	Memobust definition (2014)		Method: Generalised regression estimator
ETL	Extract Transform Load. A set of operations to make an external data set suitable for further processing, e.g. at a statistical office.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Evaluation	The systematic and objective assessment of an on-going statistical production process, its design, implementation and results. The aim is to determine the relevance and fulfillment of objectives, development efficiency, effectiveness, impact and sustainability.	GSBPM (2009)		Theme: Evaluation of Business Statistics
Experiment embedded	The sample of a survey is randomly divided into several groups, which are differently treated and then compared with regard to a hypothesis about treatment effects.	Memobust definition (2014)		Theme: Repeated Surveys
Exports of goods and services	Exports of goods and services consist of transactions in goods and services (sales, barter, and gifts) from residents to non-residents.	ESA (2010)		Theme: Manual Integration
Failure rate	The proportion of records in the unedited data that fail a given edit.	Memobust definition (2014)		Method: Manual Editing
False match rate	number of incorrectly linked record pairs divided by the total number of linked record pairs	Memobust definition (2014)	False positive rate	Method: Fellegi-Sunter and Jaro Approach to Record Linkage
False matches	Matched records which do not represent the same entity	Memobust definition (2014)	mismatch, false positive match, Type I error	Theme: Probabilistic Record Linkage
False negative match	See Missed Match	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
False non-match rate	number of incorrectly unlinked record pairs divided by the total number of true match record pairs	Memobust definition (2014)	False negative rate, Missed match rate	Method: Fellegi-Sunter and Jaro Approach to Record Linkage
False nonmatches	Unmatched records not correctly classified, that imply truly matched entities were not linked.	Memobust definition (2014)	missed match, false negative match, Type II error	Theme: Probabilistic Record Linkage
False positive match	See Mismatch	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Fatal edit rule	see Hard edit rule	Memobust definition (2014)		

Feasible matching graph	A subgraph of an MC graph that satisfies the criteria that are established for the matching graph. These criteria relate at least to the maximum degree of the points or a part thereof (degree restrictions). The word 'feasible' is used in the sense of 'feasible solution'.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Fellegi-Sunter method	Matching method described in Fellegi and Sunter (1969).	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
First order autoregressive process AR(1)	Model belonging to the class of autoregressive (AR), in which the current level is modelled on the basis of the previous levels.	Memobust definition (2014)		Method: Small area estimation methods for time series data
Flow variable	A flow variable is measured over an interval of time. (see also stock variable)	Memobust definition (2014)		Theme: Macro Integration
Focus area	Combination of an object and one accompanying attribute. Examples: accuracy of estimates, soundness of methodology, clarity of a description.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Follow-up	The work performed by an editor during manual editing to handle an edit failure.	Memobust definition (2014)		(1) Method: Manual Editing; (2) Theme: Macro-Editing
Follow-up	A further attempt to obtain information from an individual or a reporting unit in a survey or field experiment because the initial attempt has failed or later information is available.	SDMX (2009)		(1) Method: Preliminary estimates with design-based methods; (2) Method: Subsampling for Preliminary Estimates; (3) Theme: Data Collection; (4) Theme: Data Collection: Techniques and Tools
Foreign key	A key value that occurs in a record but is not suitable to identify the record itself. A foreign key is therefore located outside the key of a data set. The purpose of a foreign key is to make a match with a record in another data set which, for example, includes additional data based on that key.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Frame	A list, map or other specification of the units, which define a population to be completely enumerated or sampled	SDMX (2009), CODED		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Sample selection; (3) Theme: Asymmetry in Statistics – European Register for Multinationals (EGR)
Frame error	Error caused by imperfections in the frame (business register, population register, area sample, etc.) from which units are selected for inclusion in surveys or censuses.	NQAF (2012)		Theme: Quality and Risk Management Models

Frame population	Frame population is the set of population units described in the survey frame. Remark: Because of the coverage error of the frame population (reference scope) the frame population and the target population is not overlap each other. The part of the frame population belonging to the target population is the survey population	Memobust definition (2014)	Reference scope	(1) Theme: Statistical Registers and Frames – Survey frames for business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Asymmetry in Statistics – European Register for Multinationals (EGR); (4) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames
Frequency	The time interval at which observations occur over a given time period.	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Theme: Macro Integration
Frequency of register maintenance	Frequency of register maintenance is the time interval of the register content alterations. Remark: Registers can be maintained from different sources with different frequencies. In such cases, the most frequently used source determines the frequency of the register maintenance	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – Statistical register and survey frame design
Frequency tables	See: Tables of frequency (count) data	Memobust definition (2014)		Theme: Statistical disclosure control methods for quantitative tables
Fuzzy string matching	The comparison of two texts, for which the outcome (usually) is a scalar that indicates the extent to which the texts are similar.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Automatic coding based on semantic networks; (3) Method: Computer-assisted coding
Gazelle	A gazelle is a high-growth enterprise that is up to 5 years old.	Eurostat-OECD Manual on Business Demography Statistics.		Theme: Business Demography
Generalised regression estimator	An estimator that can be written as the sum of the Horvitz Thompson estimator (HT) and a weighted difference between known totals and their HT estimator.	Memobust definition (2014)	GREG	Method: Generalised regression estimator
Global score function	A global score function is the combination of all defined local score functions, i.e., score functions defined for individual variables.	EDIMBUS Manual		Theme: Selective Editing
Gross burden	All additional costs to businesses arising from their inclusion in a survey if all sampled businesses respond.	Willeboordse <i>et al.</i> (1997), DETI (2009)		Theme: Response Burden

Gross domestic product	Gross Domestic Product (GDP) is one of the key aggregates in the ESA. GDP is a measure of the total economic activity taking place on an economic territory which leads to output meeting the final demands of the economy. There are three ways of measuring GDP at market prices: (1) The production approach, as the sum of the values added by all activities which produce goods and services, plus taxes less subsidies on products; (2) The expenditure approach, as the total of all final expenditures made in either consuming the final output of the economy, or in adding to wealth, plus exports less imports of goods and services; (3) The income approach, as the total of all incomes earned in the process of producing goods and services plus taxes less subsidies on products.	Memobust definition (2014)	GDP	Theme: Manual Integration
Gross measurement errors	are observations that are not true values	Memobust definition (2014)		Method: Outlier Treatment
GSBPM	The Generic Statistical Business Process Model provides a framework to describe the statistical production process in terms of standard components (phases and sub-processes).	GSBPM (2009)		(1) Theme: Specification of User Needs for Business Statistics; (2) Theme: Dissemination of Business Statistics; (3) Theme: Evaluation of Business Statistics
Hamming distance	Distance between two records on a matching key, measured by counting the number of variables with different scores.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching; (5) Theme: Probabilistic Record Linkage
Hard Constraint	A constraint that should hold exactly	Memobust definition (2014)		Method: Denton for Benchmarking
Hard edit rule	An edit rule that identifies data errors with certainty.	EDIMBUS Manual	Fatal edit rule, Logical edit rule	(1) Method: Automatic Editing; (2) Method: Manual Editing; (3) Theme: Statistical Data Editing
Heteroscedasticity	A collection of random variables is heteroscedastic if there are sub-populations that have different variabilities than others.	Memobust definition (2014)		Method: Generalised regression estimator
High-growth enterprise	A high-growth enterprise is an enterprise with average annualised growth greater than 20% per annum, over a three year period. Growth can be measured by the number of employees or by turnover.	Eurostat-OECD Manual on Business Demography Statistics.		Theme: Business Demography
History	Step 9 in the OQRM model, where the history of the focus area is formulated.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Hit rate	(1) The proportion of error flags that an edit generates which point to true errors. Or (2) the proportion of error flags generated by an edit that are associated with adjustments made to the data. [Note: this is a practical approximation to the formal definition given under (1).]	EDIMBUS Manual		Method: Manual Editing

Horizontal aggregation	Horizontal aggregation: aggregation, e.g. by country	European Communities (2001)		Theme: Seasonal adjustment – introduction and general description
Horvitz-Thompson estimator	Weighted sum with weights given by the inverse of inclusion probabilities.	Module XIX.0	HT	Method: Generalised regression estimator
Hot-deck imputation	A donor record is found from the same survey as the record with the missing item(s). This donor record is used to supply values for the missing or inconsistent data item(s).	EDIMBUS Manual	Donor imputation	(1) Method: Minimum Adjustment Methods; (2) Method: Reconciling Conflicting Micro-Data; (3) Method: Statistical Matching Methods; (4) Theme: Data fusion at micro level; (5) Theme: Statistical Matching; (6) Theme: Donor Imputation
Hypercube method	A heuristic method for protecting tables through cell suppression.	Memobust definition (2014)		Theme: Statistical disclosure control methods for quantitative tables
Hypernym	A generalisation of a term or a more general term. Opposite of 'hyponym'.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on semantic networks; (2) Method: Computer-assisted coding
Hyponym	A specialisation of a term or a more specific term. Opposite of 'hypernym'.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on semantic networks; (2) Method: Computer-assisted coding
Importance	Step 7 in the OQRM model, where the importance of the focus area will be determined, related to the output quality or other objectives.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Imports of goods and services	Imports of goods and services consist of transactions in goods and services (purchases, barter, and gifts) from non-residents to residents.	ESA (2010)		Theme: Manual Integration
Imputation	(1) A procedure for entering a value for a specific data item where the response is missing or unusable. Or (2) a value that is filled in during the process described under (1).	UN/ECE Glossary of Terms on Statistical Data Editing (2007), CBS Methods Series Glossary.	Imputing, Imputed value	(1) Theme: Donor Imputation; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Imputation under Edit Constraints; (5) Theme: Model-based Imputation; (6) Method: Minimum Adjustment Methods; (7) Method: Reconciling Conflicting Micro-Data; (8) Method: Statistical Matching Methods; (9) Theme: Data fusion at micro level; (10) Theme: Design of Estimation – Some Practical Issues; (11) Theme: Quality of Statistics; (12) Theme: Statistical Data Editing
Imputation class	A subpopulation for which imputation is carried out, without using any information from the rest of the population. Different imputation methods can be used for different imputation classes.	CBS Methods Series Glossary	Imputation group	(1) Theme: Donor Imputation; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Model-based Imputation

Imputed value	see Imputation (2)	Memobust definition (2014)	Imputation	
Imputing	see Imputation (1)	Memobust definition (2014)	Imputation	
In control	Step 6 in the OQRM model, where is determined if the requirements for a focus area are met and/or if the residual risk is acceptable.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Inclusion probability	For a sampling design without replacement, the probability that a particular unit from the population is drawn. This probability may vary between units, depending on the sampling design.	CBS Methods Series Glossary		(1) Theme: Imputation; (2) Method: Balanced Sampling for Multi-Way Stratification
Inclusion probability	The probability that a member of a population will appear in a given sample.	Memobust definition (2014)		(1) Method: Calibration; (2) Method: Synthetic Estimators for Small Area Estimation
Indicator	A data element that represents statistical data for a specified time, place, and other characteristics, and is corrected for at least one dimension (usually size) to allow for meaningful comparisons.	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Theme: Macro Integration; (4) Theme: Data Collection: Techniques and Tools
Indirect estimator	An estimator that “borrows strength” by taking into account values of the variable under study from outside the domain or time period. These values are brought into the estimation process through a properly chosen model and may come from different sources, for instance censuses or administrative registers.	Memobust definition (2014)		Method: Synthetic Estimators for Small Area Estimation
Industry	A group of producing units, having similar output and production processes; the classification of industries is based on NACE	Memobust definition (2014)		Theme: Manual Integration
Influential error	An error that has a significant influence on figures to be published.	CBS Methods Series Glossary		(1) Theme: Editing Administrative Data; (2) Theme: Editing for Longitudinal Data; (3) Theme: Macro-Editing; (4) Theme: Selective Editing; (5) Theme: Statistical Data Editing
Input editing	Editing that is performed as data is input, e.g., during an interview.	EDIMBUS Manual		Theme: Selective Editing

Institutional unit	The institutional unit is an elementary economic decision-making centre characterized by uniformity of behavior and decision-making autonomy in the exercise of its principal function. A unit is regarded as constituting an institutional unit if it has decision-making autonomy in respect of its principal function and keeps a complete set of accounts. In order to be said to have autonomy of decision in respect of its principal function, a unit must be responsible and accountable for the decisions and actions it takes. In order to be said to keep a complete set of accounts, a unit must keep accounting records covering all its economic and financial transactions carried out during the accounting period, as well as a balance sheet of assets and liabilities. Remark: According to the Regulation on statistical units an institutional unit corresponds to an enterprise in the corporate enterprises sector.	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex, Section III B.		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Statistical Registers and Frames – The statistical units and the business register
Interaction burden	A product of the relationship between respondent burden and design burden, e.g. requirement concerning memory and effort to be made, familiarity of the respondent with IT methods and tools, etc.	Hedlin <i>et al.</i> (2005)		Theme: Response Burden
Interactive coding	Coding using an interactive program, which presents the necessary background or other information to a coder, who makes all the coding decisions. The program also processes the answers (and the possible reason for the choices as indicated by the coder).	Hacking & Willenborg (2012)		(1) Method: Computer-assisted coding; (2) Theme: Coding
Interactive editing	An editing method for which a computer program checks the data and a human editor adjusts the data.	CBS Methods Series Glossary	Manual editing	(1) Method: Automatic Editing; (2) Method: Manual Editing; (3) Theme: Editing Administrative Data; (4) Theme: Editing for Longitudinal Data; (5) Theme: Macro-Editing; (6) Theme: Selective Editing; (7) Theme: Statistical Data Editing
Intermediate consumption	Intermediate consumption consists of goods and services consumed as inputs by a process of production, excluding fixed assets whose consumption is recorded as consumption of fixed capital. The goods and services are either transformed or used up by the production process.	ESA (2010)		Theme: Manual Integration
Internet survey	See Web Survey.	Memobust definition (2014)		Theme: Mixed Mode Data Collection – design issues
Interviewer effect	Effects on respondents' answers stemming from the different ways that interviewers administer the same survey.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Data Collection: Techniques and Tools
Interviewer error	Effects on respondents' answers stemming from the different ways those interviewers administer the same survey. .	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics

Interviewer-administered mode	An interviewer administers and guides the respondent when answering the survey questions.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Intruder	A data user who attempts to link a respondent to a microdata record or make attributions about particular population units from aggregate data. Intruders may be motivated by a wish to discredit or otherwise harm the NSI, the survey or the government in general, to gain notoriety or publicity, or to gain profitable knowledge about particular respondents.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
Investment	Gross fixed capital formation consists of resident producers' acquisitions, less disposals, of fixed assets during a given period plus certain additions to the value of non-produced assets realised by the productive activity of producer or institutional units. Fixed assets are produced assets used in production for more than one year.	ESA (2010)	Gross fixed capital formation	Theme: Manual Integration
Inward FATS	'Inward statistics on foreign affiliates' shall mean statistics describing the activity of foreign affiliates resident in the compiling economy.	Foreign Affiliates Statistics (FATS) recommendation manual, version 2012		Theme: Asymmetry in Statistics – European Register for Multinationals (EGR)
Irregular component	This is the residual time series that results from the removal of estimated seasonal and other systematic calendar-related components of an observed time series, along with the removal of an estimated trend-cycle component	US Census Bureau		Method: Seasonal adjustment of economic time series
ITDG	Information Technology Directors Group	Eurostat website/CROS portal		Theme: The European Statistical System
Item non-response	Item non-response occurs when a respondent provides some, but not all, of the requested information, or if some of the reported information is not usable.	EDIMBUS Manual	Partial non-response	(1) Theme: Questionnaire Design; (2) Theme: Editing During Data Collection; (3) Theme: Testing the Questionnaire; (4) Theme: Response Process; (5) Theme: Imputation; (6) Theme: Imputation for Longitudinal Data; (7) Method: Little and Su Method; (8) Theme: Quality of Statistics
Item response rate	The ratio of the number of units which have provided data for a given data item to the total number of units from which data are to be collected or to the number of units that have provided information at least for some data items. It can indirectly measure the level of response burden.	Eurostat (2009).		Theme: Response Burden
Iterative proportional fitting	See multiplicative weighting.	Memobust definition (2014)		(1) Method: RAS; (2) Method: Stone

Jaro distance	Distance counts the number of common characters and the number of transpositions of characters (same character with a different position in the string) between two strings;	Memobust definition (2014)		Theme: Probabilistic Record Linkage
Joining	A form of matching used for databases and in which matching is based on matching keys being identical.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Kalman filter	An iterative technique of dynamic linear modelling, used mainly for estimating the parameters of autoregressive moving-average time series models with Gaussian residuals.	Memobust definition (2014)		Method: Preliminary estimates with model-based methods
Key	See: Object identifier	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Key word	Word in a description that is usable for coding, in contrast to a stop word.	Hacking & Willenborg (2012)		Theme: Coding
Kind-of-activity unit	The kind of activity unit (KAU) groups all the parts of an enterprise contributing to the performance of an activity at class level (4-digits) of NACE Rev. 2 and corresponds to one or more operational subdivisions of the enterprise. The enterprise's information system must be capable of indicating or calculating for each KAU at least the production value, intermediate consumption, manpower costs, the operating surplus and employment and gross fixed capital formation.	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III D		(1) Theme: Derivation of Statistical Units; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of busine; (4) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (5) Theme: Statistical Registers and Frames – The statistical units and the business register.
K-Nearest neighbor imputation	the imputed value is an average of the closest k donors chosen in such a way that some measure of distance between the donors and the recipient is minimized.	EDIMBUS Manual	Distance hot deck	Theme: Statistical Matching
Lagrange multiplier technique	In mathematical optimization, the technique of Lagrange multipliers (named after Joseph Louis Lagrange) provides a strategy for finding the maxima and minima of a function subject to constraints.	Memobust definition (2014)		Method: Denton for Benchmarking
Large outlier	the Y values are extremely larger than the other Y values of the “normal” units	Memobust definition (2014)		Method: Outlier Treatment
Least median of squares	statistical method that attempts to minimise the median of all sample squared residuals	Memobust definition (2014)		Method: Outlier Treatment

Least squares method	One of the most popular methods of finding estimates based on fitting a mathematical model to data, aiming at minimizing the sum of squares of deviations between observed and fitted values.	Memobust definition (2014)		Method: Synthetic Estimators for Small Area Estimation
Legal form	The legal form is defined according to national legislation. It is useful for eliminating ambiguity in identification searches and as the possible criterion for selection or stratification for surveys. It is also used for defining the institutional sector. Statistics according to legal form are produced in business demography. The character of legal or natural person is decisive in fiscal terms, because the tax regime applicable to the unit depends on this. It means that any statistical register fed with fiscal records will have that information. Experience has shown that legal form will often be useful to make adjustments to information collection processes and questionnaires on the legal unit operating an enterprise. A code representing the legal form should therefore be recorded in accordance with the classification of legal forms or categories. The following legal forms can be found in most Member States: Sole proprietorship, Partnership, Limited liability companies, Co-operative societies, Non-profit making bodies, Enterprises with other forms of legal constitution.	Business Register Recommendations Manual (edition 2010), chapter 5, characteristic 1.6	Legal status	Theme: Statistical Registers and Frames – Survey frames for business surveys
Legal local unit	A legal local unit is a part of a legal unit that is located at a certain address. A legal local unit can operate in several different industries. In practice, a legal local unit is the same as a local unit.	Memobust definition (2014)		Theme: Derivation of Statistical Units
Legal unit	Legal units include: - Legal persons whose existence is recognized by law independently of the individuals or institutions which may own them or are members of them. - Natural persons who are engaged in an economic activity in their own right. The legal unit always forms, either by itself or sometimes in combination with other legal units, the legal basis for the statistical unit known as the 'enterprise'.	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section II A 3 - 4.		(1) Theme: Derivation of Statistical Units; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys
Levenshtein distance	Distance measure between two strings, defined as the minimum number of mutations needed to transform one string into the other. A mutation is one of three operations: insertion, deletion or substitution of a character/l into a stringl.	Hacking & Willenborg (2012)	Damerau-Levenshtein distance	(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Automatic coding based on semantic networks; (3) Method: Computer-assisted coding; (4) Theme: Object matching; (5) Method: Object Identifier Matching; (6) Method: Unweighted Matching; (7) Method: Weighted Matching
Leverage	Outlier in the x-direction	Memobust definition (2014)		Method: Outlier Treatment

Linear mixed model (LMM)	Linear model containing both fixed and random effects.	Memobust definition (2014)		Method: Small area estimation methods for time series data
Linked tables	A set of tables with common cells.	Memobust definition (2014)		Theme: Statistical disclosure control methods for quantitative tables
Local kind-of-activity unit	The local kind-of-activity unit (local KAU) is the part of a KAU which corresponds to a local unit. The local KAU corresponds to the operational definition of the establishment. According to the European System of Accounts (ESA) the local KAU is called the establishment in the SNA and ISIC Rev. 3.	Council Regulation (EEC) No 696/93, of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III G, and European System of Accounts (ESA) 1995, [2.106], footnote 15		(1) Theme: Derivation of Statistical Units; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business; (4) Theme: Statistical Registers and Frames – Survey frames for business surveys; (5) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (6) Theme: Statistical Registers and Frames – The statistical units and the business register
Local unit	The local unit is an enterprise or part thereof (e.g. a workshop, factory, warehouse, office, mine or depot) situated in a geographically identified place. At or from this place economic activity is carried out for which - save for certain exceptions - one or more persons work (even if only part-time) for one and the same enterprise.	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III F.		(1) Theme: Derivation of Statistical Units; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The statistical units and the business register
Local unit of homogeneous production	The local unit of homogeneous production (local UHP) is the part of a unit of homogeneous production which corresponds to a local unit.	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III H		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Statistical Registers and Frames – The statistical units and the business register
LOCF	Last Observation Carried Forward. One method of handling missing data based on existing data.	Memobust definition (2014)		Theme: Imputation for Longitudinal Data
Log	A file that contains log information.	Memobust definition (2014)		Theme: Logging
Log information	Metadata produced during a specific run of a process.	Memobust definition (2014)	A set of logging indicators	Theme: Logging
Logging	Activity of producing log information in a log	Memobust definition (2014)	Tracing	Theme: Logging

Logging indicator	A variable that is logged.	Memobust definition (2014)	Log item	Theme: Logging
Logical edit rule	see Hard edit rule	Memobust definition (2014)	Hard edit rule	
Logical imputation	see Deductive imputation	Memobust definition (2014)	Deductive imputation	(1) Method: Deductive Imputation; (2) Theme: Imputation; (3) Theme: Imputation under Edit Constraints
Longitudinal data	Longitudinal data occurs when the same variables of the same units are measured several times at different moments.	Memobust definition (2014)		(1) Theme: Design of Estimation – Some Practical Issues; (2) Theme: Repeated Surveys; (3) Method: Little and Su Method; (4) Method: Subsampling for Preliminary Estimates; (5) Theme: Different types of surveys; (6) Method: Preliminary estimates with design-based methods
Longitudinal imputation	An umbrella term for imputation methods that make use of observed values for the same variable at other times, either for the same object or for different objects.	CBS Methods Series Glossary		(1) Theme: Imputation; (2) Theme: Imputation for Longitudinal Data
Longitudinal sampling design	Sampling design over time of a given unit of the population.	Memobust definition (2014)		Method: Sample co-ordination using Poisson sampling with permanent random numbers
Lower bound	The lowest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Macro integration	Integrating data from different sources on an aggregate level, to enable a coherent analysis of the data, and to increase the accuracy of estimates.	Memobust definition (2014)	Balancing	(1) Method: Denton for Benchmarking; (2) Method: RAS; (3) Method: Stone; (4) Theme: Macro Integration; (5) Theme: Manual Integration
Macrodata	The result of a statistical transformation process in the form of aggregated information.	SDMX (2009)	Tabular data	(1) Method: Denton for Benchmarking; (2) Method: RAS; (3) Method: Stone; (4) Theme: Macro Integration; (5) Theme: Statistical disclosure control methods for quantitative tables
Macro-editing	An umbrella term for editing methods that (initially) check the data on an aggregate level.	CBS Methods Series Glossary	Output editing	Theme: Editing Administrative Data
Macro-editing	A procedure for tracking suspicious data by checking aggregates or applying statistical methods on all records or on a subset of them.	SDMX (2009)	Output editing	(1) Theme: Macro-Editing; (2) Theme: Statistical Data Editing
Manual coding	Coding performed by a coder, without substantial support from a program.	Hacking & Willenborg (2012)		(1) Method: Computer-assisted coding; (2) Theme: Coding
Manual editing	see Interactive editing	Memobust definition (2014)	Interactive editing	

Marginal table	Table derived from a bigger table by aggregation.	Memobust definition (2014)	Sub table	Theme: Statistical disclosure control methods for quantitative tables
Master frame	Master frame is a snapshot of a register (union of registers) to assign the survey frames based on the given register (registers). Remark: An example of the master frame is the snapshot of the business register to define the survey frames of different economic statistical data collections. Another example can be the snapshot of the address register to make a common frame for population surveys. The common master frame, the common reference period helps the integration and linking of statistical data coming from different surveys.	Handbook on the design and implementation of business surveys		(1) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Asymmetry in Statistics – European Register for Multinationals (EGR)
Matching	The process of bringing together data (represented in records) relating to units and spread over two data sets, based on common or very similar characteristics in the form of primary or object characteristic values.	Memobust definition (2014)	Record linkage, object linkage	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Matching candidate graph	A bipartite graph that represents the possible matches between records from two data sets. A bipartite graph is one where the set of points is the union of two disjoint sets, such that each edge has its endpoints in each of these sets.	Memobust definition (2014)	MC graph	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Matching key	One or multiple key variables that are used in two or more data sets to be matched.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Matching noise	Discrepancy between the data generation mechanism and the imputation generation mechanism. The larger the matching noise, the more distant the usual inferences on the matched data set will be from the inferences that could have been done if the sample was completely observed	Memobust definition (2014)		Theme: Statistical Matching
Matching variables	Common identifiers, either quantitative or qualitative, chosen in order to compare records among files	Memobust definition (2014)	Matching keys	(1) Theme: Probabilistic Record Linkage; (2) Method: Fellegi-Sunter and Jaro Approach to Record Linkage
Matching weight	A nonnegative function defined on the edges of a graph, which associates a non-negative value G with each edge of the G . When matching, this weight expresses how well/poorly records match.	Memobust definition (2014)	Weight	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching

Maximum Likelihood	Method to estimate a parameter of a probability distribution. More specifically it is the value that maximizes the likelihood function.	Memobust definition (2014)	ML	(1) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot); (2) Method: EBLUP Unit level for Small Area Estimation; (3) Method: Small area estimation methods for time series data
MC graph	See: Matching candidate graph	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Mean Square Error	Expected value of the square of the difference between the estimator and the parameter. It is the sum of variance and squared bias.	Eurostat's Concepts and Definitions Database (2013)	MSE	(1) Theme: Weighting and Estimation; (2) Method: EBLUP Unit level for Small Area Estimation; (3) Theme: Estimation with administrative data; (4) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot); (5) Theme: Quality of Statistics
Measure	Step 5 in the OQRM model where action are determined to manage the focus area in order to be in control of the focus area. Context: Preventive, corrective and signalling measures can be distinguished.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Measurement error	Error in reading, calculating or recording numerical value.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Measurement error	Measurement is characterized as the difference between the observed value of a variable and the true, but unobserved, value of that variable.	Measuring and Reporting Sources of Error in Surveys, FCSM 2001		Theme: Data fusion at micro level
Measurement error	Error in reading, calculating or recording numerical value. Context: Measurement errors occur when the response provided differs from the real value. Such errors may be attributable to the respondent, the interviewer, the questionnaire, the collection method or the respondent's record-keeping system. Errors may be random or they may result in a systematic bias if they are not random.	SDMX (2009)		Theme: Quality and Risk Management Models
Measurement error	The difference between the observed value of a variable and the true, but unobserved, value of that variable.	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) Theme: Macro Integration
Measurement errors	Measurement errors occur when the response provided differs from the real value; such errors may be attributable to the respondent, the interviewer, the questionnaire, the collection method or the respondent's record-keeping system. Such errors may be random or they may result in a systematic bias if they are not random.	Statistics Canada, "Statistics Canada Quality Guidelines", 4th edition, October 2003, page 59.		(1) Theme: Questionnaire Design; (2) Theme: Editing During Data Collection; (3) Theme: Testing the Questionnaire; (4) Theme: Response Process

Merger	This event can be seen as the opposite of a break-up. It involves a consolidation of the production factors of two or more enterprises into one new enterprise, in such a way that the previous enterprises are no longer recognisable. There is no continuity or survival, but the closures of the previous enterprises are not considered to be deaths. Similarly the new enterprise is not considered to be a birth.	Eurostat-OECD Manual on Business Demography Statistics (chapter 4).		Theme: Business Demography
Metadata	Information that is needed to be able to use and interpret statistics.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Methodology	A structured approach to solve a problem.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Metric	A metric d for a set X is defined as nonnegative function that measures how far two points in X are apart.	Memobust definition (2014)	Distance; Distance Function	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Micro data	Non-aggregated observations or measurements of characteristics of individual units.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Micro editing	An exhaustive check to find errors by inspecting each individual observation.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Micro integration	A method that matches data on individual statistical units from different sources, to obtain a combined data file with better information. The quality of the data is measured in terms of validity, reliability and consistency.	Memobust definition (2014)		(1) Method: Denton for Benchmarking; (2) Method: Stone
Microdata	Non-aggregated observations or measurements of characteristics of individual units.	SDMX (2009)		(1) Method: Denton for Benchmarking; (2) Method: RAS; (3) Method: Stone; (4) Theme: Macro Integration; (5) Theme: Methods and Quality; (6) Theme: Statistical Disclosure Control
Micro-selection	see Selective editing	Memobust definition (2014)	Selective editing	Theme: Statistical Data Editing
Misclassification	Erroneous classification of a subject into a category in which the subject does not belong. For instance, a business is classified in Trade instead of Industry.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Mismatch	A match that has been made erroneously.	Memobust definition (2014)	False positive match; Type I error	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Missed error rate	The proportion of errors in the unedited data that were not flagged by any edits in a given set.	Memobust definition (2014)		Method: Manual Editing

Missed match	A match that should have been made but was not.	Memobust definition (2014)	False negative match; Type II error	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Missing data	Observations which were planned and are missing	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Method: RAS; (4) Method: Stone; (5) Theme: Macro Integration
Mixed-mode survey	A survey where multiple modes are used to collect data from the sampled units in the data collection period of one survey.	Memobust definition (2014)		Theme: Mixed Mode Data Collection – design issues
Mode effect	The effect that using a specific mode has on the responses that are obtained in that mode. Mode effects may be interpreted as a form of measurement bias.	De Leeuw, Hox & Dillman (2008)		Theme: Mixed Mode Data Collection – design issues
Mode effect	A pure mode effect is essentially a measurement bias that is specifically attributable to the mode. In some surveys the mode effects are small because the same questionnaire can be used across all modes. Most problems occur when mail is combined with an interviewer-administered mode.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Mode of data collection	Mode refers to what medium is used when contacting the sample members to get their responses.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Model assumption error	Error that occurs due to the use of methods, such as calibration, generalized regression estimator, calculation based on full scope or constant scope, benchmarking, seasonal adjustment and other models not included in the preceding accuracy components, in order to calculate statistics or indexes.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Model assumption error	Model assumption errors occur with the use of methods, such as calibration, generalised regression estimator, calculation based on full scope or constant scope, benchmarking, seasonal adjustment and other models not included in the preceding accuracy components, in order to calculate statistics or indexes.	SDMX (2009)		(1) Theme: Methods and Quality; (2) Theme: Quality and Risk Management Models
Model based imputation	Imputation based on an explicitly described statistical model. E.g. use of averages, medians, regression equations, etc. to impute a variable.	EDIMBUS Manual		(1) Method: Statistical Matching Methods; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Imputation under Edit Constraints; (5) Theme: Model-based Imputation
Movement preservation principle	The property that <i>all</i> changes of the sub annual data are kept as much as possible at their initial values.	Memobust definition (2014)		Method: Denton for Benchmarking

Moving holiday effects	These are systematic changes in the values of a time series that are associated with the timing of moving holidays, i.e. holidays whose dates vary from year to year, such as Easter, Passover, Ramadan, Chinese New Year and U.S. Labor Day. Estimates of one or a combination of such effects define the moving holiday component of time series	US Census Bureau		Method: Seasonal adjustment of economic time series
Moving seasonality	Moving seasonality is a form of seasonality that accounts for the variability in the seasonal component of a time series from year to year	ABS (2008)		(1) Method: Seasonal adjustment of economic time series; (2) Theme: Seasonal adjustment – introduction and general description
Multiple activity business	A business operating in several economic activities	Memobust definition (2014)		Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered
Multiple imputation	An observation with failing and/or missing values is imputed several times stochastically. Multiple imputation allows under certain conditions the correct estimation of the variance due to imputation. This estimation is based on a combination of the within and the between variance of the multiply imputed data.	EDIMBUS Manual		Theme: Imputation
Multiple location business	A business operating in several geographical locations	Memobust definition (2014)		Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered
Multiplicative weighting	A form of weighting for which the weights are obtained by multiplying relevant weight factors, determined in an iterative process. Multiplicative weighting is also referred to as raking or iterative proportional fitting.	Memobust definition (2014)		(1) Method: RAS; (2) Method: Stone
Multistage sampling	A complex form of cluster sampling	Wikipedia Multistage Sampling		Theme: Sample selection
Multivariate imputation	Imputing several missing values in a record.	CBS Methods Series Glossary		Theme: Imputation
NACE	Classification of economic activities in the European Community (referred to as 'NACE Rev. 1' or 'NACE Rev. 1.1').	Council Regulation (EEC) No 3037/90.		Theme: Small area estimation
NACE	NACE (Statistical classification of economic activities) is the European standard classification of productive economic activities. NACE presents the universe of economic activities partitioned in such a way that a NACE code can be associated with a statistical unit carrying them out. NACE provides the framework for collecting and presenting a large range of statistical data according to economic activity in the fields of economic statistics.	NACE Rev.2		(1) Theme: Different types of surveys; (2) Theme: Estimation with administrative data

NACE	Nomenclatures statistique des activités économiques dans la Communauté Européenne.	NACE Rev. 2 (ISSN 1977-0375)		Theme: The European Statistical System
NACE	General Industrial Classification of Economic Activities within the European Communities (1970 version); Statistical classification of economic activities in the European Community (after 1970)	RAMON, Eurostat's metadata server -classification	NACE Rev. 2	(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Statistical Registers and Frames – Survey frames for business surveys; (4) Theme: Statistical Registers and Frames – The statistical units and the business register
Nearest neighbor imputation	The donor is chosen in such a way that some measure of distance between the donor and the recipient is minimized.	EDIMBUS Manual	Distance hot deck	Method: Statistical Matching Methods
Negative co-ordination	Minimize the overlap between samples	Memobust definition (2014)		Theme: Sample co-ordination
Net burden	The opposite of gross burden – the total costs actually incurred by responding businesses; this type of burden accounts for “benefits” enjoyed by respondents for their contribution whereas gross burden ignores them.	Willeboordse <i>et al.</i> (1997)		Theme: Response Burden
no terms	no terms	Memobust definition (2014)		Theme: GSBPM: Generic Statistical Business Process Model
No-answer	In telephone surveys: the line sounds but nobody answer the telephone	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools
Non probability sample	A sample in which the selection of units is based on factors other than random chance, e.g. convenience, prior experience or the judgement of a researcher.	SDMX (2009)		Theme: Weighting and Estimation
Non response	A form of non observation present in most surveys. Non response means failure to obtain a measurement on one or more study variables for one or more elements k selected for the survey. The term encompasses a wide variety of reasons for non observation: "impossible to contact", "not at home", "unable to answer", "incapacity", "hard core refusal", "inaccessible", "unreturned questionnaire", and others. In the first two cases contact with the selected element is never established.	SDMX (2009)		Method: Subsampling for Preliminary Estimates
Non response error	Error that occurs when the survey fails to get a response to one, or possibly all, of the questions.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics

Non response rate	In sample surveys, the failure to obtain information from a designated individual for any reason (death, absence or refusal to reply) is often called a non-response and the proportion of such individuals of the sample aimed at is called the non-response rate.	SDMX (2009)		Method: Subsampling for Preliminary Estimates
Nonbinding benchmarking	A benchmarking problem with at least one nonbinding annual alignment constraint.	Memobust definition (2014)		Method: Denton for Benchmarking
Nonbinding constraint	See soft constraint	Memobust definition (2014)		Method: Denton for Benchmarking
Non-probability sample	A sample in which the selection of units is based on factors other than random chance, e.g. convenience, prior experience or the judgement of a researcher.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Method: Subsampling for Preliminary Estimates; (3) Theme: Sample selection
non-representative outlier	are unique in population (in the sense that there is no other unit like them)	Memobust definition (2014)		Method: Outlier Treatment
Non-response	A form of non observation present in most surveys. Nonresponse means failure to obtain a measurement on one or more study variables for one or more elements k selected for the survey. The term encompasses a wide variety of reasons for non observation: "impossible to contact", "not at the address", "unable to answer", "incapacity", "hard core refusal", "inaccessible", "unreturned questionnaire", and others. In the first two cases contact with the selected element is never established.	SDMX (2009)		(1) Theme: Questionnaire Design; (2) Theme: Editing During Data Collection; (3) Theme: Testing the Questionnaire; (4) Theme: Response Process; (5) Method: Preliminary estimates with design-based methods
Non-response error	Context: Non-sampling error may arise from many different sources such as defects in the sampling frame, faulty demarcation of sample units, defects in the selection of sample units, mistakes in the collection of data due to personal variations, misunderstanding, bias, negligence or dishonesty on the part of the investigator or of the interviewer, mistakes at the stage of the processing of the data, etc.	Memobust definition (2014)		Theme: Quality and Risk Management Models
Non-response error	Non- sampling errors may be categorised as: § Coverage errors (or frame errors) due to divergences between the target population and the frame population ; § Measurement errors occurring during data collection. § Non-response errors caused by no data collected for a population unit or for some survey variables. § Processing errors due to errors introduced during data entry, data editing, sometimes coding and imputation. § Model assumption errors.	Memobust definition (2014)		Theme: Quality and Risk Management Models
Non-response error	Error that occurs when the survey fails to get a response to one, or possibly all, of the questions.	NQAF (2012)		Theme: Quality and Risk Management Models
Non-response error	Error in sample estimates which cannot be attributed to sampling fluctuations.	SDMX (2009)		Theme: Quality and Risk Management Models

Non-response rate	In sample surveys, the failure to obtain information from a designated unit for any reason is often called a nonresponse and the proportion of such units of the sample aimed at is called the nonresponse rate.	SDMX (2009)		Method: Preliminary estimates with design-based methods
Non-sampling error	Error in sample estimates which cannot be attributed to sampling fluctuations.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Non-sampling error	An error in sample estimates which cannot be attributed to sampling fluctuations.	The International Statistical Institute, The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press, 2003.		Theme: Editing During Data Collection
Normal distribution	One of the most widely known and used of all distributions, sometimes referred to as Gaussian distribution. The continuous probability distribution with two parameters: the expected value and the variance.	Memobust definition (2014)		Method: Synthetic Estimators for Small Area Estimation
Not at Random sample	A sample in which the selection of units is based on factors other than random chance, e.g. convenience, prior experience or the judgement of a researcher	SDMX (2009)		Method: Subsampling for Preliminary Estimates
Nowcast	A forecast relating to the current time (or, rather, to the recent past) and produces an estimate for the period just behind us, but for which no direct statistical observation has been made.	Daas and Arends-Toth (2012)		(1) Theme: Collection and Use of Secondary Data; (2) Theme: Estimation with administrative data
NSA	National Statistical Authority: a non-NSI also responsible for official statistics.	Memobust definition (2014)		Theme: The European Statistical System
NSI	A National Statistical Institute is the leading statistical agency within a national statistical system.	OECD Glossary of Statistical Terms		(1) Theme: Specification of User Needs for Business Statistics; (2) Theme: Dissemination of Business Statistics; (3) Theme: Evaluation of Business Statistics; (4) Theme: Estimation with administrative data; (5) Theme: Different types of surveys; (6) Theme: Response Burden; (7) Theme: The European Statistical System; (8) Theme: Statistical Disclosure Control
NSO	National statistical office - NSI or other office producing official statistics	ESS Handbook for Quality Reports. Eurostat Methodologies and Working Papers. Luxembourg: Office for Official Publications of the European Communities.		Theme: Different types of surveys
NSTR	Nomenclature uniformes des marchandises pour les Statistiques de Transport, Révisé	Eurostat website/CROS portal		Theme: Coding

NUTS	Common regional classification, (called 'Nomenclature of territorial units for statistics' or NUTS).	Council Regulation (EEC) No. 1059/2003.		Theme: Small area estimation
NUTS	The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU.	NUTS classification		Theme: Different types of surveys
Object	Everything that can be perceived or conceived. Examples: output, process, input, staff, software, methodology, document. Context: For an organization, a specific set of objects are relevant like customers, products, processes, input, data, software, staff, etc.	ISO 1179 (2004)	Component or entity	(1) Theme: Methods and Quality; (2) Theme: Quality and Risk Management Models; (3) Theme: Quality of Statistics
Object characteristic	A combination of variables that can be used in the identification of units, but which are not used as object identifier. Often, this concerns variables (or a combination thereof) such as name, address, place of residence, date of birth, profession, education, gender, etc. None of these variables can identify the record by themselves, but the combination can be used as a proxy for a object identifier, if this is missing.	Memobust definition (2014)	Secondary key	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Object identifier	In database technology, the object identifier is the name for a variable or a combination of variables that satisfy the following requirements: - the value of the variable (or the combination of variables) is unique in the table (or data set) and therefore unambiguously defines the record in which it occurs. - the variable (or the combination of variables) is filled in everywhere and therefore cannot be empty. - The combination of variables is minimal: by eliminating one of the variables, the record is no longer unambiguously defined. If related tables refer to the table in which the variable (or combination) of variables occur, this is used to establish a relationship between tables.	Memobust definition (2014)	Primary key; Key	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Objective burden	Burden referring directly to the actual cost of completing questionnaires by respondents; subjective burden reflects their perception	Willeboordse <i>et al.</i> (1997)		Theme: Response Burden
Observation unit	An observation unit represents an identifiable entity about which data can be obtained and for which data is recorded. It should be noted that this may, or may not be, the same as the reporting unit. Remark: It may not be known in advance (e.g. commodities).	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Data Collection: Techniques and Tools
OECD	Organisation for Economic Co-operation and Development	OECD		Theme: The European Statistical System

On-site facility	A facility that has been established on the premises of several NSIs. It is a place where external researchers can be permitted access to potentially disclosive data under contractual agreements which cover the maintenance of confidentiality, and which place strict controls on the uses to which the data can be put. The on-site facility can be seen as a 'safe setting' in which confidential data can be analysed. The on-site facility itself would consist of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. Both the physical and the IT aspects of security would be considered here. The on-site facility also includes administrative and support facilities to external users, and ensures that the agreed conditions for access to the data were complied with.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
Open-ended question	Question that let respondent answer using their own words	Memobust definition (2014)		(1) Theme: Data Collection: Techniques and Tools; (2) Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Opportunities	Step 8 in the OQRM model, where opportunities are analysed if a focus area meets the requirements and more.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
OQRM	Object-oriented Quality and Risk Management	Van Nederpelt (2012)		(1) Theme: Methods and Quality; (2) Theme: Quality and Risk Management Models
Ordinary rounding	See: Conventional rounding.	Glossary on Statistical Disclosure Control (2014)	Conventional rounding	Theme: Statistical disclosure control methods for quantitative tables
OS, PS, TS	Observed Sample - respondent units for the final estimates; Preliminary Sample - quick respondent units; Theoretical Sample-planned sample.	Memobust definition (2014)		Method: Preliminary estimates with model-based methods
Outlier	An outlier is an observation which is not fitted well by a model for the majority of the data. For instance, an outlier may lie in the tail of the statistical distribution or "far away from the centre" of the data.	EDIMBUS Manual		(1) Theme: Statistical Data Editing; (2) Theme: Design of Estimation – Some Practical Issues; (3) Theme: Editing Administrative Data; (2) Theme: Macro-Editing; (4) Method: Outlier Treatment
Outlier in the x-direction	See x-outliers	Memobust definition (2014)		
Outlier in the y-direction	See y-outliers	Memobust definition (2014)		
Outliers	An outlier is a data value that lies in the tail of the statistical distribution of a set of data values.	OECD (2006)		(1) Method: Seasonal adjustment of economic time series; (2) Theme: Issues on Seasonal Adjustment; (3) Theme: Seasonal adjustment – introduction and general description

Output editing	A procedure for tracking suspicious data by checking aggregates or applying statistical methods on all records or on a subset of them.	SDMX (2009)	Macro editing	(1) Theme: Editing for Longitudinal Data; (2) Theme: Selective Editing; (3) Theme: Statistical Data Editing; (4) Theme: Macro-Editing
Outward FATS	'Outward statistics on foreign affiliates' shall mean statistics describing the activity of foreign affiliates abroad controlled by the compiling economy.	Foreign Affiliates Statistics (FATS) recommendation manual, version 2012		Theme: Asymmetry in Statistics – European Register for Multinationals (EGR)
Over-coverage	Over-coverage arises from the presence in the frame of units not belonging to the target population and of units belonging to the target population that appear in the frame more than once.	Eurostat, "Assessment of Quality in Statistics: Glossary",		(1) Theme: Weighting and Estimation; (2) Theme: Quality of Statistics; (3) Theme: Sample selection; (4) Theme: Design of Estimation – Some Practical Issues
Overediting	Editing of data beyond a certain point after which as many errors are introduced as are corrected.	UN/ECE Glossary of Terms on Statistical Data Editing (2007)		(1) Method: Automatic Editing; (2) Method: Manual Editing
Panel	A set of units, which is included several times in a repeated survey according to a specified pattern	Memobust definition (2014)		(1) Method: Little and Su Method; (2) Theme: Imputation for Longitudinal Data; (3) Theme: Design of Estimation – Some Practical Issues
Panel survey	A survey where elements are followed over time.	Memobust definition (2014)	Longitudinal survey	Theme: Weighting and Estimation
Paper and Pencil Interviewing	"Paper" is a method of data collection without the assistance of an interviewer. A questionnaire is sent to respondents, they write in their responses and send it back to the data collection organization.	Memobust definition (2014)	PPI	Theme: Mixed Mode Data Collection – design issues
PAPI	Pencil And Paper Interviewing.	Hacking & Willenborg (2012)		(1) Theme: CATI Allocation; (2) Theme: Data Collection; (3) Theme: Coding
Paradata	Paradata, also termed process data contain information about the primary data collection process (e.g. survey duration, interim status of a case, navigational errors in a survey questionnaire). They can provide a means of additional control over or understanding of the quality of the primary data (the responses to the survey questions).	Memobust definition (2014)		Theme: Data Collection
Parallel mixed mode	Using two or more modes at the same time, e.g. letting the respondent choose his preferred mode.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 2) – Contact strategies
Partial non-response	Also known as item non-response, defines the case of unit may that may respond to the questionnaire incompletely	Memobust definition (2014)	Item non-response	Theme: Data Collection: Techniques and Tools
Pencil And Paper Interviewing.	See PAPI	Memobust definition (2014)		(1) Theme: CATI Allocation; (2) Theme: Data Collection; (3) Theme: Coding

Perceived burden	Burden felt subjectively by the respondent, e.g. connected with the length of the questionnaire, difficulty of the questions, effort required to answer these questions, time spent, etc. or disadvantageous perception of the survey by some respondents, i.e. weak willingness to respond, insufficient awareness of the usefulness of participation, etc.	Willeboordse <i>et al.</i> (1997), Hedlin <i>et al.</i> (2005)	Subjective burden	Theme: Response Burden
Permanent Random Number	A unique random number permanently associated with a unit in a register	Memobust definition (2014)		(1) Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered; (2) Method: Sample co-ordination using simple random sampling with permanent random numbers; (3) Theme: Sample co-ordination
Pilot survey	A survey, usually on a small scale, carried out prior to the main survey, primarily to gain information to improve the efficiency of the main survey. For example, it may be used to test a questionnaire, to ascertain the time taken by field procedure or to determine the most effective size of sampling unit.	Dictionary of Statistical Terms, 5th edition, prepared for the International Statistical Institute by F.H.C. Marriott, Longman Scientific and Technical	Exploratory survey	Theme: Testing the Questionnaire
Planning period	Period in which CATI interviewers are scheduled. This can be a period, of say, 4 weeks starting from the current date, or a calendar month, depending on the allocation variant applied.	Memobust definition (2014)		Theme: CATI Allocation
Poisson sampling design	Sampling design whereby the selection of any unit of the population into the sample is decided independently from the selection of other units.	Memobust definition (2014)		Method: Sample co-ordination using Poisson sampling with permanent random numbers
Population	Population is the total membership or population or "universe" of a defined class of people, objects or events. There are two types of population, viz, target population and survey population. A target population is the population outlined in the survey objects about which information is to be sought and a survey population is the population from which information can be obtained in the survey. The target population is also known as the scope of the survey and the survey population is also known as the coverage of the survey. For administrative records the corresponding populations are: the "target" population as defined by the relevant legislation and regulations, and the actual "client population".	RAMON, Eurostat's metadata server		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design
Positive co-ordination	Maximize the overlap between samples	Memobust definition (2014)		Theme: Sample co-ordination
Positive predictive value	number of correctly linked record pairs divided by the total number of linked record pairs (one minus the false match rate)	Memobust definition (2014)	Precision	Method: Fellegi-Sunter and Jaro Approach to Record Linkage

p-percent rule	A (p,q) rule where q is 100 %, meaning that from general knowledge any respondent can estimate the contribution of another respondent to within 100 % (i.e., knows the value to be nonnegative and less than a certain value which can be up to twice the actual value).	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
PPOS	Planned Preliminary Observed Sample: The respondents of the PTS.	Memobust definition (2014)		Method: Subsampling for Preliminary Estimates
pps	Probability proportional to size	Memobust definition (2014)		Theme: Sample selection
PRB	Perceived Response Burden Study. A survey aimed at recognition and assessment of perceived response burden conducted using some common methodological framework, e.g. PRB Core Questions are a basis to construct relevant target-adjusted questionnaires on burden perceived by respondents in relation to a given statistical survey.	Dale and Haraldsen (2007)		Theme: Response Burden
Precision	number of correctly linked record pairs divided by the total number of linked record pairs	Memobust definition (2014)	Positive predicted value	Method: Fellegi-Sunter and Jaro Approach to Record Linkage
Precision rate	Percentage correct coded texts on the total of coded texts	D'Orazio M. and Macchia S (ROS) (2002)	Accuracy	Theme: Measuring Coding Quality
Predicted values	See anticipated values.	Memobust definition (2014)		Theme: Selective Editing
Preliminary estimates	Estimates based on a preliminary sample	Memobust definition (2014)		(1) Theme: Weighting and Estimation; (2) Theme: Estimation with administrative data
Preliminary sample	Partial sample based on early respondents.	Memobust definition (2014)		Theme: Weighting and Estimation
Preventive measure	Measure to avoid a quality problem.	Memobust definition (2014)		Theme: Quality and Risk Management Models
Price index	The result of a formula in which price changes of various goods and services are weighed together in order get an index for the aggregate.	Memobust definition (2014)		Theme: Manual Integration
Primacy effect	A given response alternative is more likely to be chosen when presented at the beginning rather than at the end of a list of response alternatives.	Memobust definition (2014)	Primacy	Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Primary confidentiality	It concerns tabular cell data, whose dissemination would permit attribute disclosure. The two main reasons for declaring data to be primary confidential are: - too few units in a cell; - dominance of one or two units in a cell. The limits of what constitutes "too few" or "dominance" vary between statistical domains.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Primary data	Data collected on behalf of an NSI and for which the NSI has defined the conceptual and process metadata	Daas and Arends-Toth (2012)		Theme: Collection and Use of Secondary Data
Primary data collection	The gathering of primary data by an NSI	Daas and Arends-Toth (2012)		Theme: Collection and Use of Secondary Data

Primary key	See Object identifier	Memobust definition (2014)		
Primary protection	Protection using disclosure control methods for all cells containing small counts or cases of dominance.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Primary research	Research that uses primary data	Golden (1976)		Theme: Collection and Use of Secondary Data
Primary source	A source containing primary data	Golden (1976)		Theme: Collection and Use of Secondary Data
Primary suppression	This technique can be characterized as withholding all disclosive cells from publication, which means that their value is not shown in the table, but replaced by a symbol such as 'x' to indicate the suppression. According to the definition of disclosive cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or representing cases of dominance have to be primary suppressed.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Principal activity	The principal (or main) activity is identified as the activity which contributes most to the total value added of a unit under consideration. The principal activity so identified does not necessarily account for 50 % or more of the unit's total value added. The classification of principal activity is determined by reference to NACE Rev. 2, first at the highest level of classification and then at more detailed levels ("top-down" method).	Business Register Recommendations Manual (edition 2010), chapter 5, characteristic 2.6, 3.6, 4.7		(1) Theme: Derivation of Statistical Units; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (4) Theme: Statistical Registers and Frames – Survey frames for business surveys; (5) Theme: Statistical Registers and Frames – The statistical units and the business register
Prior-posterior rule	See: (p,q) rule.	Glossary on Statistical Disclosure Control (2014)	(p,q) rule; ambiguity rule	Theme: Statistical disclosure control methods for quantitative tables
Probabilistic record linkage	Linkage method that makes an explicit use of probabilities for deciding when a given pair of records is actually a match or not	Memobust definition (2014)	Weighted matching	Theme: Probabilistic Record Linkage
Probability sample	A sample selected by a method based on the theory of probability (random process), that is, by a method involving knowledge of the likelihood of any unit being selected.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Theme: Sample co-ordination; (3) Theme: Sample selection; (4) Theme: Weighting and Estimation
Probing	Follow-up questions that interviewers can ask in addition to those written on the questionnaire to get more adequate information from respondents.	Memobust definition (2014)		(1) Theme: Data Collection: Techniques and Tools; (2) Theme: Design of data collection (part 1) – Choosing the appropriate data collection method

Processing error	Error in final survey results arising from the faulty implementation of correctly planned implementation methods. Context: In survey data, for example, processing errors may include transcription errors, coding errors, data entry errors and errors of arithmetic in tabulation.	NQAF (2012)		(1) Theme: Quality and Risk Management Models; (2) Theme: Quality of Statistics
PRODCOM	A classification of industrial products	Eurostat website/CROS portal		Theme: Coding
Production	Production is an activity carried out under the control, responsibility and management of an institutional unit that uses inputs of labour, capital and goods and services to produce outputs of goods and services.	ESA (2010)	Output	Theme: Manual Integration
Profiling	Profiling is a method to analyse the legal, operational and accounting structure of an enterprise group in order to establish the statistical units within that group and their links and the most efficient structures for the collection of statistical data	Business Register Recommendations Manual (edition 2010), chapter 19B		Theme: Derivation of Statistical Units
Pro-rata method	A straightforward, widely known benchmarking method that achieves consistency between annual and sub annual time series by multiplying all sub annual periods by correction factors defined by the ratio between an annual value and the sum of all underlying sub annual values. These correction factors are called <i>proportional annual discrepancies</i> .	Memobust definition (2014)		Method: Denton for Benchmarking
Provider load	The effort, in terms of time and cost, required for respondents to provide satisfactory answers to a survey.	Australian Bureau of Statistics, Service Industries Statistics, "Glossary of Terms"; unpublished on paper	Respondent burden	(1) Theme: Testing the Questionnaire; (2) Theme: Response Process
PTS	Preliminary Theoretical Sample. The planned sample draws to obtain the provisional estimates	Memobust definition (2014)		Method: Subsampling for Preliminary Estimates
Punctuality	Time lag between the release date of data and the target date on which they were scheduled for release as announced in an official release calendar.	ESS Handbook for Quality Reports (2009)		(1) Theme: Quality of Statistics; (2) Theme: Overall Design
Punctuality of a log	The period between the delivery time of the log and the planned delivery and time.	Memobust definition (2014)		Theme: Logging
PUPOS	Partially Unplanned Preliminary Observed Sample. A subset of the final sample with a specific follow-up plan. Usually the large units of the final sample	Memobust definition (2014)		Method: Subsampling for Preliminary Estimates
Purposive sample	See non-random sample	Memobust definition (2014)		Method: Subsampling for Preliminary Estimates

Qualitative data	Data describing the attributes or properties that an object possesses.	Economic Commission for Europe of the United Nations (UNECE), "Glossary of Terms on Statistical Data Editing", Conference of European Statisticians Methodological material, Geneva (2000)		Theme: Testing the Questionnaire
Quality	Quality is the degree to which a set of (inherent) characteristics fulfils requirements.	Eurostat's Concepts and Definitions Database (2013), ISO 9000 (2005)		(1) Theme: Quality of Statistics; (2) Theme: Quality and Risk Management Models; (3) Theme: Quality and Risk Management Models
Quality assurance	Part of quality management focused on providing confidence that quality requirements will be fulfilled.	ISO 9000 (2005)		Theme: Overall Design
Quality Circles	Structured employee involvement groups operating in designated work areas that meet regularly to identify work related problems and to suggest solutions or improvements to management.	OECD Glossary of Statistical Terms		Theme: Evaluation of Business Statistics
Quality control	Part of quality management focused on fulfilling quality requirements.	ISO 9000 (2005)		Theme: Overall Design
Quality dimension	Characteristic	Memobust definition (2014)	Criterion, Quality component, Quality aspect, Attribute	(1) Theme: Quality of Statistics; (2) Theme: Methods and Quality
Quality indicator	Variable that represents the quality of data or process.	Memobust definition (2014)		Theme: Quality and Risk Management Models
Quantitative tables	See: Tables of magnitude data	Memobust definition (2014)	Tables of magnitude data	(1) Theme: Statistical disclosure control methods for quantitative tables; (2) Theme: Statistical Disclosure Control
Query edit rule	see Soft edit rule	Memobust definition (2014)	Soft edit rule	
Question format	The way the question is structured. Possible formats: single-choice questions, multi-choice question, table, matrix, partial open-ended question (single or multi choice with other specify), open-ended question.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Data Collection: Techniques and Tools
R^2	Coefficient of determination. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model.	Memobust definition (2014)		Method: Chow-Lin Method for Temporal Disaggregation
Raking	See multiplicative weighting.	Memobust definition (2014)		(1) Method: RAS; (2) Method: Stone
Random error	Antonym of Systematic error	Memobust definition (2014)		(1) Method: Automatic Editing; (2) Method: Deductive Editing; (3) Theme: Statistical Data Editing

Random error	The degree to which the error in the estimate spreads around zero.	Van Nederpelt (2009)	Variance, Precision.	Theme: Quality of Statistics
Random hot deck	A donor record is randomly selected for each recipient record (record with missing information). Usually selection is carried out after grouping units according to some characteristics (e.g. same gender, region, etc.)	Memobust definition (2014)		Method: Statistical Matching Methods
Random rounding	In order to reduce the amount of data loss that occurs with suppression, alternative methods have been investigated to protect sensitive cells in tables of frequencies. Perturbation methods such as random rounding and controlled rounding are examples of such alternatives. In random rounding cell values are rounded, but instead of using standard rounding conventions a random decision is made as to whether they will be rounded up or down. The rounding mechanism can be set up to produce unbiased rounded results.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Random sample	A sample selected by a method based on the theory of probability (random process), that is, by a method involving knowledge of the likelihood of any unit being selected.	SDMX (2009)	Probability sample	(1) Method: Subsampling for Preliminary Estimates; (2) Method: Balanced Sampling for Multi-Way Stratification
random walk model	Model formalization of a random path consisting of a succession of random steps.	Memobust definition (2014)		Method: Small area estimation methods for time series data
Rank hot deck	The donor is chosen in such a way that some measure of distance among percentage points of the empirical distribution is minimized.	Memobust definition (2014)		Method: Statistical Matching Methods
Raw description	Description recorded in an interview or specified by a respondent and that has not been (thoroughly) checked. This may contain various errors, along with insufficient or unnecessary (stop words) information. This is why descriptions are first subjected to several grammatical treatments. This creates a clean or cleansed string, which is used for automatic coding. This string is not intended to be readable, but is utilised as input for the coding program used.	Hacking & Willenborg (2012)		Method: Manual coding
Real-time dataset	dataset showing how estimates change over time providing further information about the dissemination policy, the timing of revisions, the explanation of revision sources, the status of the published data	OECD (2006)	Revision triangle	Theme: Revisions of Economic Official Statistics
Recall	number of correctly linked record pairs divided by the total number of true match record pairs	Memobust definition (2014)	Sensitivity	Method: Fellegi-Sunter and Jaro Approach to Record Linkage
Recency effect	A given response alternative is more likely to be chosen when presented at the end rather than at the beginning of a list of response alternatives.	Memobust definition (2014)	Recency	Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Recipient file	File where one variable (say Z) is completely missing, and that will be imputed making use of the observed Z in the donor file	Memobust definition (2014)		Theme: Statistical Matching

Reconciliation	The series of a system must be reconciled in order to satisfy cross-sectional (contemporaneous) aggregation constraints (see aggregation above)	Dagum and Cholette (2006)	Data reconciliation	(1) Theme: Issues on Seasonal Adjustment; (2) Theme: Seasonal adjustment – introduction and general description
Record linkage	See: matching.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Reference period	The period of time or point in time to which the measured observation is intended to refer.	RAMON, Eurostat's metadata server - Statistical concept		(1) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (4) Theme: Statistical Registers and Frames – Survey frames for business surveys; (5) Theme: Statistical Registers and Frames – Statistical register and survey frame design
References	Step 11 in the OQRM model where Information is collected related to the focus area.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Referential integrity	In a relational database, this is the basic principle that is required for internal consistency of the different tables in that database. This means that a table always has a key if it is referenced by another table in a key field, possibly a foreign key field. Database systems guarantee consistency and ensure that a transaction that violates the consistency cannot be performed.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Refusal rate	The proportion of observation units for which the reporting unit has been successfully contacted, but has refused to give the information sought.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Data Collection: Techniques and Tools
Reg-ARIMA	In the seasonal adjustment context, a hybrid model in which some features of the time series, such as moving holiday, trading day and outlier effects, are modeled with linear regression variables while the remaining features (those of the regression residuals, including trend, cycle and seasonal components) are modelled with a seasonal ARIMA model	US Census Bureau		Method: Seasonal adjustment of economic time series

Register	A written and complete record containing regular entries of items and details on particular set of objects. Administrative registers come from administrative sources and become statistical registers after passing through statistical processing in order to make them fit for statistical purposes (production of register based statistics, frame creation, etc.).	Business Register Recommendations Manual (edition 2010), Glossary		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Statistical Registers and Frames – Survey frames for business surveys; (4) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (5) Theme: Statistical Registers and Frames – The statistical units and the business register
Register	A set of files (paper, electronic, or a combination) containing the assigned data elements and the associated information.	SDMX (2009)		Theme: Sample selection
Register	A systematic collection of unit-level data organized in such a way that updating is possible. Updating is the processing of identifiable information with the purpose of establishing, bringing up to date, correcting or extending the register, i.e. keeping track of any changes in the data describing the units and their attributes. As a rule, a register will contain information on a complete group of units, a target population (e.g. persons, buildings, firms). These units are defined by a precise set of rules (for instance resident population in a country), and the attributes are updated in line with changes undergone by the units.	UN/ECE Glossary of Terms on Statistical Data Editing (2007)		Theme: Collection and Use of Secondary Data
Register unit	Register unit is the unit, entity of the register population with related descriptive information on identification, accessibility and other attributes. Remark: Register unit type – that is the collection of a given type of individual units – and register unit instance – that is a concrete, individual register unit – are distinguished. In the surveying process, data processing and dissemination phases, register units might function as data supplier, data provider or statistical (reporting, observation, analytical, dissemination) units.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design
Regression	A statistical technique for estimating the relationships among variables. In the univariate case only one explanatory variable is used. For the multivariate case, the number of explanatory variables equals two or more.	Memobust definition (2014)		Method: Chow-Lin Method for Temporal Disaggregation
Rejection region	antonym of Acceptance region	Memobust definition (2014)		Method: Manual Editing

Relevance	The degree to which statistical outputs meet current and potential user needs.	ESS Handbook for Quality Reports (2009)	Usability	(1) Theme: Quality of Statistics; (2) Theme: Overall Design
Relevance of log information	The degree to which log information is useful.	Memobust definition (2014)	Usability	Theme: Logging
Reliability	Closeness of the initial estimate to subsequent (revised) estimates	OECD (2006)		(1) Theme: Quality of Statistics; (2) Theme: Revisions of Economic Official Statistics; (3) Theme: Overall Design
Remote access	On-line access to protected microdata.	Memobust definition (2014)		Theme: Statistical Disclosure Control
Remote execution	Submitting scripts on-line for execution on disclosive microdata stored within an institute's protected network. If the results are regarded as safe data, they are sent to the submitter of the script. Otherwise, the submitter is informed that the request cannot be acquiesced. Remote execution may either work through submitting scripts for a particular statistical package such as SAS, SPSS or STATA which runs on the remote server or via a tailor made client system which sits on the user's desk top.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
Repeated survey	A survey which is carried out more than once, often regularly and often designed with some overlap over time between sampled units, taking both accuracy and response burden into account.	Memobust definition (2014)		Theme: Design of Estimation – Some Practical Issues
Repeated survey	A survey which is carried out more than once, often regularly and often designed with some overlap over time between sampled units, taking both accuracy and response burden into account.	Memobust definition (2014)		Theme: Repeated Surveys
Reporting unit	A unit that supplies the data for a given survey instance. The reporting unit is the unit about which data are reported. When, for a specific survey, the book keeping office completes questionnaires for each of the locations of a business, these locations are the reporting units.	Memobust definition (2014)		Theme: Data Collection
Reporting unit	The unit to which the questionnaire is tied and for which the questionnaire is filled in. It may be the observation unit, or it may be a means to reach the observation units.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (4) Theme: Data Collection: Techniques and Tools
representative outlier	represent other population units similar in value to the observed outliers	Memobust definition (2014)		Method: Outlier Treatment

Requirement	Step 3 in the OQRM model where the requirements for the focus area are formulated. Related: norm, standard, prescription, rule, principle, and indicator.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Respondent	Respondents are businesses, authorities, individual persons, etc, from whom data and associated information are collected for use in compiling statistics.	Memobust definition (2014)		Theme: Data Collection
Respondent	The physical person at the data provider who answers the questionnaire.	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools
Respondent	The physical person who answers the questionnaire. This is a person at the data provider.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys
Respondent burden	Burden concerning behavioural and attitudinal attributes of respondents that affect the survey and cannot be changed by the supervisor or organiser of the survey. This concept also includes attitudes towards the survey itself such as the belief in the usefulness of surveys in general.	Hedlin et al. (2005)		Theme: Response Burden
Respondent burden	The effort, in terms of time and cost, required for respondents to provide satisfactory answers to a survey.	SDMX (2009)	Respondent/provider load	(1) Theme: Data Collection; (2) Theme: Data Collection: Techniques and Tools; (3) Theme: Sample selection; (4) Theme: Design of Estimation – Some Practical Issues
Response	The reaction of an individual unit to some form of stimulus. It may be to a drug, as in bioassay, or the reaction to a request for information, as in sample surveys of human beings.	A Dictionary of Statistical Terms, 5th edition, prepared for the International Statistical Institute by F.H.C. Marriott. Published for the International Statistical Institute by Longman Scientific and Technical.		Theme: Response Process
Response	In classical statistics we talk of response when each subject, or experimental units, gives rise to a single (case univariate) or vector (case multivariate) measurement on some relevant variables.	Memobust definition (2014)		Method: Little and Su Method

Response burden	The effort, in terms of time and cost, required for respondents to provide satisfactory answers to a survey.	SDMX (2009)	Statistical burden, Respondent burden	(1) Method: Sample co-ordination using simple random sampling with permanent random numbers; (2) Theme: Sample co-ordination; (3) Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered; (4) Theme: Design of Estimation – Some Practical Issues; (5) Theme: Data Collection: Techniques and Tools; (6) Method: Balanced Sampling for Multi-Way Stratification; (7) Theme: Response Burden
Response process	The result of the interaction between a respondent and a questionnaire	Edwards W.S. & Cantor D. <i>Towards a Response Model in Establishment Surveys</i> In P. P. Biemer, et al., eds., <i>Measurement Error in Surveys</i> , New York: John Wiley & Sons, pp. 211-233		(1) Theme: Testing the Questionnaire; (2) Theme: Response Process
Response rate	The number of observation units for which data have been received, as a proportion of the number of observation units for which data was sought.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Data Collection: Techniques and Tools
Response variable	A variable that is used to define the values in a table. The other kind of variable used to define a table is a spanning variable.	Memobust definition (2014)		Theme: Statistical disclosure control methods for quantitative tables
Responsibilities	Step 2 in the OQRM model, where the distribution of responsibilities of a focus area are determined. Context: There must be at least an owner of each focus area.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Responsive design	There is more than one phase of the data collection and, according to the design, changes between phases are made, based on observed process data, typically indicators of quality and costs.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 2) – Contact strategies; (3) Theme: Overall Design
Restricted (or Residual) Maximum Likelihood (REML)	Particular form of maximum likelihood estimation. It is based on maximizing a likelihood of transformed data not depending on nuisance parameters. In the case of estimation of variance components, the nuisance parameters are the regression coefficients.	Memobust definition (2014)		(1) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot); (2) Method: Small area estimation methods for time series data

Revision	This is a regular procedure in case of unadjusted (raw) data and seasonally adjusted data. Raw data may be revised due to improved information set (in terms of coverage and/or reliability). Revisions of seasonally adjusted data can also take place because of a better estimate of the seasonal pattern due to new information provided by new components. A revision shows the degree of closeness of an initial estimate to a subsequent or final estimate.	ESS Guidelines (2009), ESS Handbook on Quality Reports (2009)		(1) Theme: Overall Design; (2) Theme: Repeated Surveys; (3) Theme: Design of Estimation – Some Practical Issues; (4) Theme: Issues on Seasonal Adjustment
Revision	Difference between revised and preliminary estimate (Lt - Pt)	OECD (2006)		Theme: Revisions of Economic Official Statistics
Revision error	Difference between final and preliminary estimate	Memobust definition (2014)		(1) Theme: Weighting and Estimation; (2) Theme: Estimation with administrative data
Risk analysis	Step 4 in the OQRM model, where possible causes and possible effects with problems with a focus area are analysed. Example: Software errors cause problems with the accuracy of estimates.	Van Nederpelt (2012)		Theme: Quality and Risk Management Models
Risky cells	The cells of a table which are non-publishable due to the risk of statistical disclosure are referred to as risky cells. By definition there are three types of risky cells: small counts, dominance and complementary suppression cells.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Rotating panel	Limiting the length of time in which units stay in the survey panel by dropping a proportion of them after a certain period of time and replacing them with new ones. It is generally done only with the smaller respondents, for whom it is felt that responding to surveys imposes a significant burden. Rotation is designed to keep the sample up to date. It also helps to alleviate the problems caused by sample depletion.	Memobust definition (2014)		Method: Subsampling for Preliminary Estimates
Rotating panel survey	A panel survey where a portion (e.g. 25%/ of elements are replaced regularly.	Memobust definition (2014)		Theme: Weighting and Estimation
Rounding	Rounding belongs to the group of disclosure control methods based on output-perturbation. It is used to protect small counts in tabular data against disclosure. The basic idea behind this disclosure control method is to round each count up or down either deterministically or probabilistically to the nearest integer multiple of a rounding base. The additive nature of the table is generally destroyed by this process. Rounding can also serve as a recoding method for microdata.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Safe data	Microdata or macrodata that have been protected by suitable Statistical Disclosure Control methods.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
Safe setting	An environment such as a microdata lab whereby access to a disclosure dataset can be controlled.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control

Safety interval	The minimal calculated interval that is required for the value of a cell that does not satisfy the primary suppression rule.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Sample	A subset of a frame where elements are selected based on a randomised process with a known probability of selection.	SDMX (2009)		Theme: Sample selection
Sample co-ordination	<i>From topic Sample selection, but long there now</i>	Memobust definition (2014)		Theme: Repeated Surveys
Sample size	The number of observation units which are to be included in the sample.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Method: Subsampling for Preliminary Estimates
Sample size dependent estimator	A sample size dependent estimator is a composite estimator with a subjectively chosen weight for the direct component which depends on true and estimated domain population sizes.	Memobust definition (2014)		Method: Composite Estimators for Small Area Estimation
Sample splitting	statistical method that splits the data into two halves, a regression model is performed on each statistically independent sub-sample	Memobust definition (2014)		Method: Outlier Treatment
Sampling	The process of selecting a number of cases from all the cases in a particular group or universe.	SDMX (2009)		Theme: Sample selection
Sampling design	Design that provides information on the target and final sample sizes, strata definitions and the sample selection methodology.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Method: Subsampling for Preliminary Estimates
Sampling error	An error caused by the fact that only a sample of values is observed and therefore there is a difference between a population value and an estimate.	Memobust definition (2014)		Method: Synthetic Estimators for Small Area Estimation
Sampling error	That part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a sample of values is observed; as distinct from errors due to imperfect selection, bias in response or estimation, errors of observation and recording, etc. The totality of sampling errors in all possible samples of the same size generates the sampling distribution of the statistic which is being used to estimate the parent value.	The International Statistical Institute, The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press, 2003.		(1) Theme: Quality and Risk Management Models; (2) Theme: Quality of Statistics; (3) Theme: Editing During Data Collection
Sampling fraction	The ratio of the sample size to the population size.	SDMX (2009)		Method: Balanced Sampling for Multi-Way Stratification

Sampling frame	A list, map or other specification of the units which define a population to be completely enumerated or sampled.	SDMX (2009)		(1) Theme: Statistical Registers and Frames – Survey frames for business surveys; (2) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (4) Method: Balanced Sampling for Multi-Way Stratification
Sampling strategy	Sampling design and estimation methodology	Memobust definition (2014)		Theme: Weighting and Estimation
Satellite register	Satellite register records a given subpopulation of the business register and fulfill the following conditions: - They are not an integral part of the statistical business register as referred to in the business registers Regulation, but are capable of being linked to it. - They are more limited in scope than the statistical business register, e.g. in terms of NACE, but within that scope they may have more extensive coverage of units and/or variables. - They contain one or more variables that are not found in the statistical business register. Such variables are generally capable of being used for stratification purposes.	Business Register Recommendations Manual (edition 2010), paragraph 20.40 - modified	Associated register	Theme: Statistical Registers and Frames – The statistical units and the business register
SBS	Structural Business Statistics. SBS are statistical surveys covering industry, construction, trade and services. They are conducted in each Member State of the European Union (UE) in order to describe the structure, conduct and performance of businesses across the EU.	Council Regulation (EC, Euratom) No 58/97 of 20 December 1996 concerning structural business statistics amended by Council Regulation (EC, Euratom) No 410/98 of 16 February 1998		Theme: Different types of surveys
Scheduling of interviewers	Production of a planning which indicates which interviewers work on what days and part of day (DPoD) combinations in the planning period.	Memobust definition (2014)		(1) Theme: CATI Allocation; (2) Theme: Data Collection
Scheduling system	IN CATI surveys is the IT module to manage telephone contacts.	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools
SCM	Standard Cost Model – an international method model aimed at reducing administrative burdens in the business environment by adopting a policy based on costs of regulations	ISCM (2003)		Theme: Response Burden

Scope of data suppliers	Scope of the data supplier is the set of entities of the frame population assigned for data reporting from which data can be retrieved for the investigated population (statistical and observation units). Remark: In full scope data collection, the scope of data suppliers corresponds to the frame population. In representative or combined data collection, the scope of data suppliers is only a part of the frame population. It doesn't contain the statistical units of the frame population not selected into the sample.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys
SDC	See: Statistical Disclosure Control	Memobust definition (2014)	Statistical Disclosure Control	(1) Theme: Statistical Disclosure Control; (2) Theme: Statistical disclosure control methods for quantitative tables
Seasonal adjustment	Seasonal adjustment is a statistical technique to remove the effects of seasonal calendar influence operating on series	OECD (2006)	SA	(1) Method: Seasonal adjustment of economic time series; (2) Theme: Issues on Seasonal Adjustment; (3) Theme: Seasonal adjustment – introduction and general description
Seasonal adjustment software	There is a wide range of software and interfaces available to perform seasonal adjustment. For official statistics, the two most commonly used seasonal adjustment methods are X-12-ARIMA (US Census Bureau) and TRAMO-SEATS (Bank of Spain). Recently, Eurostat has released a new software (in which both X-12-ARIMA and TRAMO-SEATS are available), called DEMETRA+	Memobust definition (2014)		Method: Seasonal adjustment of economic time series
Seasonal component	A time series whose values quantify (usually in percents or in the units of data measurement, e.g. dollars) variations in the level of the observed series that recur with the same direction and a similar magnitude at time intervals of length one year. (Length is measured in the calendar units of the observed series--usually quarters or months, sometimes semesters, weeks, or other units.)	US Census Bureau	Seasonality	(1) Method: Seasonal adjustment of economic time series; (2) Theme: Seasonal adjustment – introduction and general description
Secondary data	Data that is collected by others (i.e. not the NSI), used by an NSI for producing statistics and where the NSI has not defined the conceptual or process metadata	Daas and Arends-Toth (2012)		Theme: Collection and Use of Secondary Data
Secondary data collection	The acquisition of secondary data by an NSI	Daas and Arends-Toth (2012)		Theme: Collection and Use of Secondary Data
Secondary Key	See Object characteristic	Memobust definition (2014)		
Secondary research	Research that uses secondary sources	Golden (1976)		Theme: Collection and Use of Secondary Data
Secondary source	A source containing secondary data	Golden (1976)		Theme: Collection and Use of Secondary Data

Secondary suppression	To reach the desired protection for risky cells, it is necessary to suppress additional non- risky cells, which is called secondary suppression or complementary suppression. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the disclosive cells at the highest level of information contained in the released statistics.	Glossary on Statistical Disclosure Control (2014)	Complementary suppression	Theme: Statistical disclosure control methods for quantitative tables
Segmentation effect	It is a characteristics typical of electronic questionnaire and consists in the display of one question per screen thus restricting the view of the questionnaire	Memobust definition (2014)		Theme: Data Collection: Techniques and Tools
Selective editing	An umbrella term for methods that select records which are likely to contain influential errors for interactive editing, on a record-by-record basis.	CBS Methods Series Glossary	Micro-selection, significance editing	(1) Theme: Editing for Longitudinal Data; (2) Theme: Selective Editing
Selective editing	A procedure which targets only some of the micro data items or records for review by prioritizing the manual work and establishing appropriate and efficient process and edit boundaries.	UN/ECE Glossary of Terms on Statistical Data Editing (2007)	Micro-selection	(1) Method: Automatic Editing; (2) Theme: Editing Administrative Data; (3) Theme: Macro-Editing; (4) Theme: Statistical Data Editing
Self-administered mode	The questions in the survey are administered and answered by the respondent without any assistance or help from an interviewer.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Self-Administered Questionnaire	A questionnaire used in Paper and Pencil interviewing.	Memobust definition (2014)		Theme: Mixed Mode Data Collection – design issues
Semantic network	A network(or grph) consisting of words and concepts and semantic relationships between them. Examples of such relationships are synonyms, hypernyms and hyponyms.	Hacking & Willenborg (2012)		Method: Computer-assisted coding
Semi-automatic coding	Synonymous with computer-supported coding and computer-assisted coding.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Computer-assisted coding; (3) Theme: Coding
Sensitivity	number of correctly linked record pairs divided by the total number of true match record pairs. Sensitivity measures the percentage of correctly classified record matches,	Memobust definition (2014)	Recall	Method: Fellegi-Sunter and Jaro Approach to Record Linkage
Sequential mixed mode	Using different modes one after another, maximizing the use of one mode before switching to another.	Memobust definition (2014)		(1) Theme: Data Collection; (2) Theme: Design of data collection (part 2) – Contact strategies
Shrinkage factor	The parameter used in composite estimator formulas to decide about the contribution of the direct and synthetic estimators.	Memobust definition (2014)		Method: Composite Estimators for Small Area Estimation
SIC	Standard Industrial Classification, a classification of industries by type of economic activity created and maintained by the Department of Labour, United States of America and used also e.g. in UK.	US Department of Labour		Theme: Response Burden

Signalling measure	Measure to detect a quality problem.	Memobust definition (2014)		Theme: Quality and Risk Management Models
Significance editing	Synonym of Selective editing.	Memobust definition (2014)	Selective editing	Theme: Selective Editing
Similarity measure	A measure that indicates the extent to which two units are similar. This type of measure (or its complement: the dissimilarity measure) is also used in the multivariate analysis.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Simple random sampling	A sampling design in which the inclusion probability of each unit of the population is given by the sampling fraction.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Method: Subsampling for Preliminary Estimates
Single activity business	A business operating in only one economic activity	Memobust definition (2014)		Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered
Single location business	A business operating from only one geographical location	Memobust definition (2014)		Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered
Skewness	Measure of the asymmetry of a distribution.	Memobust definition (2014)		(1) Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot); (2) Theme: Weighting and Estimation
Small outlier	the Y values are extremely smaller than the other Y values of the “normal” units	Memobust definition (2014)		Method: Outlier Treatment
Smith-Waterman Distance	Distance that uses dynamic programming to find the minimum cost to convert one string into the corresponding string of the compared record; the parameters of this algorithm are the insertions cost, deletions cost and transposition cost	Memobust definition (2014)		Theme: Probabilistic Record Linkage
Snapshot of register	Snapshot of a register is its frozen state on a given date. Remark: Instead of a register, snapshots are used for statistical processing because, unlike register units (that can be updated frequently), population units and their attributes must be unchanged during the data collection and statistical processing.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design
Social desirability bias	Systematic underreporting of something to “fit in” in what the respondent thinks is “normal” or accepted in society. For instance, alcohol consumption is often underreported to avoid embarrassment.	Memobust definition (2014)	Social desirable answers	Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Soft constraint	A constraint that does not have to hold exactly, but approximately.	Memobust definition (2014)		Method: Denton for Benchmarking

Soft edit rule	An edit rule whose failure indicates an error with probability less than 1.	EDIMBUS Manual	Query edit rule	(1) Method: Automatic Editing; (2) Method: Manual Editing; (3) Theme: Statistical Data Editing
Soundex	Indexing technique based on the sound (or pronunciation) of words (and not how they are written), originally only for English, but later developed for Dutch as well.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Computer-assisted coding
Soundex algorithm	Originally a phonetic algorithm to index names based on sound (in English). Later, a similar algorithm was developed for words in the Dutch language. Improvements of the Soundex algorithm for English include Metaphone and Double Metaphone.	Memobust definition (2014)		(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Soundness of methodology	The extent to which the methodology used to compile statistics complies with the relevant international standards, including the professional standards enshrined in the Fundamental Principles for Official Statistics.	SDMX (2009)		(1) Theme: Methods and Quality; (2) Theme: Quality of Statistics
Source	A specific data set, metadata set, database or metadata repository from where data or metadata are available.	SDMX (2009)		Theme: Collection and Use of Secondary Data
Source Data	Characteristics and components of the raw statistical data used for compiling statistical aggregates.	SDMX (2009)		(1) Method: RAS; (2) Method: Stone
Spanning variable	A variable that is used to define the rows, columns etc. of a table. The other kind of variable used to define a table is a response variable.	Memobust definition (2014)		Theme: Statistical disclosure control methods for quantitative tables
Specificity	number of correctly unlinked record pairs divided by the total number of true non-match record pairs. Specificity measures the percentage of correctly classified non-matches.	Memobust definition (2014)		Method: Fellegi-Sunter and Jaro Approach to Record Linkage
Split-off	This event is similar to a break-up, but in this case the original enterprise does survive in a recognisable form, and therefore there is both continuity and survival. There is no death, but one or more new enterprises are created.	Eurostat-OECD Manual on Business Demography Statistics (chapter 4).		Theme: Business Demography
Spreading activation	Method to search in a semantic network.	Hacking & Willenborg (2012)		Method: Computer-assisted coding
State space model	A time-series model that predicts the future state of a system from its previous states probabilistically, via a process model. The state space models describes mathematically how observations of the state of the system are generated via an observation model.	Memobust definition (2014)		(1) Method: Preliminary estimates with model-based methods; (2) Method: Small area estimation methods for time series data
Statistical burden	The burden of the sampled unit to respond to the survey questionnaire.	Memobust definition (2014)		Method: Balanced Sampling for Multi-Way Stratification
Statistical data	Data that are collected and/or generated by statistics in process of statistical observations or statistical data processing.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics

Statistical data collection	Statistical data collection is the operation of statistical data processing aimed at gathering of statistical data and producing the input object data of a statistical survey.	Terminology on Statistical Metadata, Conference of European Statisticians Statistical Standards and Studies, No. 53, UNECE, Geneva 2000,		Theme: Testing the Questionnaire
Statistical data editing	The process of editing a data file for statistical purposes.	Memobust definition (2014)		Theme: Statistical Data Editing
Statistical disclosure control	Statistical Disclosure Control techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released.	Glossary on Statistical Disclosure Control (2014)	Statistical disclosure limitation; SDC;SDL	(1) Theme: Statistical Disclosure Control; (2) Theme: Statistical disclosure control methods for quantitative tables
Statistical edit	A statistical edit is a set of checks based on statistical analysis of respondent data, e.g., the ratio of two fields lies between limits determined by a statistical analysis of that ratio for presumed valid reporters. A statistical edit may incorporate cross-record checks, e.g., the comparison of the value of an item in one record against a frequency distribution for that item for all records. A statistical edit may also use historical data on a firm-by-firm basis in a time series modeling procedure.	Glossary of Terms Used in Statistical Data Editing Located on K-Base, the knowledge base on statistical data editing, UNECE Data Editing Group		Theme: Editing During Data Collection
Statistical matching	Matching records with information from units which do not necessarily have to be the same, but are similar. In terms of intention, this method deals with an entirely different problem than is discussed in this report. This is actually an imputation method. This method is not further discussed in this report for this reason.	Memobust definition (2014)	Synthetic matching	(1) Theme: Object matching; (2) Method: Object Identifier Matching; (3) Method: Unweighted Matching; (4) Method: Weighted Matching
Statistical measure	A summary (means, mode, total, index, etc.) of the individual quantitative variable values for the statistical units in a specific group (study domains).	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Statistical output	Results from a statistical process to be accessed by the final users. Context: Can take the form of aggregate statistics, analysis, and microdata releases and can include different forms of media.	NQAF (2012)		Theme: Methods and Quality
Statistical output	Results from a statistical process to be accessed by the final users..	NQAF (2012)		Theme: Quality of Statistics
Statistical register	Statistical register is a continuously updated set of objects for a given population containing information on identification, accessibility of population units and other attributes, supporting the surveying process of the population. The register contains the current and historical statuses of the population and the causes, effects and sources of alterations in the population. Register data of population units are stored in a structured database	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys

Statistical register	A regularly updated list of units and their characteristics to be used for statistical purposes.	SDMX (2009)		Theme: Collection and Use of Secondary Data
Statistical source	A source containing information collected and maintained for statistical purposes. It contains statistical units and statistical variables	Parallel definition to administrative source		Theme: Collection and Use of Secondary Data
Statistical unit	Statistical units are defined on the basis of three criteria: Legal, accounting or organizational criteria; Geographical criteria; Activity criteria. The relationship between different types of statistical units can be summarized in the following way: Units with one or more activities and one or more locations; Enterprise; Institutional unit; Units with one or more activities and a single location; Local unit; Units with one single activity and one or more locations; KAU; UHP; Units with one single activity and one single location; Local KAU; Local UHP. The Council Regulation (EEC), No 696/93 of 15 March 1993 on statistical units for the observation and analysis of the production system in the Community lays down a list of eight (types of) statistical units: The enterprise; The institutional unit; The enterprise group; The kind-of-activity unit (KAU); The unit of homogeneous production (UHP); The local unit; The local kind-of-activity unit (local KAU); The local unit of homogeneous production (local UHP).	Council Regulation (EEC), No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section.		(1) Theme: Statistical Registers and Frames – The statistical units and the business register; (2) Theme: Derivation of Statistical Units
Step problem	The phenomenon of a large gap between the last sub annual period of one annual period and the first sub annual period of the next annual period. (for instance: a large gap between December and Januar). Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.	Memobust definition (2014)		Method: Denton for Benchmarking
Stochastic imputation	In stochastic imputation the imputed value contains a random component. Repetition of the imputation leads to a different result.	EDIMBUS Manual		(1) Theme: Donor Imputation; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Model-based Imputation
Stochastic regression imputation	Model based imputation method: : imputes the missing value with a value obtained as the sum of the predicted value by the regression model being considered and a random error term	Memobust definition (2014)		Method: Statistical Matching Methods
Stock Variable	A stock variable is measured at one specific time, and represents a quantity existing at that point in time. See also flow variable	Memobust definition (2014)		Theme: Macro Integration
Stop word	Word in a description that does not contain any information or contains too little information, because it occurs too frequently. A stop word can therefore be deleted by an automatic coding system.	Hacking & Willenborg (2012)		Theme: Coding

Stratification	A sampling procedure in which the population is divided into homogeneous subgroups or strata and the selection of samples is done independently in each stratum.	SDMX (2009)		(1) Method: Sample co-ordination using simple random sampling with permanent random numbers; (2) Theme: Sample selection
Stratified simple random sampling	A sampling design in which the population is divided into homogeneous subgroups or strata and the selection of samples is done independently in each stratum.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Method: Subsampling for Preliminary Estimates
Structural zero (cell)	A zero in a table cell corresponding to a situation where there can be no population elements, because this is impossible, on logical or as a matter of fact or principle. (For instance: a pregnant man.)	Memobust definition (2014)		Theme: Statistical disclosure control methods for quantitative tables
STS	Short-Term Statistics. STS are statistical surveys conducted in each Member State of the UE with a monthly or quarterly frequency. Output data (indicators) deliver information on supply, demand, factors of production and prices in four main domains: industry, construction, retail trade and other services.	Regulation EC No 1165/98, amended by Regulation EC No 1158/2005 and the regulations implementing and amending these two instruments		(1) Theme: Different types of surveys; (2) Theme: The European Statistical System; (3) Theme: Estimation with administrative data
Study domains	A segment of the population for which separate statistics are needed.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Subadditivity	One of the properties of the (n,k) rule or (p,q) rule that assists in the search for complementary cells. The property means that the sensitivity of a union of disjoint cells cannot be greater than the sum of the cells' individual sensitivities (triangle inequality). Subadditivity is an important property because it means that aggregates of cells that are not sensitive are not sensitive either and do not need to be tested.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Subpopulation	Subpopulation is a subset of a population. Remark: Subpopulation refers to populations that require different handling in the statistical working process. Subpopulations are usually specified to understand the distinguishing characteristics of these populations	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design
Supplier manager	Person responsible to acquire and manage products and or resources to needed to run a business	Memobust definition (2014)		Theme: Collection and Use of Secondary Data
Supply use tables	An accounting framework in which supply and use of goods and services and the generation of value added is described, detailed to commodities and industries. It is the ideal framework for making estimates of gross domestic product (GDP)	Memobust definition (2014)		Theme: Manual Integration

Survey	Survey is an investigation on the characteristics of a given population by means of collecting data and estimating their characteristics through the systematic use of statistical methodology. Remark: Included are: - censuses, which attempt to collect data from all members of a population; - sample surveys, in which data are collected from a (usually random) sample of population members. Surveys can be unique in time or repeated with regular or irregular periodicity. A single wave of a repeated survey is called survey instance. A wider definition under which the term survey covers any activity that collects or acquires statistical data (including censuses, sample surveys, the collection of data from administrative records and derived statistical activities) has also been proposed. (see Statistics Canada, "Statistics Canada Quality Guidelines", 4th edition, October 2003, page 7, available at http://www.statcan.ca:8096/bsolc/english/bsolc?catno=12-539-X&CHROPG=1).	RAMON, Eurostat's metadata server – Statistical concepts		Theme: Statistical Registers and Frames – The populations, frames and units of business surveys
Survey	A investigation about the characteristics of a given population by means of collecting data from a sample of that population and estimating their characteristics through the systematic use of statistical methodology.	SDMX (2009)		Theme: Sample selection
Survey (1)	Survey is an investigation on the characteristics of a given population by means of collecting data and estimating their characteristics through the systematic use of statistical methodology. Remark: Included are: censuses, which attempt to collect data from all members of a population; - sample surveys, in which data are collected from a (usually random) sample of population members. Surveys can be unique in time or repeated with regular or irregular periodicity. A single wave of a repeated survey is called survey instance.	RAMON, Eurostat's metadata server – Statistical concepts		Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames
Survey (2)	A wider definition under which the term survey covers any activity that collects or acquires statistical data (including censuses, sample surveys, the collection of data from administrative records and derived statistical activities) has also been proposed. (see Statistics Canada, "Statistics Canada Quality Guidelines", 4th edition, October 2003, page 7, available at http://www.statcan.ca:8096/bsolc/english/bsolc?catno=12-539-X&CHROPG=1)	Memobust definition (2014)		Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames

Survey feedback	Information obtained from a survey used to update the Register	Memobust definition (2014)		(1) Theme: Sample co-ordination; (2) Theme: Repeated Surveys
Survey frame	Survey frame is the set of survey population units together with their attributes referring to a given reference period. The frame contains the identification, contact, classification attributes of the frame units for a given reference period.	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys; (3) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (4) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (5) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames
Survey instance	Survey instance is a particular survey and reference period in which data are collected from respondents	RAMON, Eurostat's metadata server - UN metadata terminology		(1) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (4) Theme: Statistical Registers and Frames – Survey frames for business surveys
Survey population	Survey population is the population for which information during the survey process can be obtained. Remark: Concurrence or difference of survey and target populations is measured by coverage	Handbook on the design and implementation of business surveys		(1) Theme: Asymmetry in Statistics – European Register for Multinationals (EGR); (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys
Surveying department	The surveying department is the unit of the statistical office that responsible for the data collection phase of the survey	Memobust definition (2014)		(1) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (2) Theme: Statistical Registers and Frames – Survey frames for business surveys

Survival rate	The survival rate of newly born enterprises in a given reference period is the number of enterprises that were born in year t-i (i=1,...,n) and survived to year t as a percentage of all enterprises born in year t-i.	Memobust definition (2014)		Theme: Business Demography
Surviving enterprise	In the Business Demography context, survival occurs if an enterprise is active in terms of employment and/or turnover in the year of birth and the following year(s). Two types of survival can be distinguished: 1) An enterprise born in year t-1 is considered to have survived in year t if it is active in terms of turnover and/or employment in any part of year t (= survival without changes). 2) An enterprise is also considered to have survived if the linked legal unit(s) have ceased to be active, but their activity has been taken over by a new legal unit set up specifically to take over the factors of production of that enterprise (= survival by take-over).	Eurostat-OECD Manual on Business Demography Statistics		Theme: Business Demography
Swapping (or switching)	Swapping (or switching) involves selecting a sample of the records, finding a match in the data base on a set of predetermined variables and swapping all or some of the other variables between the matched records. Swapping (or switching) was illustrated as part of the confidentiality edit for tables of frequency data.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Synonym	Word or concept with the same meaning as another word, possibly in a special context.	Hacking & Willenborg (2012)		(1) Method: Computer-assisted coding; (2) Theme: Coding
Synthetic estimator	An indirect estimator based on the assumption that small areas have the same characteristics as a large area and a reliable direct estimator for the large area is used in the estimation process for small areas.	Memobust definition (2014)		Method: Synthetic Estimators for Small Area Estimation
Synthetic matching	See: Statistical matching	Memobust definition (2014)		
Systematic error	(1) An error reported consistently over time and/or between responding units. Or (2) a type of error for which the error mechanism and the imputation procedure are known.	UN/ECE Glossary of Terms on Statistical Data Editing, EDIMBUS Manual.		(1) Method: Automatic Editing; (2) Method: Deductive Editing; (3) Theme: Editing Administrative Data; (4) Theme: Statistical Data Editing
Systematic error	The systematic deviation of the estimate from the true value.	Van Nederpelt (2009)	Bias, purity	Theme: Quality of Statistics
Table	A special form of aggregate data, where the information is divided into cells, each corresponding to a group of individual entities	Memobust definition (2014)		Theme: Statistical Disclosure Control
Table redesign	See: table restructuring	Memobust definition (2014)	Table restructuring	Theme: Statistical disclosure control methods for quantitative tables
Table restructuring	A technique to produce safe tables by combining rows or columns	Memobust definition (2014)	Table redesign	Theme: Statistical disclosure control methods for quantitative tables

Tables of frequency (count) data	These tables present the number of units of analysis in a cell. When data are from a sample, the cells may contain weighted counts, where weights are used to bring sample results to the population levels. Frequencies may also be represented as percentages.	Glossary on Statistical Disclosure Control (2014)	Frequency tables	(1) Theme: Statistical Disclosure Control; (2) Theme: Statistical disclosure control methods for quantitative tables
Tables of magnitude data	Tables of magnitude data present the aggregate of a "quantity of interest" over all units of analysis in the cell. When data are from a sample, the cells may contain weighted aggregates, where quantities are multiplied by units' weights to bring sample results up to population levels. The data may be presented as averages by dividing the aggregates by the number of units in their cells.	Glossary on Statistical Disclosure Control (2014)	Quantitative tables	(1) Theme: Statistical Disclosure Control; (2) Theme: Statistical disclosure control methods for quantitative tables
Tabular data	Aggregate information on entities presented in tables.	Glossary on Statistical Disclosure Control (2014)	Macrodata	Theme: Statistical disclosure control methods for quantitative tables
Take-over	This event can be seen as the opposite of a split-off. Enterprises taken over are not considered to be deaths. In this case, one of the original enterprise does survive in a recognisable form, and therefore there is both continuity and survival. The remaining original enterprises are closed.	Eurostat-OECD Manual on Business Demography Statistics (chapter 4).		Theme: Business Demography
Target population	Target population is the set of units about which information is wanted and estimates are required. Remark: We differentiate the ideal and the intended target population. The ideal target population is the user demand, the intended target population is the realisable population of the survey.	CODED – Statistical concept - modified		(1) Theme: Statistical Registers and Frames – Main module; Theme: Statistical Registers and Frames – Quality of statistical registers and frames; (2) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (3) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (4) Theme: Statistical Registers and Frames – Survey frames for business surveys; (5) Theme: Statistical Registers and Frames – Statistical register and survey frame design; (6) Theme: Asymmetry in Statistics – European Register for Multinationals (EGR)
Target variable	A variable that is observed or derived and that measures an aspect of a phenomenon of interest during a survey; a goal of the survey will be to estimate population parameters for such a variable.	CBS Methods Series Glossary		(1) Theme: Donor Imputation; (2) Theme: Imputation; (3) Theme: Imputation for Longitudinal Data; (4) Theme: Model-based Imputation

t-ARGUS	t-Argus is a specialized software tool for the protection of tabular data. t-Argus is used to produce safe tables. t-Argus uses the same two main techniques as μ -Argus: global recoding and local suppression. For t-Argus the latter consists of suppression of cells in a table.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical Disclosure Control
TDE	Telephone/Touchtone Data Entry is a data entry mode in which a telephone is used by the respondent to communicate his/her answers. It is a form of self-administered telephone survey that does not require interviewer assistance.	Memobust definition (2014)		Theme: Mixed Mode Data Collection – design issues
Temporal constraint	Constraints in the same time-series for different periods	Memobust definition (2014)		Method: Denton for Benchmarking
Temporal Disaggregation	Deriving sub annual data (for instance quarterly data) from annual data, by using indicators of the sub annual data (i.e. related time series), see disaggregation. Annual and sub annual are used in a broad sense here. It can be any combination of two periods with a difference frequency, such that one annual period covers a whole number of sub annual periods.	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Theme: Macro Integration
Test variable	A component of an edit rule that defines, for a given edit group, the expression (in terms of one or more observed variables) that is to be evaluated with respect to the acceptance regions for edit groups.	Norberg (2011)		(1) Method: Manual Editing; (2) Theme: Editing for Longitudinal Data
TF-IDF Distance	Distance that is used to match strings in a document. It assigns high weights to frequent tokens in the document and low weights to tokens that are also frequent in other documents	Memobust definition (2014)		Theme: Probabilistic Record Linkage
Threshold rule	Usually, with the threshold rule, a cell in a table of frequencies is defined to be sensitive if the number of respondents is less than some specified number. Some agencies require at least five respondents in a cell, others require three. When thresholds are not respected, an agency may restructure tables and combine categories or use cell suppression, rounding or the confidentiality edit, or provide other additional protection in order to satisfy the rule.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Time series	A set of ordered observations on a quantitative characteristic of an individual or collective phenomenon taken at different points of time.	SDMX (2009)		(1) Method: Chow-Lin Method for Temporal Disaggregation; (2) Method: Denton for Benchmarking; (3) Theme: Macro Integration
Time series	A sequence of measurements of an economic (or other) variable made at approximately equally spaced times. It is important that the definition of the variable and the method used to measure it be consistent over time	US Census Bureau		Method: Seasonal adjustment of economic time series

Timeliness	The length of time between the event or phenomenon the statistical outputs describe and their availability.	ESS Handbook for Quality Reports (2009)		(1) Theme: Quality of Statistics; (2) Theme: Overall Design
Timeliness	The lapse of time between the end of a reference period and availability of the data.	SDMX (2009)		(1) Method: Subsampling for Preliminary Estimates; (2) Method: Preliminary estimates with design-based methods
Top-of-the-head responses	The respondent is feeling stressed and pressured to give an quick answer and therefore picks the first response category presented to them.	Memobust definition (2014)		Theme: Design of data collection (part 1) – Choosing the appropriate data collection method
Total survey error	The accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data.	Biemer (2010)		Theme: Quality of Statistics
Training set	A corpus where the codes linked to the descriptions are verified. The codes originate from a classification. A training set is used in the coding methods that are based on supervised classification.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Computer-assisted coding
Transversal sampling design	Sampling design of one of the surveys, at one sampling occasion.	Memobust definition (2014)		Method: Sample co-ordination using Poisson sampling with permanent random numbers
Trend-cycle	The trend is the underlying long-term movement lasting many years. The cycle, also called business-cycle, is a quasi-periodic oscillation lasting for more than a year around the long-term trend. It is characterized by alternating periods of expansion and contraction. The trend and the cycle are difficult to estimate separately and thus are considered and analysed as a whole as the trend-cycle	Statistics Canada (2009)	TC	(1) Method: Seasonal adjustment of economic time series; (2) Theme: Seasonal adjustment – introduction and general description
Trigram	String consisting of three consecutive characters. They are used in fuzzy string matching. The more trigrams two strings have in common, compared to the trigrams they have not in common, the more similar they are.	Hacking & Willenborg (2012)		(1) Method: Automatic coding based on pre-coded datasets; (2) Method: Automatic coding based on semantic networks; (3) Method: Computer-assisted coding
Trimmed least absolute value	robust statistical method that attempts to minimise the sum of absolute deviation (residuals) over a subset of k points which yields the lowest sum of absolute residuals ($k < n$)	Memobust definition (2014)		Method: Outlier Treatment
Trimmed least square	robust statistical method that attempts to minimise the sum of squared residuals over a subset, k points which yields the lowest sum of squared residuals ($k < n$)	Memobust definition (2014)		Method: Outlier Treatment
True value	The actual population value that would be obtained with perfect measuring instruments and without committing any error of any type, both in collecting the primary data and in carrying out mathematical operations.	Eurostat's Concepts and Definitions Database (2013)		Theme: Quality of Statistics
Type I error	See: Mismatch	Memobust definition (2014)		
Type II error	See: Missed match	Memobust definition (2014)		

Unbiased	Estimator whose bias is zero.	Memobust definition (2014)		Method: Generalised regression estimator
Unbiasedness	An estimator is said to be unbiased if the bias (difference between its mathematical expectation and the true value it estimates) is zero.	The International Statistical Institute, "The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press (2003).		Theme: Small area estimation
Under-coverage	There are target population units that are not accessible via the frame.	ESS Handbook on Quality Reports (2009)		Theme: Design of Estimation – Some Practical Issues
Under-coverage	Under-coverage results from the omission from the frame of units belonging to the target population.	OECD Glossary		Theme: Weighting and Estimation
Under-coverage	Failure to include required units in the frame, which results in the absence of information for those units.	SDMX (2009)		(1) Theme: Quality of Statistics; (2) Theme: Sample selection
Unequal probability sampling	A sampling design in which the inclusion probability may be different for each unit of the population.	SDMX (2009)		(1) Method: Balanced Sampling for Multi-Way Stratification; (2) Method: Subsampling for Preliminary Estimates
Unit	Units refer to entities, respondents to a survey or things used for the purpose of calculation or measurement. Their statistics are collected, tabulated and published. They include, among others, businesses, government institutions, individual organisations, institutions, persons, groups, geographical areas and events. They form the population from which data can be collected or upon which observations can be made. Remark: In this handbook chapter the unit can belong to the population, frame, register. The type of unit can be statistical unit, collection unit, reporting unit, observation unit, analytical unit, legal unit.	RAMON, Eurostat's metadata server		Theme: Statistical Registers and Frames – The populations, frames and units of business surveys
Unit non-response	The event that no data are obtained from a unit that was supposed to be observed.	CBS Methods Series Glossary		(1) Theme: Imputation; (2) Theme: Imputation for Longitudinal Data; (3) Theme: Quality of Statistics; (4) Theme: Data Collection: Techniques and Tools

Unit of homogeneous production	The unit of homogeneous production (UHP) is characterised by a single activity which is identified by its homogeneous inputs, production process and outputs. The products which constitute the inputs and outputs are themselves distinguished by their physical characteristics and the extent to which they have been processed as well (as) by the production technique used, by reference to a product classification. The unit of homogeneous production may correspond to an institutional unit or a part thereof; on the other hand, it can never belong to two different institutional units	Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community, Annex Section III E; SBS Regulation No 58/97, variable (12 11 0)		(1) Theme: Statistical Registers and Frames – Building and maintaining statistical registers to support business surveys; (2) Theme: Statistical Registers and Frames – The populations, frames and units of business surveys; (3) Theme: Statistical Registers and Frames – The statistical units and the business register
Unit of measurement error	An error that occurs when respondents report values that are consistently too high or too low by a constant factor.	Memobust definition (2014)		(1) Method: Deductive Editing; (2) Theme: Statistical Data Editing
Unit response rate	The ratio of the number of units for which data for some variables have been collected to the total number of units from which data are to be collected. It can indirectly measure response burden.	Eurostat (2009)		Theme: Response Burden
Unit types	A Business Register generally consists of several unit types, for example the enterprise unit, the kind of activity unit	Memobust definition (2014)		Method: Assigning random numbers when co-ordination of surveys based on different unit types is considered
Unity measure error	An error that occurs when respondents report the value of a variable in a wrong unity measure.	EDIMBUS Manual		Theme: Editing for Longitudinal Data
UPOS	Unplanned Preliminary Observed Sample. Early respondents used for provisional estimates. No specific follow-up has been planned	Memobust definition (2014)		Method: Subsampling for Preliminary Estimates
Upper bound	The highest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
User	A person or an organization that employs or applies statistical information, within or outside an NSI. Institutional users can also be stakeholders in the specification of which statistical information is produced.	Memobust definition (2014)		Theme: Specification of User Needs for Business Statistics
User needs	User needs refer to the data and metadata requirements of persons or organizations to meet a particular use or set of uses. Such needs may be specified in terms of the quality dimensions promulgated by international organizations or national agencies.	OECD Glossary of Statistical Terms		Theme: Specification of User Needs for Business Statistics
Value index	The ratio of transaction in current prices of the present and previous period	Memobust definition (2014)		Theme: Manual Integration
Variance	Expectation of the square difference between the estimates and its means value.	ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys		Method: Generalised regression estimator

Variance	The variance is the mean square deviation of the variable around the average value. It reflects the dispersion of the empirical values around its mean.	Eurostat's Concepts and Definitions Database (2013)	Precision, random error.	Theme: Quality of Statistics
Variance	Expectation of the square difference between the estimates and its means value over the possible values.	See also Glossary of the Handbook on precision requirement and variance estimation for household surveys.		Theme: Weighting and Estimation
Vertical aggregation	Vertical aggregation: aggregation by sector or branch	European Communities (2001)		Theme: Seasonal adjustment – introduction and general description
Volume index	The result of a formula in which volume changes of various goods and services are weighed together in order get an index for the aggregate.	Memobust definition (2014)		Theme: Manual Integration
VVK	The Dutch Association of Chambers of Commerce. VVK means Vereniging Van Kamers van Koophandel	Hacking & Willenborg (2012)		Method: Automatic coding based on semantic networks
Waiver approach	Instead of suppressing tabular data, some agencies ask respondents for permission to publish cells even though doing so may cause these respondents' sensitive information to be estimated accurately. This is referred to as the waiver approach. Waivers are signed records of the respondents' granting permission to publish such cells. This method is most useful with small surveys or sets of tables involving only a few cases of dominance, where only a few waivers are needed. Of course, respondents must believe that their data are not particularly sensitive before they will sign waivers	Glossary on Statistical Disclosure Control (2014)		Theme: Statistical disclosure control methods for quantitative tables
Web forms	A form on a website that enables visitors to communicate with the host by filling in the fields and submitting the information. Information received via a form can be received by email and processed by other specific software.	OECD, 2004, Promise and Problems of E-Democracy: Challenges of Online Citizen Engagement, OECD, Paris, Annex 1: Commonly used E-Engagement Terms.		(1) Theme: Questionnaire Design; (2) Theme: Editing During Data Collection; (3) Theme: Testing the Questionnaire
Web Survey	A form of CASI in which a computer administers a questionnaire on a web site. In on-line surveys the questions are viewed and answered using a standard web browser on a PC, laptop or tablet. In an off-line survey the electronic questionnaire is downloaded and completed off-line. The responses are transferred through the internet to the server.	Memobust definition (2014)	CAWI	Theme: Mixed Mode Data Collection – design issues
Weight	The importance of an object in relation to a set of objects to which it belongs.	SDMX (2009)	Matching weight	(1) Method: Denton for Benchmarking; (2) Theme: Weighting and Estimation
Weight trimming	reduction of weights larger than some value	Memobust definition (2014)		Method: Outlier Treatment

Weighted Least Square	The parameter are obtained as those value the minimize a weighted square of distance between predicted and observed.	Memobust definition (2014)		Method: EBLUP Unit level for Small Area Estimation
Weighted Least Square (WLS)	Parameter estimates are obtained as the values maximizing the weighted square of distance between predicted and observed values.	Memobust definition (2014)		Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)
Weighting	The act of assigning weights to survey respondents, which are then used to obtain estimates of population parameters by calculating weighted sums of observed values.	CBS Methods Series Glossary		(1) Theme: Imputation; (2) Theme: Imputation for Longitudinal Data
Winsorization	modifying values in the sample so that the estimator becomes robust and isn't affected by large residuals	Memobust definition (2014)		Method: Outlier Treatment
Working/trading day effects	These are systematic effects in monthly times series related to changes in the day-of-week composition of each month and, in some cases, also to changes in the length of February. For flow series (monthly accumulations of daily activity e.g. monthly sales), the increases or decreases from average day-of-week activity associated with the days that occur five times in the month in a given year are important. For flow series, the length of February can have an impact. For stock series, such as end-of-month inventories, the extent to which inventories tend to rise or fall on the day of measurement (e.g. the last day of the month) can have an impact that is different from year to year. Attempts to measure analogous effects in quarterly series are seldom successful. A series of estimated trading day effects defines a trading day component for the time series	US Census Bureau		Method: Seasonal adjustment of economic time series
X-outliers	the X values of a few sample units that are very distant from the X-values of the other sample units.	Memobust definition (2014)	Outlier in the x-direction	Method: Outlier Treatment
Y-outliers	the Y values of a few sample units that are very distant from the Y-values of the other sample units.	Memobust definition (2014)	Outlier in the y-direction	Method: Outlier Treatment
T-ARGUS	Software program designed to protect statistical tables.	Argus (2013)		Theme: Logging

Source	Website
Argus (2013)	http://neon.vb.cbs.nl/casc/glossary.htm . Retrieved 25 October 2013
ESA (2010)	Eurostat (2010), European system of accounts (forthcoming).
ESS Regulation No 223 (2009)	http://www.ons.gov.uk/ons/about-ons/what-we-do/relationships-abroad/european-statistical-system--ess-/index.html . Regulation No 223/2009
Eurostat's Concepts and Definitions Database (2013)	http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GL_OSSARY&StrNom=CODED2&StrLanguageCode=EN . Retrieved 25 October 2013
Glossary on Statistical Disclosure Control (2014)	http://neon.vb.cbs.nl/casc/index.htm
Glossary, Adapting new technologies to census operations (2001)	Adapting new technologies to census operations, Arij Dekker, Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects, Statistics Division, Department of Economic and Social Affairs, United Nations Secretariat New York, 7-10 August 2001, Glossary.
Golden (1976)	Golden, M.P. (1976), <i>The research experience</i> . F.E. Peacock Publishers Inc., Itasca, Illinois, USA
Hacking & Willenborg (2012)	Hacking, W. and L. Willenborg (2012), <i>Coding – interpreting short descriptions using a classification</i> . Translation of a contribution to the CBS Methods Series, Report, Statistics Netherlands, The Hague and Heerlen.
Hacking, W. and L. Willenborg (2012) Memobust definition (2014)	Hacking, W. and L. Willenborg (2012), <i>Coding – interpreting short descriptions using a classification</i> . Translation of a contribution to the CBS Methods Series, Report, Statistics Netherlands, The Hague and Heerlen. http://www.cros-portal.eu/content/memobust
NACE Rev.2	http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-015/EN/KS-RA-07-015-EN.PDF
NUTS classification	http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction
OECD Glossary	http://stats.oecd.org/glossary/detail.asp?ID=5069
RAMON, Eurostat's metadata server	
SDMX (2009)	http://sdmx.org

SDMX (2009)	http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf
US Census Bureau	US Census Bureau http://www.census.gov/srd/www/x13as/glossary.html
US Department of Labour	https://www.osha.gov/pls/imis/sic_manual.html
Wikipedia Cluster Sampling	http://en.wikipedia.org/wiki/Cluster_sampling
Wikipedia Multistage Sampling	http://en.wikipedia.org/wiki/Multistage_sampling



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Sample Selection – Main Module

Contents

General section	3
1. Summary	3
2. General description.....	3
2.1 Probability and non-probability sampling	3
2.2 Stratified simple random sampling	4
2.3 Probability proportional to size (pps) sampling	5
2.4 Cut-off sampling.....	6
2.5 Cluster or multistage sampling	6
2.6 Systematic sampling.....	6
2.7 Balanced sampling.....	7
2.8 The use of panels and rotation groups	7
3. Design issues	7
4. Available software tools	7
5. Decision tree of methods	8
6. Glossary	8
7. References	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

Sample selection in business statistics can be challenging because of several reasons. The population is often skewed, new businesses are created or they go out of business, and businesses may be related to each other in different ways. The use of a stratified simple random sampling design can enable researchers to draw inferences about specific subgroups that may be lost in a more generalised random sample, but this requires the selection of relevant stratification variables. An important option here, which is commonly used for business surveys whenever element size varies greatly, is probability proportional to size (pps) sampling, often in combination with cut-off sampling. This method can improve accuracy for a given sample size by concentrating the sample on large elements that have the greatest impact on population estimates. An alternative to stratified simple random sampling is systematic sampling. Cluster or multistage sampling is motivated by the need for practical, economical and sometimes administrative efficiency. The use of fixed panels will produce very efficient estimates of periodic change. In most periodic surveys sample rotation is used in order to reduce response burden.

2. General description

Most samples for business surveys are sampled from lists (list frames) or registers. Developing the associated sample designs can be challenging because business populations can have the following characteristics (Sigman and Monsour, 1995):

- *Skewness*. A small number of businesses account for a large proportion of the population total.
- *Dynamic membership*. Businesses are created, go out of business, change their type or level of activity, or change their identity.
- *Inter-business relationships*. Businesses may be related to each other in such a way that they are owned by the same legal entity, they employ the same accountant, or their activities are combined in a common set of financial records.

2.1 Probability and non-probability sampling

In a **probability sampling** scheme every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Data collected by statistical offices are often not consistent with the accounting rules. This happens, for example, because economic data are frequently collected by different methods, using different sample surveys and different data processing methods and because of estimation error in case of missing data.

Probability sampling includes stratified simple random, probability proportional to size, systematic and balanced sampling. These various ways of probability sampling have two things in common:

- Every element in the population of interest has a known non-zero probability of being sampled.

- They involve random selection at some points.

Non-probability sampling can be divided into two categories. The first category is when we have a situation where some elements of the population have *no* chance of selection (these are sometimes referred to as ‘out of coverage’, ‘undercovered’, ‘take no’). Cut-off sampling, see Section 2.4, is an example of such a scheme, which is commonly applied in business statistics.

The second category is where the probability of selection cannot be accurately determined. This involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is non-random, non-probability sampling does not depend on the rationale of probability theory, thus it does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population. Non-probability also sampling includes accidental (haphazard, convenience) and purposive sampling. These methods are usually not applied in business statistics and will therefore not be discussed further.

In addition to cut-off sampling, a non-probability sampling scheme may also be applied when the goal is to find preliminary estimates (see the module “Sample Selection – Subsampling for Preliminary Estimates”), which merely involves the use of quick respondent units.

2.2 *Stratified simple random sampling*

Whenever the population embraces a number of distinct categories, the frame can be organised by these categories into separate ‘strata’. Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. Questions that need to be answered with this design include:

- How should strata be constructed?
- How should the sample be allocated to strata?

There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalised random sample. Methods for stratum construction in the case of one characteristic of interest and one continuous stratification variable are described by Dalenius and Hodges (1959), Cochran (1977, pp. 127–133), Godfrey et al. (1984), Kott (1985), Hidirolou (1986) and Detlefsen and Veum (1991), among others.

Second, utilising a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than simple random sampling would, provided that each stratum is proportional to the group’s size in the population.

Third, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can on the one hand increase the cost and complexity of sample selection, on the other hand lead to an increased complexity of population estimates. Second, examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than other methods would (although in most cases, the required sample size would be smaller than in case of simple random sampling).

A stratified sampling approach is most effective when the following conditions are met:

- Variability within strata is minimised.
- Variability between strata is maximised.
- The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages of the approach over other sampling methods are:

- It focuses on important subpopulations and overshadows, possibly ignores irrelevant ones.
- It allows the use of different sampling techniques for different subpopulations.
- It improves the accuracy/efficiency of estimation.
- It permits balancing the statistical power of tests of the various strata by departing from proportional sampling to a greater extent.

Its disadvantages are:

- It requires the selection of relevant stratification variables which can be difficult.
- It is not useful when there are no homogeneous subgroups.
- Its implementation can be expensive.

In some cases the sample designer has an access to an auxiliary variable or the size measure, believed to be correlated with the variable of interest, for each element in the population. It can be used to improve accuracy in sample design. A possible option is to use the auxiliary variable as a basis for stratification, as discussed above. Another option is probability proportional to size sampling (see Section 2.3).

Note that stratification (or *prestratification*) should not be confused with *poststratification*. The latter uses a discrete auxiliary variable to stratify the sample data *after* the sample has been selected. Its purpose is to improve efficiency of an estimator, see “Weighting and Estimation – Main Module”.

2.3 *Probability proportional to size (pps) sampling*

The *pps* approach can improve accuracy for a given sample size by concentrating the sample on large elements that have the greatest impact on population estimates. The *pps* sampling design is commonly used for business surveys whenever element size varies greatly and/or auxiliary information is available – for instance, a survey attempting to measure the number of guest-nights spent in hotels might use each hotel’s number of rooms as an auxiliary variable. In other typical examples the

auxiliary information can be the number of employees or turnover as measuring size. In some cases, a former measurement of the variable of interest can be used as an auxiliary variable for attempting to produce more current estimates. Poisson sampling (Hájek, 1960) is a *pps* sampling design with random sample sizes. This tends to be less efficient than *pps* designs with fixed sample sizes, but has the main advantage that it is easy to coordinate the samples, that is, to minimise or maximise overlap between samples selected from the same population. See “Sample Selection – Sample Co-ordination”.

2.4 *Cut-off sampling*

The *pps* approach can be used in combination with *cut-off sampling*. This is often applied to highly skewed populations, such as populations of businesses with a few large units (e.g., defined by the number of employees), and more and more, smaller and smaller values. Therefore, most of the volume for a given variable will be covered by a relatively small number of businesses, hence individual small businesses will have a little impact on population estimates. Together with the fact that the respondent burden on those businesses will be relatively high, we deliberately exclude businesses with size below a certain *cut-off threshold* from the population, i.e., give them a selection probability of zero. A short introduction to cut-off sampling is given by Knaub (2008).

Although cut-off sampling is common among practitioners, its theoretical foundations are weak. Elisson and Elvers (2001) performed a univariate analysis that compared cut-off sampling with simple stratified sampling. They conclude that the dimensional variable determining the cut-off threshold has a relevant impact on the result, so they stress that great care must be employed in choosing this variable. Benedetti et al. (2010) proposes a framework that justifies cut-off sampling and provides means for determining cut-off thresholds. They also compute the variance of the resulting estimator and its bias.

2.5 *Cluster or multistage sampling*

Cluster sampling is an alternative approach for using multiple stratification variables. It is motivated by the need for practical, economical and sometimes administrative efficiency. An important advantage of cluster sampling is that a sampling frame at the element level is not needed. Thus, in multistage sampling, a subsample is drawn from the sampled clusters at each stage except the last. At this stage all the elements from the sampled clusters can be taken in an element level sample, or a subsample of the elements can be drawn (Lehtonen and Pahkinen, 2004, p. 70). For example, in two-stage sampling the first stage can be a frame of geographical areas from which areas (*first-stage units*) are selected, and the second stage a list of businesses (*primary sampling units*) from areas selected in the first stage. Colledge et al. (1987) and Armstrong and St-Jean (1993) give examples on two-stage sampling in business statistics. For mathematical details, see, e.g., Thompson (1997, section 2.6).

A recommended method of a clustering algorithm for stratum construction is given by Jarque (1981).

2.6 *Systematic sampling*

A popular alternative to simple random sampling is systematic sampling (Cochran 1977, pp. 205–232). Stratified systematic sampling often leads to more efficient estimation than stratified simple random sampling because it can incorporate an additional level of implicit stratification within explicit strata, see example by Garrett and Harter (1995). Systematic sampling is often performed according to

an order based on random numbers. In these cases the notion *systematic* is misleading, because this sampling is random in essentials.

2.7 *Balanced sampling*

A balanced sampling design (see also “Sample Selection – Balanced Sampling for Multi-Way Stratification”) has the property that the estimators of the totals for a set of auxiliary variables are equal to the totals we want to estimate (Tillé, 2006, p. 147). Many types of sampling designs can be interpreted as balanced sampling such as simple random sampling, sampling with fixed size, stratified simple random sampling and unequal probability sampling.

2.8 *The use of panels and rotation groups*

A *panel* is defined as the collection of all units in the survey for a given period, e.g., a week, a month, a quarter, or whatever. The exclusive use of a fixed panel produces very efficient estimates of periodic change. In most periodic surveys *sample rotation* is used in order to reduce response burden (see “Response – Response Burden”). Sample rotation is closely connected to sample co-ordination (see “Sample Selection – Sample Co-ordination”). A more detailed description of the various forms of rotation sampling is given by Wolter (1979), Sigman and Monsour (1995) and Srinath and Carpenter (1995), among others.

3. **Design issues**

Not applicable.

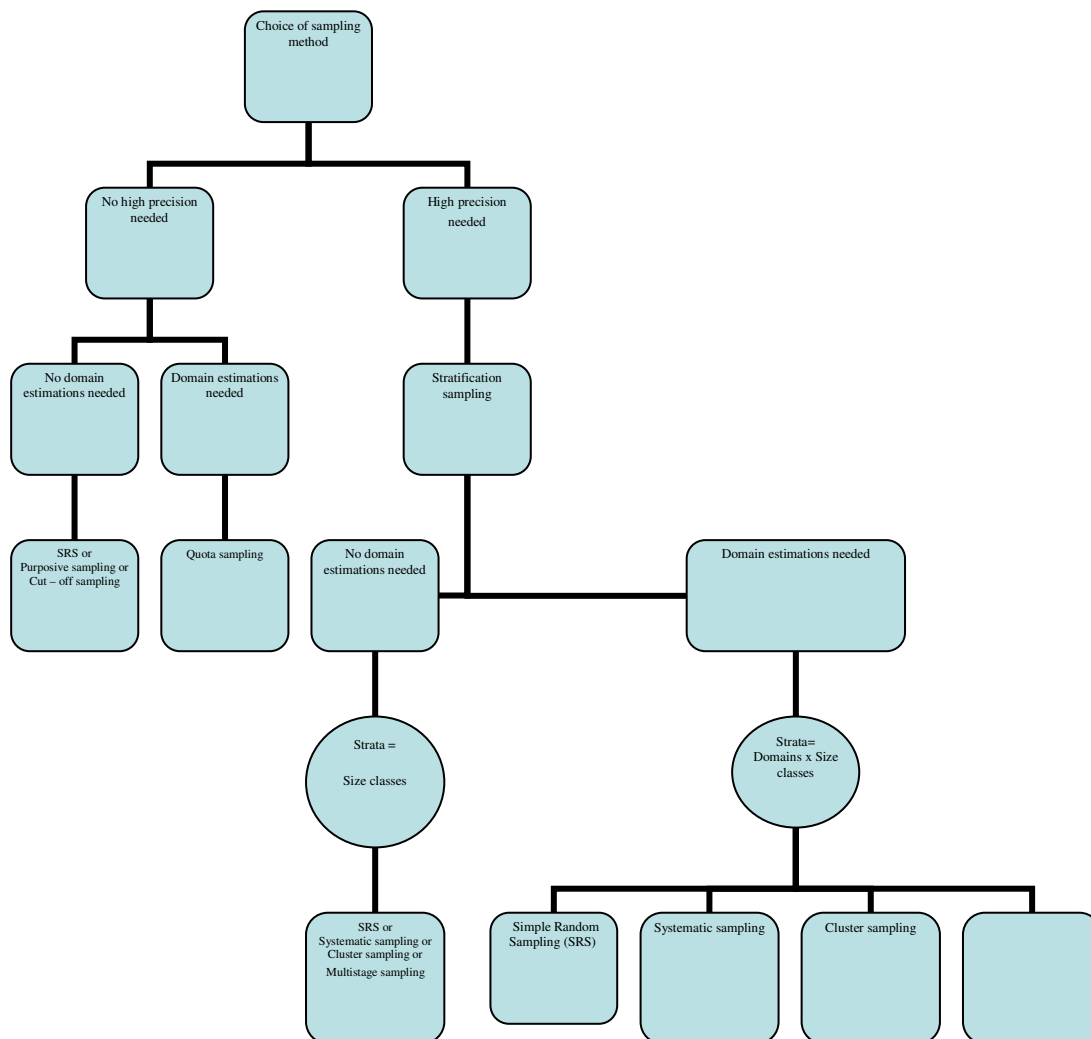
4. **Available software tools**

Packages for sample designs (<http://www.hcp.med.harvard.edu/statistics/survey-soft/>):

- [AM Software](#) from American Institutes for Research.
- [Bascula](#) from Statistics Netherlands.
- [CENVAR](#) from U.S. Bureau of the Census.
- [CLUSTERS](#) from University of Essex.
- [Epi Info](#) from Centers for Disease Control.
- [Generalized Estimation System \(GES\)](#) from Statistics Canada.
- [IVEware](#) from University of Michigan.
- [PCCARP](#) from Iowa State University.
- [R survey package](#) from the R Project.
- [SAS/STAT](#) from SAS Institute.
- [SPSS Complex Samples](#) from SPSS Inc.
- [Stata](#) from Stata Corporation.
- [SUDAAN](#) from Research Triangle Institute.
- [VPLX](#) from U.S. Bureau of the Census.

- [WesVar](#) from Westat, Inc.

5. Decision tree of methods



6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Armstrong, J. and St-Jean, H. (1993), Generalized Regression Estimation for a Two-Phase Sample of Tax Records. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, Alexandria, VA, 402–407.
- Benedetti, R., Bee, M., and Espa, G. (2010), A Framework for Cut-off Sampling in Business Survey Design. *Journal of Official Statistics* **4**, 651–671.
- Cochran, W.G. (1977), *Sampling Techniques*. Wiley, New York.
- Colledge, M., Estavao, V., and Foy, P. (1987), Experience in Coding and Sampling Administrative Data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 529–534.

- Dalenius, T. and Hodges, J. L. (1959), Minimum Variance Stratification. *Journal of the American Statistical Association* **54**, 88–101.
- Detlefsen, R. E. and Veum, C. S. (1991), Design Issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 592–596.
- Elisson, H. and Elvers, E. (2001), Cut-off Sampling and Estimation. *Proceedings of Statistics Canada Symposium*.
- Garrett, J. K. and Harter, R. M. (1995), Sample Design Using Peano Key Sequencing in Market Research. In: *Business Survey Methods* (eds. B. G. Cox et al.), Wiley, New York, 205–217.
- Godfrey, J., Roshwalb, A., and Wright, R. (1984), Model-Based Stratification in Inventory Cost Estimation. *Journal of Business and Economic Statistics* **2**, 1–9.
- Hájek, J. (1960), Limiting Distributions in Simple Random Sampling from a Finite Population. *Publications of the Mathematical Institute of the Hungarian Academy of Science* **5**, 361–374.
- Hidiroglou, M. A. (1986), The Construction of a Self-Representing Stratum of Large Units in Survey Design. *The American Statistician* **40**, 27–31.
- Jarque, C. M. (1981), A Solution to the Problem of Optimum Stratification in Multivariate Sampling. *Applied Statistics* **30**, 163–169.
- Knaub Jr., J. R. (2008), Cutoff Sampling. In: *Encyclopedia of Survey Research Methods* (ed. P. J. Lavrakas), Sage, London.
- Kott, P. S. (1985), A Note on Model-Based Stratification. *Journal of Business and Economic Statistics* **3**, 284–288.
- Lehtonen, R. and Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, Chichester.
- Sigman, R. S. and Monsour, N. J. (1995), Selecting Samples from List Frames of Businesses. In: *Business Survey Methods* (eds. B.G. Cox et al.), Wiley, New York, 133–152.
- Srintah, K. P. and Carpenter, R. M. (1995), Sampling Methods for Repeated Business Surveys. In: *Business Survey Methods* (eds. B.G. Cox et al.), Wiley, New York, 171–184.
- Tillé, Y. (2006), *Sampling Algorithms*. Springer, New York.
- Thompson, M. E. (1997), *Theory of Sample Surveys*. Chapman and Hall, London.
- Wolter, K. M. (1979), Composite Estimation in Finite Populations. *Journal of the American Statistical Association* **74**, 604–613.

Interconnections with other modules

8. Related themes described in other modules

1. Sample Selection – Sample Co-ordination
2. Weighting and Estimation – Main Module
3. Response – Response Burden

9. Methods explicitly referred to in this module

1. Sample Selection – Balanced Sampling for Multi-Way Stratification
2. Sample Selection – Subsampling for Preliminary Estimates

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

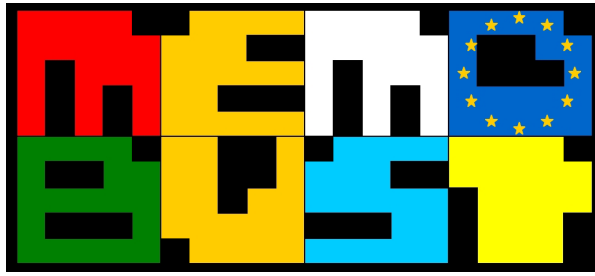
Sample Selection-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	18-02-2013	first version	Magnar Lillegård	SSB
0.2	02-04-2013	second version	Magnar Lillegård	SSB
0.2.1	06-09-2013	preliminary release		
0.2.2	09-09-2013	page numbering adjusted		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:41



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Balanced Sampling for Multi-Way Stratification

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	4
4. Examples – not tool specific.....	5
4.1 Example: the multi-way stratification design for controlling the sample size	5
4.2 Example: the multi-way stratification design to retain the sample allocation.....	5
4.3 Example: the multi-way stratification design for reducing the response burden	6
5. Examples – tool specific.....	6
6. Glossary.....	7
7. References	7
Specific section.....	8
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

Balanced sampling is a class of techniques using auxiliary information at the sampling design stage. Many types of sampling designs can be interpreted as balanced sampling, such as simple random sampling with fixed size, stratified simple random sampling and unequal probability sampling.

Furthermore, the balanced sampling can be applied to define a multi-way stratification design also known as incomplete stratification or marginal stratification. Multi-way stratification allows to plan the sample sizes of the domains of interest belonging to two or more non-nested partitions of the population in question without using the standard solution based on a stratified sample in which strata are identified by cross-classifying the variables defining the different partitions (one-way stratified design). The standard solution in many Structural Business Surveys (SBSs) may have drawbacks from the view-point of cost-effectiveness. In fact, SBSs produce typically estimates for a great number of very detailed domains forming several non-nested partitions of the population and creating really small cross-classified strata.

2. General description of the method

Balanced sampling can be applied according to two different inferential approaches: the model based approach (Royall and Herson, 1973, Valliant *et al.* 2000) and the design based or randomisation assisted approach (Deville and Tillé, 2004). The first approach bases the inference on a statistical superpopulation model and it may be performed by probability or non-probability sample. In this framework a sample is balanced when the sample means of a set of auxiliary variables (balancing variables) are equal to the known population means (Valliant *et al.*, 2000). Balanced samples are used to follow a robust sampling strategy. The design based approach needs a sampling frame and uses a probability sample to make inferences. In this second context a sample is balanced when the Horvitz-Thompson (H-T) sample estimates for the auxiliary variables are equal to their known population totals. The selection of a balanced sample generally improves the efficiency of the sampling estimates (Cochran, 1977). This section focuses on this second inferential approach.

A widely used application of the method is stratified simple random sampling. As known, it has been introduced in the sampling methodology to enhance the efficiency of the estimates. Nevertheless, stratified sampling can be used as an operative tool in the surveys as well. An instrumental use of stratified sampling is when the objective of the survey is to produce estimates for some subpopulations (or domains) forming two or more non-nested partitions of the population and a fixed or planned sample size for each domain is required. A standard sampling design solution defines strata by cross-classifying the variables defining the different partitions. In this case stratification is not strictly used to improve estimation quality. It is used to implement a random selection method guaranteeing the selected sample sizes corresponding to the planned ones. This standard solution, hereinafter denoted as one-way stratified design, may have some drawbacks, especially in the SBSs.

When the number of cross-classified strata is too large, there are some immediate consequences, described as follows:

- (i) the overall sample size could easily be too large for the survey economic constraints;

- (ii) when the population size in many strata is small, the stratification scheme becomes inefficient; in other words the sample allocation may be far from the theoretically desired allocation;
- (iii) when there are strata containing only few units in the population, a not equally distributed response burden may arise in surveys repeated over time.

Many methods have been proposed in the literature to keep that the sample size under control in all the domains without using one-way stratified designs. This means that sample size of each cross-classified stratum is a random variable. These approaches may be roughly divided into two main categories. The first category contains methods commonly known as controlled selection. Seminal papers have been proposed by Bryant et al. (1960) and Jessen (1970). Other methods based on controlled rounding problems via linear programming have been proposed by Causey et al. (1985), Rao and Nigam (1990; 1992), Sitter and Skinner (1994) and Winkler (2001).

In the second category there are methods based on sample coordination. A separate sample is selected for each partition in order to guarantee the maximum overlap among the different samples (Ohlsson, 1995; Ernst and Paben, 2002). We define all these methods as multi-way stratified designs.

Literature shows that these methods pose theoretical and operative problems especially for large scale surveys as in the SBSs. A recently proposed method, the Cube algorithm (Deville and Tillé, 2004) overcomes these drawbacks. The method, included in the first category, has been originally defined for drawing balanced samples with a large number of balancing variables for large population size. Multi-way stratification is a special case of balanced sampling. Given the population U of size N , let π_k be the inclusion probability of k -th population unit ($k=1, \dots, N$) and let δ_{dk} be the value of the indicator variable of the domain U_d , being $\delta_{dk}=1$ if the unit k belongs to domain U_d and equal to zero otherwise. Then, by definition the sample size of U_d is $\sum_{k=1}^N \delta_{dk} \pi_k = n_d$. The Cube method assumes that the inclusion probabilities are known, and it selects a random sample achieving the consistency among the known totals and the H-T estimates. We define the following auxiliary variables $z_{dk} = \delta_{dk} \pi_k$ for each domain. When the sample is balanced on the z variables the H-T estimate $\hat{Z}_d = \sum_{k=1}^N s_k z_{dk} / \pi_k$ has to be equal to the known population total $Z_d = n_d$ with s_k being a random variable equal to 1 if the k -th unit belongs to the sample and equal to 0 otherwise. For satisfying the balancing equations, $\hat{Z}_d = Z_d$, the Cube algorithm has to select n_d units from U_d . When the expected sample sizes are integer numbers the Cube algorithm applied to obtain a multi-way stratification always finds the solution. Some illustrative examples of multi-way sampling designs are given in Falorsi and Righi (2008).

3. Preparatory phase

4. Examples – not tool specific

4.1 Example: the multi-way stratification design for controlling the sample size

In order to explain the problem, we consider the population of 165 schools reported in Table 1 (Cochran, 1977, p. 124). We assume that the parameters of interest are the totals of a variable, related to school, separately for the *Size of city* (5 categories: *I,II,III,IV,V*) and for the *Expenditure per pupil* (4 categories: *A,B,C,D*). Two distinct partitions of the population are defined: the size of city (first partition) defining 5 non-overlapping domains, and the expenditure per pupil defining 4 domains. We have 9 domains of interest.

Table 1.

		<i>Expenditure per pupil</i>				
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	Totals
<i>Size of city</i>	<i>I</i>	15	21	17	9	62
	<i>II</i>	10	8	13	7	38
	<i>III</i>	6	9	5	8	28
	<i>IV</i>	4	3	6	6	19
	<i>V</i>	3	2	5	8	18
	Totals	38	43	46	38	165

The standard one-way stratified design (or cross-classification design) defines $20=5 \times 4$ strata by crossing the categories of the domains of the two partitions. Due to budgetary constraints, we suppose that the sample size could be up to 10 units. Nevertheless, in each stratum at least one school should be selected (or two schools for estimating the sampling variance without any bias) and, consequently, according to this design the sample size should amount to 20 (or 40) schools at least. Hence, the cross-classification design becomes unfeasible.

4.2 Example: the multi-way stratification design to retain the sample allocation

We consider the above population of schools. We plan a sample of 20 schools and we want to allocate the sample proportionally to the domain size. Table 2 shows the planned size and the integer rounded sample allocation. We note that the sample sizes in the cross-classified strata are not constrained to be integers.

Table 2.

		<i>Expenditure per pupil</i>				Domain weight	Rounded planned sample size
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>		
<i>Size of city</i>	<i>I</i>	0.0909	0.1273	0.1030	0.0545	0.3757	8
	<i>II</i>	0.0606	0.0485	0.0788	0.0424	0.2303	5
	<i>III</i>	0.0364	0.0545	0.0303	0.0485	0.1697	3
	<i>IV</i>	0.0242	0.0182	0.0364	0.0364	0.1152	2
	<i>V</i>	0.0182	0.0121	0.0303	0.0485	0.1091	2
	Domain weight	0.2303	0.2606	0.2788	0.2303	1.0000	
Rounded planned sample size		5	5	5	5		20

According to the one-way stratified design of size 20 (Table 3), we obtain a sample allocation far from the planned one.

Table 3.

		<i>Expenditure per pupil</i>				Domain weight	Rounded planned sample size
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>		
Size of city	<i>I</i>	1	1	1	1	0.2000	8
	<i>II</i>	1	1	1	1	0.2000	5
	<i>III</i>	1	1	1	1	0.2000	3
	<i>IV</i>	1	1	1	1	0.2000	2
	<i>V</i>	1	1	1	1	0.2000	2
	Domain weight	0.2500	0.2500	0.2500	0.2500	1.0000	
Rounded planned sample size		5	5	5	5		20

4.3 Example: the multi-way stratification design for reducing the response burden

We consider the population of schools described in section 4.1 and we suppose that the population distribution to be fixed over time. We have to select a sample of size 40 on several survey occasions. Moreover, we want to compute unbiased variance estimates. According to the one-way stratified design we have to select 2 schools per stratum (Table 4).

Table 4.

		<i>Expenditure per pupil</i>				Domain weight	Rounded planned sample size
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>		
Size of city	<i>I</i>	2	2	2	2	0.2000	8
	<i>II</i>	2	2	2	2	0.2000	8
	<i>III</i>	2	2	2	2	0.2000	8
	<i>IV</i>	2	2	2	2	0.2000	8
	<i>V</i>	2	2	2	2	0.2000	8
	Domain weight	0.2500	0.2500	0.2500	0.2500	1.0000	
Rounded planned sample size		10	10	10	10		40

Then, we can see that the schools in stratum *V-B* are drawn with certainty on each survey occasion and the schools in strata *IV-B* and *V-A* have a high probability to be included in the samples. That happens because in stratum *V-B* there are only two schools in the population, while in the strata *IV-B* and *V-A* there are three schools in the population and the inclusion probability is 0.67. Hence, the response burden is not equally distributed in the population of schools (is high for the schools belonging to small population size strata) and this burden does not depend on efficiency issues.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Bryant, E. C., Hartley, H. O., and Jessen, R. J. (1960), Design and Estimation in Two-Way Stratification. *Journal of the American Statistical Association* **55**, 105–124.
- Causey, B. D., Cox, L. H., and Ernst, L. R. (1985), Applications Transportation Theory to Statistical Problem. *Journal of the American Statistical Association* **80**, 903–909.
- Cochran, W. G. (1977), *Sampling Techniques*. Wiley, New York.
- Deville J.-C. and Tillé, Y. (2004), Efficient Balanced Sampling: the Cube Method. *Biometrika* **91**, 893–912.
- Ernst, L. R. and Paben, S. P. (2002), Maximizing and Minimizing Overlap When Selecting Any Number of Units per Stratum Simultaneously for Two Designs with Different Stratifications. *Journal of Official Statistics* **18**, 185–202.
- Falorsi, P. D. and Righi, P. (2008), A Balanced Sampling Approach for Multi-Way Stratification Designs for Small Area Estimation. *Survey Methodology* **34**, 223–234.
- Jessen, R. J. (1970), Probability Sampling with Marginal Constraints. *Journal of the American Statistical Association* **65**, 776–795.
- Lu, W. and Sitter, R. R. (2002), Multi-Way Stratification by Linear Programming Made Practical. *Survey Methodology* **28**, 199–207.
- Ohlsson, E. (1995), Coordination of Samples using Permanent Random Numbers. In: *Business Survey Methods* (eds. Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S.), Wiley, New York, Chapter 9.
- Rao, J. N. K. and Nigam, A. K. (1990), Optimal Controlled Sampling Design. *Biometrika* **77**, 807–814.
- Rao, J. N. K. and Nigam, A. K. (1992), Optimal Controlled Sampling: a Unifying Approach. *International Statistical Review* **60**, 89–98.
- Royall, R. and Herson, J. (1973), Robust Estimation in Finite Population. *Journal of the American Statistical Association* **68**, 880–889.
- Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.
- Winkler, W. E. (2001), Multi-Way Survey Stratification and Sampling. RESEARCH REPORT SERIES, Statistics #2001-01, Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233.

Specific section

8. Purpose of the method

Balanced sampling is used for selecting a multi-way stratified design, which is a sampling design planning the sample sizes for domains of interest belonging to different partitions of the population without using a one-way stratified or cross-classified stratification design.

9. Recommended use of the method

1. The method can be applied when the one-way stratified designs (those where strata are obtained by combining the domains of different partition of the population) can be inefficient or can produce statistical burden for surveys repeated over time.
2. The method may be applied in large scale surveys, with large population and a lot of domains.
3. The method may be useful in the small area estimation problem when the membership indicator variables for small areas are known at population level. Planning the sample size for each domain allows to estimate specific small area effects improving the efficiency of indirect model based small area estimators.

10. Possible disadvantages of the method

1. The method needs to know the inclusion probabilities. The definition of the optimal inclusion probabilities is less intuitive than in case of one-way stratification design.
2. Analytic expression of the variance of the estimates is unknown. Approximations (shown in the literature) are needed.
3. Some difficulties when a complex estimator is used for the computation of sampling errors.

11. Variants of the method

1. Balanced Sampling for multi-way stratification is defined to select a planned sample size for each domain. In addition, it may be worthwhile including other balancing variables to enhance the estimation efficiency according to the calibration estimation theory.

12. Input data

1. Data including the domain membership indicator variable and the inclusion probability for each population units are needed.

13. Logical preconditions

1. Missing values
 1. Not allowed.
2. Erroneous values
 1. Not allowed.
3. Other quality related preconditions

- 1.
4. Other types of preconditions
 1. The sum of the inclusion probabilities over each domain must be an integer.
 2. The sum over population domains of the inclusion probabilities must be equal for each partitions.

14. Tuning parameters

1. Depending on the origin of the inclusion probabilities a calibration step could be needed. The calibration step modifies the probabilities for satisfying sample size consistency among the partitions and for achieving an integer expected sample size in each domain (see 13.4.1).

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. Sample membership indicator variable is added in the input data set.

17. Properties of the output data

1. The sum of the sample membership indicator variable over each domain is equal to the expected sample size.

18. Unit of input data suitable for the method

Processing full data set.

19. User interaction - not tool specific

1. Definition of the set of inclusion probabilities.
2. Before execution of the method, verify that the planned sample sizes for each domain are integer numbers and consistent.
3. When performing a multi-way stratification design considering also other balancing variables in the sample selection process, the indicators of the quality of balancing have to be analysed.

20. Logging indicators

1. No specific indicators.

21. Quality indicators of the output data

1. When used only for multi-way stratification, the theory shows that the method selects exactly a sample satisfying the planned sample size. When other balanced variables are added, the ratio among the H-T estimates and the known totals are used as quality indicators.
2. No other quality indicators are used to strictly evaluate the performances of the methods.

22. Actual use of the method

1. Balanced sampling is widely used in the Insee not specifically for implementing multi-way stratification.
2. Istat has used balanced sampling for a population survey.
3. An Istat research project is studying the optimal allocation for multi-way stratified design.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Sample Selection – Main Module
2. Sample Selection – Sample Co-ordination

24. Related methods described in other modules

- 1.

25. Mathematical techniques used by the method described in this module

1. The method mainly implements a balancing martingale theory, with the aim to round off each inclusion probabilities randomly to 0 or 1. From the mathematical point of view that corresponds to the maximisation of the entropy measure (maximisation of the randomness) under linear constraints (balancing equations).

26. GSBPM phases where the method described in this module is used

1. 2.4 Design Frame & Sample methodology
2. Partially 4.1 Select sample

27. Tools that implement the method described in this module

1. Sampling R package
2. SAS Macro downloadable Insee site

28. Process step performed by the method

Sample planning and selection

Administrative section

29. Module code

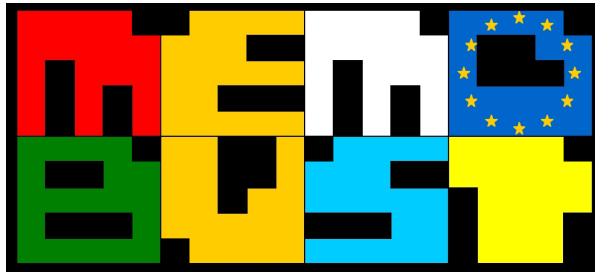
Sample Selection-M-Balanced Sampling

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	10-12-2011	first version	Paolo Righi	ISTAT
0.2	29-02-2012	second version	Paolo Righi	ISTAT
0.3	22-02-2013	third version	Paolo Righi	ISTAT
0.3.1	06-09-2013	preliminary release		
0.3.2	09-09-2013	page numbering adjusted		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:42



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Subsampling for Preliminary Estimates

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 The inferential approach.....	5
2.2 Sampling design for design-based/model-assisted approach	5
2.3 Sampling design for model-based approach.....	6
3. Preparatory phase	8
4. Examples – not tool specific.....	8
4.1 Example: Comparison of different sampling designs for a preliminary subsample based on the Italian Monthly Retail Trade Survey data	8
5. Examples – tool specific.....	15
6. Glossary.....	15
7. References	15
Specific section.....	17
Interconnections with other modules.....	19
Administrative section.....	20

General section

1. Summary

Among the main components of the quality in official statistics, the timeliness seems to be one of the most relevant both for producers and users of statistical data. In particular timeliness is becoming a pressing target especially for short term statistics (EUROSTAT, 2000). Therefore, in recent years, in many fields of official short term statistics the timeliness is becoming the driving issue, both for the increasing demand of users and the need to fill the gap comparing to data release standards already achieved by USA and other developed countries. The Amendment EU Regulation on Short Term Statistics (introduced in August 2005, EUROSTAT) requests all the statistical institutes of the EU Member States to transmit preliminary short term indicators to EUROSTAT with a reduced delay comparing to the timeliness set in the original 1998 Regulation. Frequently, in the NSIs short term statistics are based on fixed panel surveys of enterprises or rotating panels with a partial overlap from one year to another. Auxiliary variables coming from the previous survey occasions are often available.

A common approach for dealing with *preliminary estimates* focuses essentially on the study and the definition of efficient estimators, exploiting almost exclusively auxiliary information in the estimation phase. In such context sampling has a marginal role. Preliminary estimation merely involves the use of the quick respondent units. In fact, in order to obtain “good” preliminary estimates, standard survey strategy often aims to achieve high quick response rate by means of a well-structured plan of follow-up. In some surveys the “largest” units are carefully supervised. Following this approach, we point out that there is no explicit definition of sampling design for preliminary estimation, but that for the approach trying to observe large units. Hence, the preliminary estimates are usually drawn by a non-probabilistic sample design. A useful documentation on preliminary estimation problems (even though not comprehensive) can be downloaded from the OECD web site¹.

The topic investigates alternative sampling approaches for planning the subsamples for preliminary estimates. These designs try to exploit the auxiliary information in an efficient way according to the estimator used for the preliminary and final estimation. Therefore, an *overall strategy* for the production of preliminary estimates is developed, involving both the sample design and the estimator definition.

2. General description of the method

Given a sampling survey, a preliminary (or provisional) estimate is defined. It means the estimation of a parameter of interest obtained on the basis of a sample of quick respondent units available within a time lag Δ'_t after the reference time point (or end of the reference period) t of the survey, while the correspondent final estimate is based on both quick and late respondents (final sample), observed within a time lag Δ_t ($> \Delta'_t$). The indicators measuring the statistical quality of a preliminary estimation method are based on the differences between the two estimates. These differences are known as *revisions* or *revision errors*. For a detailed description of the indicators for statistical quality

¹ For the issue of the preliminary subsample the link is:

http://www.oecd.org/document/17/0,3746,en_2649_33715_30386193_1_1_1_1,00.html.

of preliminary estimations see, for instance, Di Fonzo (2005). The quick respondents can be observed according to different sampling processes. In particular we can observe the sample of quick respondents

- (a) without any sampling and follow-up plans: we denote it as *Unplanned Preliminary Observed Sample* (UPOS);
- (b) without any sampling plan but with a follow-up plan for the large final sampled units: we denote it as *Partially Unplanned Preliminary Observed Sample* (PUPOS);
- (c) with a planned subsample for preliminary estimates. Then a *Preliminary Theoretical Sample* (PTS) is drawn and an intensive follow-up of the PTS units is planned so that *Planned Preliminary Observed Sample* (PPOS) will be as close to PTS as possible.

Before exposing the topic devoted to the sampling process (c), we make some general remarks:

1. preliminary estimation has two goals: producing accurate estimates of the parameters of interest; producing estimates with small revisions comparing to the final estimates. In some sense, this second goal can be more important for a NSI than the first one because the statistical users can compare the preliminary and final estimates and they can have a concrete perception of the sampling errors. Typical examples are the trend estimates where the preliminary and final estimates could have opposite signs, although the two estimation procedures can produce accurate and unbiased estimates;
2. the preliminary estimation issue arises also in surveys based on administrative data. Many aspects of the topic are suitable for such kind of surveys, but some others not. The topic does not tackle these peculiarities. Baldi et al. (2003) gives many interesting indications illustrating the problem and the possible solutions for the Employment, Wages and Labour Cost Survey conducted by the Italian NSI;
3. few references in literature on sampling design aiming at the preliminary estimation are available (cf. OECD link; D'Alò et al., 2007; Righi and Tuoto, 2007).

Regarding the definition of the preliminary samples we point out that:

4. sampling process (b) is a special case of sampling process (c);
5. as far as sampling processes (b) and (c) are concerned, we highlight that preliminary estimation has a distinctiveness with respect to the standard estimation process (producing the final estimates). The researcher using PUPOS or PPOS will obtain responses also from other sampled units of the final sample not included in the preliminary subsample and/or in the follow-up plan;
6. the sampling process (c) assumes that there is significant difference between early respondents and late respondents. Then intensive follow-up of the PTS aims to survey all units that would belong to the two categories in a standard context. The small size of the PTS had to guarantee a small nonresponse rate;
7. using sampling process (c), the researcher can define an overall strategy taking into account the functional form of the parameters of interest (in general totals, indices or ratios) and the preliminary and final estimators. In an extensive vision of the problem,

the sampling design for PTS must be coordinated with the sampling design of the final sample and with the provisional and final estimation process. The ideal situation is to plan all these elements at the same time and, in practice, the two samples are coordinated according to an optimisation problem that takes into account of the trade-off between publishing early (risking a high revision error) and publishing late (which is not attractive) with a smaller risk for a high revision error. Nevertheless, the topic does not treat such huge context, but considers a restricted field typical for many sampling surveys. The provisional estimation goals and consequently the PTS are defined after the final sample and estimator were fixed, while the time lag of the provisional estimates is given by some legislative regulation on official statistics. In this case the possibility of defining different types of sampling strategy is restricted. If the strategy used for final estimation is optimal (within a given family of estimators and according to a design- or model-based approach), there is no particular reason for justifying the use of a quite dissimilar strategy for preliminary estimation. Secondly in order to reduce the revisions the form of the provisional estimator should be similar to the given final estimator.

The basic issue of using the PTS is the intensive follow-up that must be guaranteed for applying the method (point 6). If the PTS is affected by high non response, in general, small revisions cannot be obtained by letting the preliminary estimator resembling the final estimator. In this case it needs to define a specific provisional estimator following the approach typically used when the estimates are based on UPOS. An example of such approach is given by Rao et al. (1989). As far estimation process is concerned, the modules “Weighting and Estimation – Preliminary Estimates with Design-Based Methods” and “Weighting and Estimation – Preliminary Estimates with Model-Based Methods” shows some techniques. Here we pay the attention to the sampling phase.

Strategies for contact and follow-up of sampled units are dealt with in the module “Data Collection – Design of Data Collection Part 2: Contact Strategies”.

2.1 *The inferential approach*

The sampling design for a PTS has to be defined according to the inferential approach. We distinguish two classical alternative inferential paradigms: the design-based/model-assisted and the model-based approaches (see also “Weighting and Estimation – Main Module”). The literature has studied the two approaches for the final estimates from different points of view and currently neither of them is dominant, although in the official statistics the design-based/model-assisted prevails. However, in the preliminary estimation context the reference framework is unlike from the context considering only the final estimate. We refer to the elements described in points 2.2 and 2.3, common in the preliminary estimation and missing in the final estimation process. Such elements can drive to prefer the model-based approach.

2.2 *Sampling design for design-based/model-assisted approach*

The PTS (selected from the final sample) has been drawn according to a random selection procedure. Standard sampling designs (simple random sampling, stratified simple random sampling, unequal probability sampling design etc.) can be implemented (see “Sample Selection – Main Module”).

The choice of a sampling design depends on:

- the explicative power of the auxiliary variable (known at the design phase) for the variables of interest;
- the sampling design for the final estimates.

The second condition is established essentially to define a preliminary estimation process as similar as possible to the final estimation process. Then, if a stratified design is used for the final sample, in general it is better to use the same design for the PTS even though the stratification should be more aggregate because of a smaller sample size. The aim is to limit the revisions.

Advantages

The main advantages of a random PTS are the following:

- the inference of the design-based/model-assisted approach with the PTS does not suffer from bias;
- if the final estimates are design-based/model-assisted, it is better to use the same approach for the preliminary estimates in order to bound the revisions.

The most important condition for achieving these advantages is that the sampling design should be followed by a good follow-up plan for the preliminary and final sample. If the PTS and PPOS are quite dissimilar and the final sample has a high non response rate, the revisions can be very high and systematic, producing an highly undesirable upward or downward revisions in each survey occasion.

Disadvantages

The drawbacks of using a random subsample depend on whether the inferential approach is suitable. In fact, the preliminary estimation has a special parameter as the final estimate. Comparing the two estimates we obtain the revision. In the ideal situation, when the PPOS is the PTS and the final sample is fully observed, the revision represents simply estimate error. The inferential paradigm assures that under the preliminary sampling design the expected revision is zero. Nevertheless, such condition holds rarely and preliminary response and non-response for the units not belonging to the PTS have to be dealt with. However, there is a further complexity due to the variability of the final estimates because of final or late non-response. It means that the final estimates are random variables depending on the unknown non-response mechanism of the final sample. Since the inferential approach based on random sample does not cover the possibility of non-fixed parameters of interest (except for some special model assumptions), we have to use the model-based approach. In this case a non-random sample can be drawn.

2.3 Sampling design for model-based approach

The model-based approach does not require a random sample for making inference. The preliminary sample can be purposive, judgemental or non-random. On the other hand, the preliminary sample has to respect some features depending on the superpopulation model which generates the data for obtaining efficient estimates. In particular, it is important for the preliminary estimation to concentrate on the non-response mechanism. As mentioned earlier, in the real survey context the use of models in preliminary estimate problems is quite common because it is necessary to deal with the preliminary non-response, the preliminary response for the units not belonging to the PTS and the non-response of the final sample. We underline that in the last case, when the researcher has to deal with preliminary

estimation problem, he/she has to model the final non-response before observing it, in order to estimate the expected composition and size of the corresponding final sample.

The researcher usually does not know the non-response mechanisms, thus he/she has to make some assumptions defining the *working models*. Model-based approach makes inference on the working models, assuming that they represent satisfactory approximations of the true non-response mechanisms. Nevertheless, if the working models are seriously incorrect, the estimates can be strongly biased and the revisions can be systematically positive or negative. To avoid these problems, robust sampling strategies can be defined, in the sense that they perform well with the working and alternative models.

As far as the sample selection to protect from bias is concerned, we consider the balanced sampling design for drawing the PTS. Roughly speaking, in the model-based approach a sample is defined as balanced on a set of auxiliary variables if the sample and the known population means of the auxiliary variables are equal (Royall and Herson, 1973; Valliant et al., 2000). According to the considered estimator, different kinds of balanced samples can be used. Therefore, before defining the sampling design, the knowledge of the estimator form is necessary. The use of a balanced sample defines a *bias-robust* strategy.

The example of section 4 suggests how to implement a balanced sample for a real survey.

Advantages

If a model-based approach is used, a random or a purposive sample can be drawn. Nevertheless, there are some theoretical and operative advantages of using a purposive sample. From theoretical point of view:

- a suitable sample according to the working models used in the estimator can be drawn. Assuming that the working models hold, suitable purposive samples produce efficient estimates. When the researcher has no evidence that the working models represent satisfactory approximations, balanced samples produce robust estimates.

Considering the operative aspects, we refer to the short term statistics based on a panel component or longitudinal data. Frequently, these surveys rely on a set of sample units with high quick response rate achieved after a sensitisation work in the previous survey occasions. Then, in the perspective of renovating the preliminary sample with a non-random sample,

- it is easier to include the units that have shown high quick response probability in the preliminary subsample.

Disadvantage

The drawbacks to use a purposive or a non-random sample are linked to the inferential paradigm. If the working models are far from the actual non-response mechanisms, the inferences can be biased. On the other hand, model-based sampling theory suggests that it could be useful to select a random sample with this approach as well. Such samples could preserve the inference from the biasedness.

A second disadvantage can emerge when the final estimation is design-based/model-assisted. However, we remark that even with this approach the use of models is rather widespread.

3. Preparatory phase

4. Examples – not tool specific

4.1 Example: Comparison of different sampling designs for a preliminary subsample based on the Italian Monthly Retail Trade Survey data

The complexity of the preliminary estimation problem allows giving only a few general indications about the steps for defining a preliminary subsample for the provisional estimates. The following example shows how the process can be defined, but the main conclusion we want to highlight is that the preliminary sampling design has to take carefully into account the estimation process. The example is based on the data of the Italian Monthly Retail Trade Survey (MRTS), collected in 2004 (De Sandro and Gismondi, 2004).

4.1.1 Parameters of interest, preliminary and final estimates of the Italian Monthly Retail Trade Survey

The MRTS is based on the monthly measurement of the turnover of a stratified sample of retail enterprises (Division 52 of NACE nomenclature for a population of about 570 thousands) of different types and sizes. The sample is composed by a panel and a non-panel component, drawn every year and observed for 12 months. The survey provides provisional estimates within 30 days after the reference time and final estimates within 54 days according to the EU user needs (Eurostat, 2000). The provisional retail trade indices are referred to the domains: type of product sold (food and non-food retail enterprises) and type of distribution (large and small retail enterprises). Then the parameters of interest at the month t are defined as

$$I_d^{t,0} = \left(\sum_{h \in d} I_h^{t-12,0} R_h^t \gamma_h \right) / \sum_{h \in d} \gamma_h, \quad \text{with} \quad R_h^t = \frac{\sum_{i \in U_h^{t,t-12}} Y_i^t}{\sum_{i \in U_h^{t,t-12}} Y_i^{t-12}},$$

where d is the generic domain of interest; h is the generic stratum defined by the cross-classification of the main group of product sold, the class of employed persons and the type of distribution for 120 strata; $I_h^{t-12,0}$ is the retail trade index of the same month t of the previous year in the stratum h (with $t=13, 14, \dots, 24$)²; γ_h is a stratum weight given by the yearly turnover in 2000, derived from structural business statistics (ASIA archive); Y_i^t and Y_i^{t-12} are the total turnover variables of the unit i in month t and the same month of the previous year, respectively; $U_h^{t,t-12}$ is the longitudinal population of stratum h in time period $(t, t-12)$. The product term $(I_h^{t-12,0} R_h^t)$ represents the elementary index at stratum level.

² For instance January is indicated with $t=13$ and the same month of the previous year with $t-12=1$.

The sampling design is stratified simple random sampling, with about 7,500 units. Each year about 30% of the sample is renewed³.

In the questionnaire of the reference month t both the values of the variables Y^t and Y^{t-12} are collected with some other auxiliary variables.

Starting at the end of 2004, the evaluation of the preliminary estimates is based on an UPOS calculated after $\Delta'_t=29$ days from the end of the reference month. The estimation phase follows a complex procedure. Here the main steps, used in the simulation study, are sketched.

All the non-respondents within Δ'_t are imputed to obtain the provisional estimates. For each domain of interest the provisional estimation process is given by

$$\tilde{I}_d^{t,0} = \left(\sum_{h \in d} \tilde{I}_h^{t-12,0} \tilde{R}_h^t \gamma_h \right) / \sum_{h \in d} \gamma_h, \quad (1)$$

with

$$\tilde{R}_h^t = \frac{\sum_{i \in s_{ah(t)}^t} y_i^t + \sum_{i \in (\tilde{s}_h^t - s_{ah(t)}^t)} \tilde{y}_i^t}{\sum_{i \in s_{ah(t-12)}^t} y_i^{t-12} + \sum_{i \in (\tilde{s}_h^{t-12} - s_{ah(t-12)}^t)} \tilde{y}_i^{t-12}}, \quad (2)$$

where $\tilde{I}_h^{t-12,0}$ is the estimate of $I_h^{t-12,0}$; y_i^t and y_i^{t-12} are the observed values of Y_i^t and Y_i^{t-12} , \tilde{y}_i^t and \tilde{y}_i^{t-12} are the imputed values for the non-respondents; $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ are the respective sample units giving information for the preliminary estimates about the variables Y^t and Y^{t-12} in stratum h in month t with $s_{ah(t)}^t \subseteq \tilde{s}_h^t$ and $s_{ah(t-12)}^t \subseteq \tilde{s}_h^{t-12}$, where \tilde{s}_h^t is the theoretical overall sample for the final estimates in stratum h in month t , while $(\tilde{s}_h^t - s_{ah(t)}^t)$ and $(\tilde{s}_h^{t-12} - s_{ah(t-12)}^t)$ are the corresponding non-respondent samples after the time lag Δ'_t for the variables Y^t and Y^{t-12} , respectively. We suppose that unit i providing the value of y_i^t gives information on y_i^{t-12} as well; then, $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ coincide and they are indicated by s_{ah}^t .

The imputation procedure is defined by two steps:

$$\tilde{y}_i^{t-12} = a_i^{t-12} \frac{\sum_{i \in s_{ag}^t} y_i^{t-12}}{\sum_{i \in s_{ag}^t} a_i^{t-12}}, \quad (3)$$

$$\tilde{y}_i^t = \frac{\sum_{i \in s_{ag}^t} y_i^t}{\sum_{i \in s_{ag}^t} y_i^{t-12}} \frac{a_i^t}{a_i^{t-12}} \tilde{y}_i^{t-12}, \quad (4)$$

³ For practical reasons this percentage could be higher. For instance in 2004 data, analysed in the simulation study, about 50% of the sample belongs to the panel component (observed in the 2003 survey) while the other part is a new sample.

where $s_{ag}^t = \bigcup_{h \in g} s_{ah}^t$ represents the sample of the quick respondents of size n_{ag}^t belonging to the imputation cell g defined by crossing the type of distribution and the class of employed persons (8 cells, 3 for large and 5 for small retail enterprises); a_i^t and a_i^{t-12} are the numbers of persons employed in the respective months t and $t-12$ for the unit i , observed in the survey or imputed. Imputation is performed by the following procedure: the missing value of the variable a_i^{t-12} is imputed by the value a_i^t if it is not missing, otherwise it is imputed by the value of the business register; before imputing a_i^t , the outlier values considering the ratio a_i^t / a_i^{t-12} are checked. If the ratio does not belong to the interval $(0.1, 10)$, the value a_i^t is replaced by a_i^{t-12} . If a_i^t is missing, it is imputed by a_i^{t-12} . In the expressions (3) and (4) we ignore this imputation process and always consider these two variables as observed. The final estimation has the same steps as the preliminary one, working with the information of both the quick and late respondents.

Finally, let us note that, although a probabilistic sample is used and the numerator and denominator of (2) are Horvitz-Thompson estimates with the imputation of the missing values, the sampling weights are annulled. These estimates can be analysed in the model-based context as well.

4.1.2 Definition of the Preliminary Theoretical Sample in the MRTS

The task of planning a subsample integrated with the provisional estimator, defining an overall preliminary sampling strategy needs to give an explicit form of the model of the imputation procedure. In the MRTS the procedure is quite complex. To keep things simple, we consider only the imputation processes defined in (3). In the model-based approach the process is the Best Linear Unbiased provisional estimator with respect to the final estimates if and only if the following superpopulation model generating the data is:

$$\begin{cases} Y_i^{t-12} = \beta_g a_i^{t-12} + \varepsilon_i & (i \in g), \\ E(\varepsilon_i) = 0; E(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 a_i^{t-12} & \text{if } i = j, \\ 0 & \text{else.} \end{cases} \end{cases} \quad (5)$$

Denoting by $\tilde{T}_{Y(d)}^{t-12}$ the provisional estimate of $\hat{T}_{Y(d)}^{t-12}$, the final estimate of the total of the variable Y_d^{t-12} with the final theoretical sample, if (5) is the true model, the expected revision $(\tilde{T}_{Y(d)}^{t-12} - \hat{T}_{Y(d)}^{t-12})$ is zero and it has the smallest variance under (3).

The second imputation step (4) cannot be expressed in a linear superpopulation model. Nevertheless, a reduction of the imputation error in the first step is important for a “good” imputation in the second one.

Under the working model (5), the optimal sampling strategy requires that the quick respondent sample is given by the units whose a^{t-12} values are the largest (Royall and Herson, 1973). However, the (5) is just a working model that likely will be different from the true superpopulation model. When (5) is wrong, selection of the largest units produces quite biased estimates.

In the example we have compared some alternative sampling designs for selecting a preliminary subsample in a simulation. A detailed description of the simulation comparing different preliminary

sampling strategies (preliminary subsamples and preliminary estimators) are given by Righi and Tuoto (2007). In particular, we focused on the selection of balanced sampling, which allows to plan a bias-robust strategy against the model failure (Valliant *et al.*, 2000). We consider two different balanced sampling designs. The first design uses the following balancing equations:

$$\sum_{s_{ag}^t} \frac{(a_i^{t-12})^j}{n_{ag}^t} = \sum_{\tilde{s}_g^t} \frac{(a_i^{t-12})^j}{\tilde{n}_g^t} \quad (j=1, 2, \dots, J), \quad (6)$$

where \tilde{n}_g^t is the size of $\tilde{s}_g^t (= \bigcup_{h \in g} \tilde{s}_h^t)$.

The strategy defined by (3) and (6) is called as Simple Balanced (SB) strategy.

The second balanced design tries to satisfy the weighted balancing equation

$$\frac{1}{n_{ag}^t} \sum_{i \in s_{ag}^t} \frac{a_i^{t-12}}{\sqrt{a_i^{t-12}}} = \frac{\sum_{i \in \tilde{s}_g^t} a_i^{t-12}}{\sum_{i \in \tilde{s}_g^t} \sqrt{a_i^{t-12}}}, \quad (7)$$

defining a weighted balanced sample (Royall, 1992). We call this strategy as Weighted Balanced (WB) strategy. Royall (1992) and Valliant *et al.* (2000) give a deeper description of the two strategies and their properties related to the true and working superpopulation model. Here we only highlight that the imputation process (3) becomes more robust with the balanced sampling, even though the variance of the estimator increases with respect to the strategy selecting the largest sampling unit.

4.1.3 Results of the simulation

In order to carry out the simulation study, an artificial sample based on the observed final sample of 2004 has been arranged (Righi and Tuoto, 2007). The main aim of the artificial sample is to make the complete set of data available in terms of target variables and covariates, for all the 7,448 units in the final sample. Starting from the complete data set, denoted as pseudo-sample, the properties of the proposed sampling strategies have been studied in a simulative context. For each strategy 500 PTS, each one with 1,920 units, as recommended by EUROSTAT (2001), have been selected from the pseudo-sample. At each iteration the preliminary estimates are computed for the domains and the revisions are calculated comparing to the final pseudo-sample estimates.

Table 1 shows the UPOS monthly sample size distribution. We observe an average of about 2,340 units, with a maximum value equal to 2,607 and a minimum value equal to 2,068 units.

Table 1. Monthly UPOS dimensions

Month	1	2	3	4	5	6	7	8	9	10
Sample size	2,112	2,302	2,275	2,385	2,384	2,482	2,348	2,332	2,607	2,068

The experiment compares the results coming from the proposed sampling strategies, hereinafter the balanced strategies, with both the estimates based on the UPOS and the estimates obtained by the sample of 1,920 units of the largest retail enterprises in terms of turnover or number of employed persons. The largest enterprises samples may be considered as cut-off sampling (cf. “Sample Selection

– Main Module”), which is frequently used in short terms statistics. The last two samples are allocated according to the same technique defining the allocations of the balanced PTS. These three strategies represent the benchmark of the balanced strategies. The balanced samples have been selected by means of the Cube algorithm (Deville and Tillé, 2004).

For evaluating the performances in term of revision, the monthly *Mean Percentage Revision* (MPR) has been computed according to the expression

$$MPR_D^{t,0} = \sum_{d \in D} \left[\left(\frac{\tilde{I}_d^{t,0} - \hat{I}_d^{t,0}}{\hat{I}_d^{t,0}} \right) \times 100 \right] \gamma_d, \quad (8)$$

where $\hat{I}_d^{t,0}$ is the final estimate for the more disaggregated domain d (Large-non-food; Small-non-food; Large-food; Small-food) in month t , and $\tilde{I}_d^{t,0}$ assumes one of the following values:

- $(1/500) \sum_r \tilde{I}_{d,r}^{t,0}$ with the balanced strategies, $\tilde{I}_{d,r}^{t,0}$ being the provisional estimate on the domain d in the r -th replication;
- $\tilde{I}_d^{t,0} \equiv \tilde{I}_d^{t,0}$ considering the benchmark strategies.

Finally, D indicates the generic domain at the more disaggregate level ($d \equiv D$) or at aggregate level (Non-food, Food, Large, Small, Total), and $\gamma_d = \sum_{h \in d} \gamma_h$.

The yearly version of (8) is given by

$$MPR_D = \frac{1}{12} \sum_{t=13}^{24} MPR_D^{t,0}. \quad (9)$$

A second type of indicators measures the variability of the estimates by means of the *Mean Absolute Percentage Revision* (MAPR). At monthly level it is defined by

$$MAPR_D^{t,0} = \sum_{d \in D} \left[\left| \frac{\tilde{I}_d^{t,0} - \hat{I}_d^{t,0}}{\hat{I}_d^{t,0}} \right| \times 100 \right] \gamma_d. \quad (10)$$

We prefer to use the expression (10) for the balanced strategies instead of a more appropriate indicator using the term $(1/500) \sum_r \left| (\tilde{I}_{d,r}^{t,0} - \hat{I}_d^{t,0}) / \hat{I}_d^{t,0} \right| \times 100$ in the square brackets, since we observed only one preliminary sample for the benchmark strategies. Therefore, in the balanced strategies this alternative indicator catches the variability due to the iterations, not detectable in the benchmark strategies. The yearly MAPR is

$$MAPR_D = \frac{1}{12} \sum_{t=13}^{24} MAPR_D^{t,0}. \quad (11)$$

We point out that the monthly MPR and MAPR give rough measures especially for the benchmark strategies, because they are computed for few values and with only one value for the more disaggregate domains. Hence, we show the results of the statistics (9) and (11). The exhaustive description of the simulation results is given in Righi and Tuoto (2007).

Table 2.a shows the values of the statistics (9) for the preliminary domain estimates given by crossing the variables type of sold product (food and non-food retail enterprises) and type of distribution (large and small retail enterprises).

Table 2.a. Yearly Mean Percentage Revision (MPR) by Type of sold product and Type of distribution domains

Method	Large	Small	Large	Small
	non-food	non-food	food	food
Strategy using UPOS	0.674	0.236	0.505	0.021
Largest Units in terms of Employed Persons (LUEP) strategy	1.606	-0.139	-0.070	-0.592
Largest Units in terms of Turnover (LUT) strategy	0.977	0.126	0.359	-0.309
Simple Balanced (SB) strategy	0.555	0.006	0.733	-0.241
Weighted Balanced (WB) strategy	0.639	-0.077	0.197	-0.303

The table underlines that the balanced approaches using a PTS have, in general, better performances than the benchmark strategies. Especially the WB seems to be the best. The benchmark strategies present a MPR less than the WB strategy only in two cases: for large-food domain the *Largest Units in terms of Employed Persons* (LUEP) strategy has MPR = -0.070, while the WB strategy has MPR = 0.197, and for the small-food domain, where the strategy based on UPOS has MPR = 0.021, a value closer to zero than the value -0.303 of the WB strategy. The SB strategy has good performances except for the large-food domain with MPR = 0.733.

Table 2.b shows the MPR results for the aggregate domains. The findings must be analysed with caution because of the opposite signs of the MPR at the more disaggregate levels. “Good” results could actually hide an unstable strategy in term of unbiasedness, and this aspect must be taken into account in the conclusive evaluations. The WB strategy is the best method based on PTS, except for the case of small type of distribution with MPR = -0.109, while the SB strategy has MPR = -0.029. In the large domain, LUEP strategy is slightly better. Finally, for the total domain the strategy observing the LUEP has the best MPR with a value equal to -0.021. The WB strategy has MPR = 0.043.

Table 2.b. Yearly Mean Percentage Revision (MPR) by Type of sold product, Type of distribution and Total domains

Method	Type of sold product		Type of distribution		Total
	Non-food	Food	Large	Small	
Strategy using UPOS	0.293	0.396	0.540	0.205	0.334
Largest Units in terms of Employed Persons (LUEP) strategy	0.087	-0.188	0.272	-0.205	-0.021
Largest Units in terms of Turnover (LUT) strategy	0.236	0.209	0.486	0.063	0.225
Simple Balanced (SB) strategy	0.078	0.514	0.697	-0.029	0.250
Weighted Balanced (WB) strategy	0.016	0.084	0.287	-0.109	0.043

Table 3.a gives some findings about the variability of the compared strategies, computed by (11). The methods based on balanced PTS seem to have better performances, especially the WB strategy. The SB has the best MAPR for the large-non-food domain (0.998), but it has a high value for the large-food domain (0.840) with respect to some benchmark strategies. The *Largest Units in terms of Turnover* (LUT) sample strategies have the best results for small-non-food domain with MAPR = 1.001 and LUEP has MAPR = 1.143. The last strategy has the best performance also for the large-food domain (0.317). We note that when the WB has worse results than the largest units strategies, the values are close each other. On the other hand, when WB is the best strategy the MAPR is quite better than the benchmark strategies. Strategy based on UPOS, despite the greatest mean overall sample size, does not operate very well at least with respect to the balanced PTS strategies. Just for the small-non-food domain MAPR = 1.336, while the SB has MAPR = 1.369.

Table 3.a. Yearly Mean Absolute Percentage Revision (MAPR) by Type of sold product and Type of distribution domains

Method	Large	Small	Large	Small
	non-food	non-food	food	food
Strategy using UPOS	1.493	1.336	1.093	2.091
Largest Units in terms of Employed Persons (LUEP) strategy	2.263	1.143	0.317	2.949
Largest Units in terms of Turnover (LUT) strategy	2.461	1.001	0.587	2.344
Simple Balanced (SB) strategy	0.998	1.369	0.840	2.038
Weighted Balanced (WB) strategy	1.118	1.322	0.392	2.040

For the aggregate domains (Table 3.b) the use of balanced PTS still leads to the best MAPR values. In a few cases the largest units strategies achieve lower values: for the non-food domain, where the LUT strategy is better than the SB strategy (1.191 vs. 1.195) and for the large domain, where the LUEP strategy (0.715) is better than the SB approaches with MAPR values greater than 0.77.

The results in this simulation study show that the WB even though it is not always the best strategy it does appear to be the strategy with the best overall performance.

Table 3.b. Yearly Mean Absolute Percentage Revision (MAPR) by Type of sold product, Type of distribution and Total domains

Method	Type of sold product		Type of distribution		Total
	Non food	Food	Large	Small	
Strategy using UPOS	1.356	1.317	1.175	1.444	1.341
Largest Units in terms of Employed Persons (LUEP) strategy	1.288	0.909	0.715	1.403	1.139
Largest Units in terms of Turnover (LUT) strategy	1.191	0.982	0.970	1.195	1.109
Simple Balanced (SB) strategy	1.195	0.685	0.771	0.881	0.618
Weighted Balanced (WB) strategy	1.168	0.659	0.467	0.840	0.506

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Baldi, C., Ceccato, F., Congia, M. C., Cimino, E., Pacini, S., Rapiti, F., and Tuzi, D. (2003), Use of Administrative Data for Short Term Statistics on Employment, Wages and Labour Cost. *Proceedings of the “17th Roundtable on Business Survey frames”*, Rome, 26-31 October 2003. <http://www.oecd.org/dataoecd/15/62/36232440.pdf>
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*. Chapman and Hall, New York.
- D’Alò, M., De Vitiis, C., Falorsi, S., Righi, P., and Gismondi, R. (2007), Sampling Strategies for Preliminary Estimates Production in Short-Term Business Surveys. In: *Proceedings of the 2007 intermediate conference Risk and prediction*, Società Italiana di Statistica.
- De Sandro, L. and Gismondi, R. (2004), Provisional Estimation of the Italian Monthly Retail Trade Index. *Contributi-Istat*, 24/2004.
- Deville J.-C. and Tillé, Y. (2004), Efficient Balanced Sampling: the Cube Method. *Biometrika* **91**, 893–912.
- Di Fonzo, T. (2005), The OECD project on revisions analysis: First elements for discussion. Paper presented at OECD STESEG meeting, Paris 27-28 June 2005. <http://www.oecd.org/dataoecd/55/17/35010765.pdf>
- EUROSTAT (2000), *Short-term Statistics Manual*. Eurostat, Luxembourg.
- EUROSTAT (2001), Conclusion of the First Meeting of the Export Group Contro-Stratified European Sample for Retail Trade, Final Report, July 2001. Eurostat, Luxembourg.
- EUROSTAT (2005), Council Regulation No 1165/98 Amended by the Regulation No 1158/2005 of the European Parliament and of the Council – Unofficial Consolidated Version. Eurostat, Luxembourg.
- Rao, J. N. K., Srinath, K. P., and Quenneville, B. (1989), Estimation of Level and Change using Current Preliminary Data. In: *Panel Surveys* (eds. Kasprzyk, Duncan, Kalton, and Singh), John Wiley & Sons, New York, 457–485.
- Righi, P. and Tuoto, T. (2007), The planning of Preliminary Sample: methodological aspects and an application to the Italian Monthly Retail Trade Survey. In: *Rivista di Statistica Ufficiale* N. 2-3/2007.
- Royall, R. and Herson, J. (1973), Robust Estimation in Finite Population. *Journal of the American Statistical Association* **68**, 880–889.

- Royall, R. M. (1992), Robustness and Optimal Design Under Prediction Models for Finite Populations. *Survey Methodology* **18**, 179–185.
- Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Specific section

8. Purpose of the method

Selection of a preliminary subsample from the sample is used for producing preliminary or provisional estimates. Different sampling designs are suggested according to the survey context. As far as short term statistics are concerned, where the timeliness is a pressing target and the estimation is based on panel or rotating panel, the *balanced sampling design* (see the module “Sample Selection – Balanced Sampling for Multi-Way Stratification”) according to the model-based inferential framework is suggested for defining the preliminary sampling design.

9. Recommended use of the method

1. In general the method requires an intensive follow-up of the units belong to the preliminary subsample, so that the rate of quick non-response is low.
2. The method using the balanced sampling design to define the preliminary subsample exploits the time series data of the units in the panel sample.

10. Possible disadvantages of the method

1. The method has no particular theoretical disadvantage.
2. There can be operative disadvantages due to the implementation of the intensive follow-up for the preliminary subsampled units.

11. Variants of the method

1. n/a

12. Input data

1. Data including auxiliary variables for defining the sampling design.
2. In case of a balanced sampling design it is useful to collect data of the previous survey occasions for the panel units.

13. Logical preconditions

1. Missing values
 1. In practice, the method can be adapted to deal with missing values. If the non-response rate is too high it is needed to act in the estimation process.
2. Erroneous values
 1. In practice, the auxiliary variables will inevitably contain some errors. This is not ideal, but the method might still be useful in this case.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions

1.

14. Tuning parameters

1. The method needs a careful tuning phase of the parameters. The study of the time series of the sampling data (observed in the previous survey occasions) allows to define the suitable sampling design.

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. Sample membership indicator variable for the preliminary estimates is added in the input data set.

17. Properties of the output data

- 1.

18. Unit of input data suitable for the method

Processing full data set.

19. User interaction - not tool specific

- 1.

20. Logging indicators

1. No specific indicator.

21. Quality indicators of the output data

1. The main quality indicator is the revision, that is the difference between the final and the preliminary estimates.
2. In some survey occasion a simulation study such as the one described in Section 4 might also be used to obtain quality indicators.

22. Actual use of the method

1. Many NSIs base the preliminary estimates on a preliminary subsample. Because of the nature of the problem it means that an intensive follow-up is done for a subsample of the final sample.
2. Istat has used balanced sampling for the MRTS.
3. Balanced sampling that takes into account the estimation process such as in the topic has not been used yet.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Sample Selection – Main Module
2. Data Collection – Design of Data Collection Part 2: Contact Strategies
3. Weighting and Estimation – Main Module

24. Related methods described in other modules

1. Sample Selection – Balanced Sampling for Multi-Way Stratification
2. Weighting and Estimation – Preliminary Estimates with Design-Based Methods
3. Weighting and Estimation – Preliminary Estimates with Model-Based Methods

25. Mathematical techniques used by the method described in this module

1. Regression

26. GSBPM phases where the method described in this module is used

1. 4.1 Select sample

27. Tools that implement the method described in this module

1. Sampling package R
2. SAS Macro downloadable Insee site

28. Process step performed by the method

Sample planning and selection

Administrative section

29. Module code

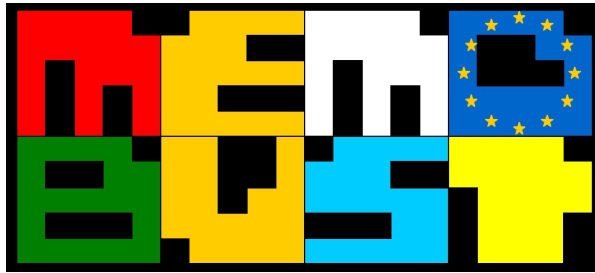
Sample Selection-M-Subsampling

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	19-03-2012	first version	Paolo Righi	ISTAT
0.2	24-04-2012	second version	Paolo Righi	ISTAT
0.3	18-05-2012	third version	Paolo Righi	ISTAT
0.4	15-04-2013	fourth version	Paolo Righi	ISTAT
0.4.1	06-09-2013	preliminary release		
0.4.2	09-09-2013	page numbering adjusted		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:42



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Sample Co-ordination

Contents

General section.....	3
1. Summary	3
2. General description.....	3
2.1 Sample co-ordination between surveys	4
2.2 Sample co-ordination over time for the same survey	4
2.3 Co-ordination of surveys based on different kind of units	5
2.4 Methods for sample co-ordination.....	5
2.5 Estimation of variances when samples are co-ordinated by using PRNs.....	5
2.6 Survey feedback into the Business Register when samples are co-ordinated	6
2.7 Negative sample co-ordination in practice	6
3. Design issues	8
3.1 Considerations before introducing sample co-ordination.....	8
3.2 Sample rotation.....	8
4. Available software tools.....	9
5. Decision tree of methods	9
6. Glossary.....	9
7. References	9
Interconnections with other modules.....	10
Administrative section.....	11

General section

1. Summary

Main objectives of sample co-ordination are to obtain comparable and coherent statistics, high precision in estimates of change over time and to spread the response burden evenly among the businesses. Sample co-ordination means to introduce dependence between two consecutive samples for the same survey or between samples for different surveys in order to minimise or maximise their overlap (number of units in common). Sample co-ordination is very useful and therefore commonly used among business surveys although sample co-ordination makes sampling and estimation more complicated (in contrast to the use of independent samples) because standard methods for sampling and estimation cannot be used in many cases. There are two main categories of methods that can be used for sample co-ordination and within each category there are a number of different methods. Many countries have implemented sample co-ordination but the specific method varies. This module includes an introduction to sample co-ordination in general, main principles to obtain different kind of co-ordination and some comments on sample co-ordination in practice.

2. General description

The overlap (number of units in common) between samples for different surveys (or between two consecutive samples for the same survey) is random when the samples are drawn independently of each other. In order to have some control over the overlap some kind of method for sample co-ordination can be used. Sample co-ordination can be used to increase the precision in estimates of change over time; the co-ordination design ensures that consecutive samples for the same survey are overlapping, although each sample is drawn from an up-to-date register. Sample co-ordination can also be used to obtain an even distribution of the response burden among the businesses¹. Businesses often perceive participating in a survey (i.e., completing questionnaires) as a burden because it takes time and effort. The opportunity to spread the response burden among the businesses and improve the precision in estimates of change makes it worthwhile to consider implementing a system for sample co-ordination. Even though sample co-ordination often means that standard methods for sampling and estimation cannot be used.

Several National Statistical Institutes (NSIs) have system and methodology implemented for sample co-ordination but system and methodology varies between the countries. The specific method implemented depends on kind of statistics to be produced, on conditions related to the production of the statistics and on what (or which) objectives of the sample co-ordination the country is focused on. For most of the methods it is easy to show that the given co-ordinated sample is a strict probability sample. However, this can be the case when there is a strong focus on response burden. This is due to the fact that strong focus on response burden often means guaranties on maximum number of surveys a specific business must participate in and/or maximum number of years a specific business must participate in the same survey. There are even sample co-ordination methods that intentionally do not give a strict probability sample in favour for the advantage to have better control over the overlap. Note that sample co-ordination introduces dependence between the obtained samples.

¹ The word “business” is used as a generic name for all unit types used in business surveys

Sample co-ordination is commonly used for business surveys but when it comes to individual and household surveys it is more unusual. This is probably related to differences between business and social statistics, differences like the very skew business population and that many countries have a business register (compared to a population register). In addition, the National Accounts has a clear need for comparable and coherent economic statistics in order to compile the Gross Domestic Product (GDP). Sample co-ordination can be obtained in two dimensions: between different surveys and over time for the same survey.

2.1 Sample co-ordination between surveys

Negative sample co-ordination between surveys means that samples for two negatively co-ordinated surveys have as few businesses in common as possible. Entirely successful negative co-ordination means of course completely separate samples. However, there are not always enough businesses to obtain complete negative co-ordination, but this sample co-ordination at least reduces the number of businesses in common. How well negative co-ordination works depends to a large extent on the size of the sample fractions in the different surveys. Note that negative sample co-ordination cannot give guaranties on maximum number of surveys a specific business must participate in when strict probability samples are used. For more information on negative co-ordination in practice, see section 2.7 further down.

Positive sample co-ordination between surveys means that samples for two positively co-ordinated surveys have as many businesses in common as possible. Positive co-ordination between surveys can be used to facilitate comparisons between variable values on the micro level even though data are collected in different surveys. However, this facility is needed most for editing purposes among large businesses and they are almost always completely enumerated and therefore included in all samples independently of co-ordination. This kind of co-ordination can also facilitate the production of comparable and coherent statistics required by (at least) the National Accounts using results from the majority of the economic surveys as building bricks when compiling the GDP. But for small businesses this kind of co-ordination could mean an unnecessary burden taking into account that coherence analysis is most focused on large businesses. And, the requirement from the National Accounts on comparable and coherent statistics can, to a large extent, be met by another kind of co-ordination, namely co-ordination of frame populations; see theme module “Statistical Registers and Frames – Survey Frames for Business Surveys” for more information.

2.2 Sample co-ordination over time for the same survey

Positive co-ordination over time for the same survey is used to obtain high precision in estimates of change over time. This can be achieved when consecutive samples for the same survey are overlapping, i.e., two consecutive samples have many businesses in common. The size of the overlap is stochastic and depends to a large extent on the sampling design, on sampling fractions as well as on changes in the business population between the two sampling occasions. Due to this positive co-ordination over time, a selected business may have to participate in a survey for many years. In order to spread the response burden among the businesses many countries have implemented some kind of method of sample rotation, for more information see section 3.2.

2.3 *Co-ordination of surveys based on different kind of units*

There is a third kind of co-ordination, not mentioned before, namely co-ordination between business surveys based on different kind of unit types. The fact that business surveys use different kind of statistical units in the BR to construct a frame population and to draw a sample implies a need for this third kind of co-ordination. There are methods for sample co-ordination that admit co-ordination between surveys based on different kind of unit types. The fact that the business population changes very fast in terms of registrations, de-registrations, mergers, split-offs, breakups and take-overs makes it a challenge to achieve a strong co-ordination lasting over time, especially when it comes to co-ordination of surveys based on different kind of units types. For more information on co-ordination of surveys based on different kind of units, see the method module “Sample Selection – Assigning Random Numbers When Co-ordination of Surveys Based on Different Unit Types is Considered”.

2.4 *Methods for sample co-ordination*

There are in principle two categories of sample co-ordination methods:

- 1) Sample co-ordination methods based on Permanent Random Numbers (PRNs)
- 2) Sample co-ordination methods *not* based on PRNs

The most common method to obtain sample co-ordination is based on the use of PRN. The basic idea is to associate an independent and unique random number, uniformly distributed over the interval (0,1), with every unit in the register. For every unit persisting in the register the same random number is used on each sampling occasion. In this way we always get a new sample from the updated register but a large overlap with the latest sample can be expected. Every new unit (births) is assigned a new random number while closed-down units (deaths) are withdrawn from the register with their random numbers.

Several countries use sample co-ordination based on PRNs and there exists many variations of this method. In the method modules listed below two different methods for sample co-ordination (based on PRN) are presented:

- “Sample Selection – Sample Co-ordination Using Simple Random Sampling with Permanent Random Numbers”
- “Sample Selection – Sample Co-ordination Using Poisson Sampling with Permanent Random Numbers”

Sample co-ordination methods *not* based on PRNs are in general not used in the NSIs. Nevertheless, one of those methods, which is based on linear programming, can be applied in business surveys (Reiss et al., 2003). A main feature with methods based on linear programming is the possibility to optimise (maximise or minimise) the overlap between two samples, two consecutive samples for the same survey or two samples for different surveys. See Ernst (1996, 1998, 1999 and 2002) for a more general description of non-PRN methods.

2.5 *Estimation of variances when samples are co-ordinated by using PRNs*

In the area of economic statistics measures of change are key parameters and it is of great importance to be able to determine whether an observed change is statistically significant or not. Sample co-ordination by PRNs makes the level estimates correlated and this correlation is quite complicated to

estimate because the use of PRNs brings an additional component of randomness to the rotation pattern (compared to ordinary panel design). The problem of estimating the variance for measures of change when samples are co-ordinated by PRNs has been addressed by several persons during the years. However, a complete and workable method for estimating this correlation under the Statistics Sweden sampling method (co-ordinated SRS and stratified SRS based on PRN) was developed in the late 1990's, see Nordberg (2000). This approach can hopefully be of interest in the context of other PRN systems. For more information on variance estimation, see theme module "Quality Aspects – Quality of Statistics".

2.6 Survey feedback into the Business Register when samples are co-ordinated

Survey feedback means that information obtained from a survey is used to update the Business Register (BR). It is not advisable to use survey feedback from co-ordinated sample surveys, especially not where positive co-ordination over time is used. This applies mainly to variables used in the survey design, variables like economic activity, number of employees and annual turnover. Information on contact variables is not equally sensitive to survey feedback.

Survey feedback implies that businesses in the BR, which are included in samples, are updated, while those not included are not updated. Furthermore, the large overlap between two consecutive samples means that the latest sample is based on a BR-version where businesses included in the previous sample are more updated compared to the rest of the businesses in the BR. Survey feedback will in this case lead to bias in the estimates because the sample is no longer representative for the whole frame population. But from this point of view it is all right to update large businesses with survey feedback because they are, almost always, completely enumerated. Nevertheless, in practice there is a strong desire to be able to use all information collected from surveys to update business information in the BR. During the years work has been done in order to estimate the magnitude of the introduced bias as well as on methods to reduce this bias in the estimation phase. However, at the moment there is no complete and workable method to recommend. In general, feedback from co-ordinated sample surveys should rather be used as quality indicators in the maintenance of the BR.

2.7 Negative sample co-ordination in practice

In recent years there has been a strong focus on lowering the response burden in all EU-countries. There is a challenge in reducing the response burden without negatively affecting the quality of the desired estimates. Sample co-ordination is a successful method only if the number of businesses among which the response burden is spread is sufficient. This is not always the case because the structure of the business population is generally very skewed, consisting of a huge number of small enterprises and a very small number of medium and large-sized enterprises (size in terms of persons employed or other economic variables). Table 1, below, shows the structure of the business population in the EU-countries (source: *Key figures on European business with a special feature on SMEs*). The division into size classes is based on number of persons employed.

Table 1. Total number of enterprises and their distribution by size class among EU countries

	Total	Distribution of enterprises by size class			
	number of	Micro	Small	Medium	Large
	enterprises	< 10	10- <50	50- <250	250-
	(thousands)	%	%	%	%
EU-27	20 994	92,0	6,7	1,1	0,2
Belgium	426	92,5	6,3	0,9	0,2
Bulgaria	270	88,7	9,2	1,9	0,3
Czech Republic	899	95,1	3,9	0,8	0,2
Denmark	211	85,0	12,2	2,4	0,4
Germany	1 880	83,0	14,1	2,4	0,5
Estonia	46	83,9	13,0	2,7	0,4
Ireland	158	87,8	9,9	1,9	0,3
Greece	:	:	:	:	:
Spain	2 653	93,1	6,0	0,8	0,1
France	:	:	:	:	:
Italy	3 947	94,3	5,1	0,5	0,1
Cyprus	47	92,3	6,4	1,1	0,2
Latvia	70	84,4	12,9	2,4	0,3
Lithuania	139	88,7	9,2	1,9	0,3
Luxembourg	17	85,8	11,5	2,2	0,5
Hungary	566	94,3	4,7	0,8	0,2
Malta	:	:	:	:	:
Netherlands	583	90,4	8,0	1,4	0,3
Austria	294	87,2	10,8	1,7	0,4
Poland	1 556	95,5	3,3	1,0	0,2
Portugal	778	94,0	5,1	0,7	0,1
Romania	506	88,9	8,8	1,9	0,4
Slovenia	93	92,4	6,1	1,3	0,3
Slovakia	59	71,2	24,2	3,7	0,9
Finland	202	91,7	6,9	1,1	0,3
Sweden	586	94,7	4,4	0,8	0,2
United Kingdom	1 731	89,3	8,8	1,5	0,4

Table 1 shows that all EU-countries have a skewed business population. Countries with a relatively small total number of businesses often meet a more difficult situation when it comes to spreading the response burden compared to countries with a large total number of businesses. This is due to the fact that quality (in terms of standard errors) in desired estimates is mainly correlated with the sample size (and not with the population size). And, even though total number of businesses is large a detailed stratification can lead to large sampling fractions in specific strata. On the other hand, detailed stratification is often needed in order to produce detailed domain estimates of high quality.

Negative co-ordination is a very effective tool to spread the response burden among small businesses. This is important because they often do not have the capacity to participate in many surveys. However, there is little room for spreading the response burden among medium-sized businesses because these businesses are few, see table 1. In addition, they have a proportionately large impact on the estimates in terms of economic variables and they must therefore often be included in samples. Medium-sized businesses could meet a heavy burden, especially in industries with few businesses. The small number of large businesses, see table 1, are of great importance for the economic statistics because they have a

large impact on the estimates in terms of economic variables. Therefore it is crucial, with few exceptions, to include all large sized businesses belonging to the frame population for a specific survey. Otherwise it would be more or less impossible to publish the survey results. The work on response burden regarding large sized businesses must mainly focus on simplifying for the respondents to supply the requested information.

The possibility of spreading the response burden depends of course, to a large extent, on the structure of the business population in a specific country. The structure is almost always skewed but if the total number of businesses (all categories) is considerably large the possibility for successful negative sample co-ordination increases.

3. Design issues

3.1 Considerations before introducing sample co-ordination

It is very important to consider some kind of optimal co-ordination (negative as well as positive) between surveys before the sample co-ordination is introduced into a system. Once a survey is placed into the system it is preferable to keep the survey in the same place because of the desired overlap between two consecutive samples for the survey. There are mechanical, as well as manual, methods that can be used to place surveys into a system. However, a number of well-considered decisions must be taken to obtain the best solution in the long run. This could be considerations like:

- optimal co-ordination between the surveys, how are the surveys related to each other
- economic activities and size classes covered by the survey
- periodicity of the survey (monthly, quarterly, annual, periodic)
- time point when questionnaires for different surveys are sent out
- survey content, in terms of number and type of variables

3.2 Sample rotation

As mentioned before, due to the positive co-ordination over time, a selected business may have to participate in a survey for many years. In order to spread the response burden among the businesses it is possible to implement some kind of sample rotation into the sample co-ordination system. The objective of this sample rotation is to keep a selected business in the sample for a pre-specified number of years and then let it rotate out of the sample (in contrast to the stochastic rotation obtained by changes in the business population). There are several methods to obtain sample rotation and Ohlsson (1995) gives a description of some of the methods. As mentioned before, a sample rotation method cannot give guaranties on maximum number of years a specific business has to participate in the same survey when strict probability samples are used.

The number of years a business should participate in a survey is a balance between response burden and the decrease in the precision of the estimates of change over time that is acceptable. And in practice, rotation works only successfully if there is room for rotation. And by successful rotation is meant that a business can rotate out of a sample after the pre-specified number of years without immediately rotating into the sample of another survey. Such room is only available among small businesses where the sampling fraction is small. It takes longer time for businesses in stratum with

larger sampling fraction to rotate out of the sample. How long depends to a large extent on the size of the sampling fraction.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Ernst, L. R. (1996), Maximizing the overlap of sample units for two designs with simultaneous selection. *Journal of Official Statistics* **12**, 33–45.
- Ernst, L. R. (1998), Maximizing and minimizing overlap when selecting a large number of units per stratum simultaneously for two designs. *Journal of Official Statistics* **14**, 297–314.
- Ernst, L. R. (1999), The maximization and minimization of sample overlap problems: a half century of results. In: *Proceedings of the International Statistical Institute*, 52nd Session, Finland, 168–182.
- Ernst, L. R. and Paben, S. P. (2002), Maximizing and minimizing overlap when selecting any number of units per stratum simultaneously for two designs with different stratifications. *Journal of Official Statistics* **18**, 185–202.
- Key figures on European business with a special feature on SMEs*. Eurostat pocketbooks, ISSN 1830-9720.
- Nordberg, L. (2000), On Variance Estimation for Measures of Change When Samples are Coordinated by the Use of Permanent Random Numbers. *Journal of Official Statistics* **16**, 363–378.
- Ohlsson, E. (1995), Coordination of samples using permanent random numbers. In: *Business Survey Methods* (eds. Cox, B. G., Binder, D. A., Chinnapa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S.), Wiley, New York, Chapter 9, 153–169.
- Reiss, P., Schiopu-Kratina, I., and Mach, L. (2003), The use of the transportation problem in coordinating the selection of samples for business surveys. In: *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Annual Meeting, June 2003.

Interconnections with other modules

8. Related themes described in other modules

1. Repeated Surveys – Repeated Surveys
2. Statistical Registers and Frames – Survey Frames for Business Surveys
3. Weighting and Estimation – Main Module
4. Quality Aspects – Quality of Statistics

9. Methods explicitly referred to in this module

1. Sample Selection – Sample Co-ordination Using Simple Random Sampling with Permanent Random Numbers
2. Sample Selection – Sample Co-ordination Using Poisson Sampling with Permanent Random Numbers
3. Sample Selection – Assigning Random Numbers When Co-ordination of Surveys Based on Different Unit Types is Considered

10. Mathematical techniques explicitly referred to in this module

- 1.

11. GSBPM phases explicitly referred to in this module

- 1.

12. Tools explicitly referred to in this module

- 1.

13. Process steps explicitly referred to in this module

- 1.

Administrative section

14. Module code

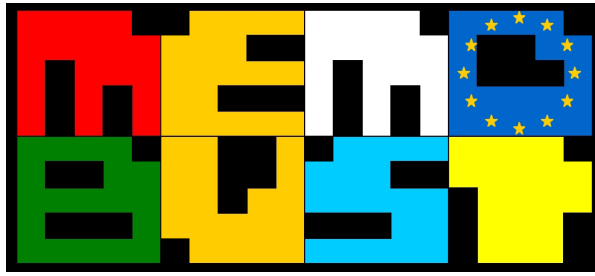
Sample Selection-T-Sample Co-ordination

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-02-2013	first version	Annika Lindblom	Statistics Sweden
0.2	30-04-2013	improvements based on the Norwegian and Swiss reviews	Annika Lindblom	Statistics Sweden
0.3	29-05-2013	improvements based on the Norwegian and Swiss reviews	Annika Lindblom	Statistics Sweden
0.3.1	18-09-2013	preliminary release		
0.4	27-09-2013	improvements based on the EB-review	Annika Lindblom	Statistics Sweden
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:43



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Sample Co-ordination Using Simple Random Sampling with Permanent Random Numbers

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Sequential simple random sampling.....	3
2.2 The “JALES” method.....	3
2.3 The Cotton & Hesse method	4
2.4 Co-ordination of stratified samples	4
2.5 Co-ordination of surveys based on different unit types.....	5
2.6 Rotation	5
3. Preparatory phase	5
4. Examples – not tool specific.....	5
4.1 Negative co-ordination with the “JALES” method of two surveys using two different starting points	5
4.2 Positive co-ordination with the “JALES” method of a single survey over time	6
4.3 Negative co-ordination over time with the Cotton and Hesse method.....	7
5. Examples – tool specific.....	8
6. Glossary.....	8
7. References	9
Specific section.....	10
Interconnections with other modules.....	11
Administrative section.....	13

General section

1. Summary

The purpose of this module is to introduce the reader to the Permanent Random Number (PRN) technique for co-ordinating (stratified) simple random samples. The concept of co-ordination based on PRNs using simple random sampling was introduced in Sweden in the early 70s (see, e.g., Ohlsson 1992). This technique provides a practical solution for controlling the overlap (the number of common units) between different samples.

The methods developed and used by Statistics Sweden (Ohlsson, 1992) and the Institut National de la Statistique et des Études Économiques (INSEE) of France (Cotton and Hesse, 1992) are both methods of co-ordination of simple random samples using PRNs and will be briefly described.

2. General description of the method

The National Statistical Institutes (NSIs) publish economic statistics based on business surveys. Often, different surveys use the same sampling frame (a register or a list frame) which leads to a need to co-ordinate sampling for the surveys. We distinguish two main kinds of co-ordination: positive and negative co-ordination which have different aims, i.e., maximising or minimising the overlap between samples. The general aspects of the sample co-ordination problem are described in detail in the theme module “Sample Selection – Sample Co-ordination”.

With the PRN technique for co-ordinating samples both positive and negative co-ordination can be obtained. Several national agencies use variations of this technique. A review of basic PRN techniques and a comparison by countries can be found in Ohlsson (1995) and Hesse (1999).

2.1 Sequential simple random sampling

Consider a population of size N . This population could be also a stratum, the principle is the same. Each unit in the population is assigned a random number, uniformly distributed over the interval $[0,1]$. The units are sorted in ascending order of the random numbers. The sample is composed of the first n units in the ordered list. This technique is described in Fan et al. (1962) who called it “sequential”, thus the name *sequential simple random sampling*. Ohlsson (1992) shows that this technique produces a simple random sampling without replacement (*srswor*). Due to the symmetry of the uniform distribution, the selection of the last n units in the ordered list of units also gives a sequential *srswor*. More precisely, selecting the first n units to the left, or to the right, of any fixed point a in $[0,1]$ also yields a *srswor*. If there are not enough units to the right (or left) of the starting point a , then, depending on the chosen direction, the selection can continue to the right (or left) of the point 0 (the point 1). Thus, the sampling frame can be viewed as a circular list.

2.2 The “JALES” method

The Swedish SAMU system uses sequential *srswor* to co-ordinate samples across surveys and over time (SAMU is an acronym for “co-ordinated samples” in Swedish). The used method is referred to as “JALES”. A full description of the system can be found in Ohlsson (1992) and Lindblom (2003). Following Ohlsson (1995), we present a brief overview of the method.

A PRN, uniformly distributed over the interval $[0,1]$, is associated with each unit in the frame. Units who stay in the frame have the same random number on each sampling occasion. Every new business in the frame (a birth) is assigned a new PRN while closed-down businesses (deaths) are withdrawn from the sampling frame together with their assigned PRN.

Co-ordination over time is done in the following way. On each sampling occasion a new sequential *srswor* using PRNs is drawn from the updated (for births and deaths) frame. However, a large overlap with the previous sample can be expected, since persistent businesses have the same PRN on each occasion. This type of co-ordination enables good precision in estimates of change over time.

In order to co-ordinate samples with desired sample sizes n_1 and n_2 for two different surveys, we choose two constants a_1 and a_2 in $[0,1]$. Then we take the units with the n_1 PRNs closest to the right (or left) of a_1 to obtain the first sample and the n_2 units closest to the right or left of a_2 to obtain the second sample. Positive co-ordination of two surveys is obtained by using the same starting point and direction for both surveys. Negative co-ordination can be achieved if the starting points and sampling directions are properly chosen, e.g., different starting points and the same direction.

2.3 *The Cotton & Hesse method*

The Cotton & Hesse method is a PRN method based on sequential *srswor*. It is fully described in Cotton and Hesse (1992). The method, which allows only for negative co-ordination, could be used for co-ordinating samples for different surveys drawn the same year or for co-ordinating samples for a single survey over time. Each unit of the population receives a PRN from a uniform distribution $U[0,1]$. Then a sequential *srswor* of size n is drawn choosing the units with the n smallest random numbers (starting point $a = 0$). Negative co-ordination is obtained by permutation of the random numbers. The reordering is done in such a way that the selected units receive the largest PRNs and the non-selected units receive the smallest PRNs. Within the two subsets of selected and non-selected units, the order of the permuted PRNs must remain unchanged. This means that if a unit was assigned the smallest of the first n ordered PRNs, after reordering it will be assigned the smallest of the n ordered largest PRNs. Then a sequential *srswor* is drawn in the reordered list of units.

Note that, after permutation the PRNs remain independent uniform random numbers. So, the successive samplings remain *srswor*. As for SAMU, a closed-down unit loses its PRN and a new unit receives a new PRN. An interesting property of the method is that the minimum of the expected overlap between two successive simple random samples is obtained. The joint sampling design of each pair of subsequent samples is the same as this obtained by sequential *srswor* when the same starting point and opposite directions are chosen (Hesse, 1999).

2.4 *Co-ordination of stratified samples*

Many business surveys use stratified *srswor* in order to improve the accuracy of estimates by dividing the frame population into homogenous sub-populations (strata). The PRN technique for co-ordination can be easily adapted to this environment. In SAMU a sequential *srswor* is drawn in each stratum. For a specific survey, if the same direction and starting point are used in all strata, then we obtain positive co-ordination. As before, negative co-ordination for two surveys can be obtained by choosing different starting points or directions. SAMU allows for positive or negative co-ordination when different stratifications are used. The Cotton and Hesse method is easily adapted to stratified *srswor*. It also allows for different stratifications (Cotton and Hesse, 1992).

2.5 Co-ordination of surveys based on different unit types

The methods of co-ordination of simple random samples using PRNs can be used for surveys based on different unit types, e.g., establishments and enterprises. The module “Sample Selection – Assigning Random Numbers When Co-ordination of Surveys Based on Different Unit Types is Considered” explains in more detail how random numbers can be assigned to the different unit types and how co-ordination is done. This kind of co-ordination has been implemented and used in SAMU (Lindblom, 2003). The method of Cotton and Hesse also allows for co-ordinating sampling units belonging to different levels (Hesse, 1999).

2.6 Rotation

The main purpose of rotation is to spread the response burden among small businesses. Rotation is meaningful mainly for units with small inclusion probabilities. Sample rotation ensures that a unit will rotate out of the sample for a certain number of years and will not be included immediately in the sample of another survey. Using methods based on PRNs, we can handle sample rotation quite easily.

Let us assume, for example, that a sample should be rotated every year and persisting units with inclusion probabilities of less than 0.10 should leave the sample after five years. The Random Rotation Cohort (RRC) method which can deal with this situation was introduced in the Swedish SAMU in 1989 (see, e.g., Ohlsson, 1992; Lindblom, 2003). The principle is to designate randomly and permanently each unit in the frame to one of five rotation cohorts (groups). The random numbers are then shifted by 0.10 only in one rotation group each year. After the first year all units in rotation group one will shift PRNs 0.10 to the left. The second year, the PRNs in rotation group two are shifted 0.10 to the left and so on. Units with inclusion probabilities less than 0.10 can be expected to be out of sample after five years. Thus we achieve an expected rotation rate of 1/5. Units with larger inclusion probabilities are also rotated but it takes more time for them, depending on the size of the inclusion probability, to leave the sample. With the RRC method the sampling procedure remains sequential *srswor* which ensures the positive and negative co-ordination between surveys.

3. Preparatory phase

4. Examples – not tool specific

4.1 Negative co-ordination with the “JALES” method of two surveys using two different starting points

A simple example which illustrates how negative co-ordination for two samples works will be given in the following table. Suppose we have a population of 10 units. Each unit has been assigned a PRN drawn independently from $U[0,1]$. The desired sample size for both samples is equal to 6. In order to reduce the overlap between the two samples we choose two different starting points, $a_1=0$ and $a_2=0.6$, and the same direction (right of the starting point). The units are ordered in ascending order of their PRNs and then the first 6 units starting from $a_1=0$ are selected to compose the first sample, and the first 6 units starting from $a_2=0.6$ are selected to compose the second sample. Thus the first sample, S_1 , is composed of units {5, 8, 1, 2, 4, 3} and the second sample, S_2 , is composed of units {3, 10, 6, 7, 9, 5}. Units 3 and 5 are in both samples.

Unit number	PRN	Ordered list of the PRNs	Units after ordering	Sample 1	Sample 2
1	0.25	0.1	5	x	x
2	0.4	0.2	8	x	
3	0.7	0.25	1	x	
4	0.5	0.4	2	x	
5	0.1	0.5	4	x	
6	0.8	0.7	3	x	x
7	0.9	0.75	10		x
8	0.2	0.8	6		x
9	1	0.9	7		x
10	0.75	1.0	9		x

4.2 Positive co-ordination with the “JALES” method of a single survey over time

In this example we illustrate how positive co-ordination with the “Jales” method works for a single survey over time, the sampling frame being updated between times $t=1$ and $t=2$. We suppose that we have a population of 10 units at time $t=1$. Each unit has been assigned a PRN drawn independently from $U[0,1]$. The desired sample size is equal to 6. The units are ordered in ascending order of their PRNs and then the first 6 units starting from $a_I=0$ are selected to compose the sample at time $t=1$. Thus the sample, S^1 , is composed of units $\{5, 8, 1, 2, 4, 3\}$. For simplification, we suppose that the population at time $t=2$ has the same number of units, thus we have as many births as deaths in the population. Deaths are denoted by - and births by +. The births receive a new PRN. The units are ordered in ascending order of their PRNs and the first 6 units starting from $a_I=0$ are selected. Thus the sample at time $t=2$, S^2 , is composed of units $\{5, 11, 13, 2, 4, 12\}$. Units 2, 4 and 5, which are persistent units in the sampling frame, are in the sample at both times $t=1$ and $t=2$.

Unit	PRN	Ordered PRNs	Unit	Sample t=1	Births/Deaths	Unit	PRN	Ordered PRNs	Unit	Sample t=2
1	0.25	0.1	5	x	-	1		0.1	5	x
2	0.4	0.2	8	x		2	0.4	0.18	11	x
3	0.7	0.25	1	x	-	3		0.28	13	x
4	0.5	0.4	2	x		4	0.5	0.4	2	x
5	0.1	0.5	4	x		5	0.1	0.5	4	x
6	0.8	0.7	3	x		6	0.8	0.58	12	x
7	0.9	0.75	10			7	0.9	0.75	7	
8	0.2	0.8	6		-	8		0.8	8	
9	1	0.9	7			9	1	0.9	9	
10	0.75	1.0	9			10	0.75	1	10	
					+	11	0.18		11	
					+	12	0.58		12	
					+	13	0.28		13	

4.3 Negative co-ordination over time with the Cotton and Hesse method

In this simple example, we consider co-ordinated sampling over time of three samples of size 3 in a population U of size 5. The initial PRNs are 0.5, 0.2, 0.4, 0.6, 0.8. The table below shows how the algorithm is executed at times $t = 1, 2$ and 3 and what the selected samples are. At time $t=1$, we order the units by ascending PRN. We select the sample by taking the first three units in the ordered list of units. Thus, $S^1 = \{2,3,1\}$. Next, at time $t=2$, the PRNs are reassigned to the units of the population in the following way: Units 2, 3 and 1, which had the smallest three PRNs at $t=1$ receive the largest PRNs, by respecting the ranks of the PRNs, respectively 0.5, 0.6 and 0.8, which are the largest three PRNs in ascending order. Units 4 and 5, which were not selected at time $t=1$, receive the smallest PRNs, respectively, 0.2 and 0.4. PRN 0.2 (the smallest of 0.2 and 0.4) is assigned to unit 4 which had a PRN equal to 0.6 at time $t=1$ (the smallest of 0.6 and 0.8). Then the units are ordered by ascending order of their new PRNs and the 3 units with the smallest new PRNs are selected to form the sample at time $t=2$. Thus, $S^2 = \{4,5,2\}$. Following the same procedure, we obtain the sample $S^3 = \{3,1,4\}$. Thus, we obtain a minimum overlap between S^1 and S^2 (unit 1) and between S^2 and S^3 (unit 4).

t=1

Unit	PRNs t=1
1	0.5
2	0.2
3	0.4
4	0.6
5	0.8

t=2

Units after ordering at t=1	New PRN t=2
2	0.5
3	0.6
1	0.8
4	0.2
5	0.4

t=3

Units after ordering at t=2	New PRN t=3
4	0.5
5	0.6
2	0.8
3	0.2
1	0.4

Ordered PRNs at time t=1	Units after ordering at t=1	Sample at time t=1
0.2	2	x
0.4	3	x
0.5	1	x
0.6	4	
0.8	5	

Ordered PRNs at time t=2	Units after ordering at t=2	Sample at time t=2
0.2	4	x
0.4	5	x
0.5	2	x
0.6	3	
0.8	1	

Ordered PRNs at time t=3	Units after ordering at t=3	Sample at time t=3
0.2	3	x
0.4	1	x
0.5	4	x
0.6	5	
0.8	2	

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Cotton, F. and Hesse, C. (1992), Co-ordinated selection of stratified samples. *Proceedings of Statistics Canada Symposium 92*, 47–54.
- Fan, C., Muller, M., and Rezucha, I. (1962), Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association* **57**, 387–402.
- Hesse, C. (1999), Sampling co-ordination: A review by country. Technical Report E9908, Direction des Statistique d'Entreprises, INSEE, Paris.
- Lindblom, A. (2003), SAMU - The system for coordination of frame populations and samples from the Business Register at Statistics Sweden. Background Facts on Economic Statistics 2003:3, Statistics Sweden.
- Ohlsson, E. (1992), SAMU. The system for Co-ordination of samples from the Business Register at Statistics Sweden. R&D report 1992:18, Statistics Sweden.
- Ohlsson, E. (1995), Coordination of samples using permanent random numbers. In: *Business Survey Methods* (eds. Cox, B. G., Binder, D. A., Chinnapa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S.), Wiley, New York, Chapter 9, 153–169.

Specific section

8. Purpose of the method

The purpose of the method is to allow for positive or negative co-ordination of (stratified) simple random samples, when co-ordination of the same survey over time is desired or co-ordination between different surveys.

9. Recommended use of the method

1. To obtain positive co-ordination of samples over time for the same survey.
2. Co-ordination of samples for different surveys.
3. Co-ordination of samples with different designs.
4. Co-ordination of surveys based on different unit types.

10. Possible disadvantages of the method

1. Information from the surveys should not be used to update the list frame, as this could introduce bias in the estimates.
2. If they are different levels in the surveys, e.g., enterprise and establishment levels, the co-ordination of samples selected at different levels is less efficient for multiple-location enterprises, unless clustered (or two-stage) sampling of local units (establishments) is used.

11. Variants of the method

1. Rotation of samples is obtained by shifting the permanent random numbers of the sampling units which are divided into rotation groups.

12. Input data

1. Ds_input1: Sampling frame, essentially a business register.
2. Ds_input2: Auxiliary information, i.e., size measures for stratification can come from additional sources, e.g., an administrative register.

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. Starting points and directions for co-ordination.

15. Recommended use of the individual variants of the method

1. Sample rotation.

16. Output data

1. Ds_output1: Selected co-ordinated sample for the current sampling occasion.

17. Properties of the output data

1. The selected sample is selected with a sequential *srswor*, thus is of fixed sample size.

18. Unit of input data suitable for the method

Incremental processing of frame units is possible since the units are treated independently.

19. User interaction - not tool specific

1. Delimitation of the sampling frame.
2. Determination of co-ordination rules, e.g., negative or positive co-ordination.
3. Determination of stratification and allocation.

20. Logging indicators

- 1.

21. Quality indicators of the output data

1. Number of repeated selections of an enterprise as a measure of response burden.
2. Size of the overlap between the current survey and the previous surveys.

22. Actual use of the method

1. The PRN technique for co-ordinating (stratified) simple random samples is actually used in the system for co-ordinating business surveys (SAMU) of Statistics Sweden (Lindblom, 2003) and in INSEE (Hesse, 1999). Other countries also use methods based on sequential *srswor*.

Interconnections with other modules**23. Themes that refer explicitly to this module**

1. Sample Selection – Sample Co-ordination

24. Related methods described in other modules

1. Sample Selection – Sample Co-ordination Using Poisson Sampling with Permanent Random Numbers

2. Sample Selection – Assigning Random Numbers When Co-ordination of Surveys Based on Different Unit Types is Considered

25. Mathematical techniques used by the method described in this module

- 1.

26. GSBPM phases where the method described in this module is used

- 1.

27. Tools that implement the method described in this module

1. SAMU (Ohlsson,1992)

28. Process step performed by the method

Sample selection

Administrative section

29. Module code

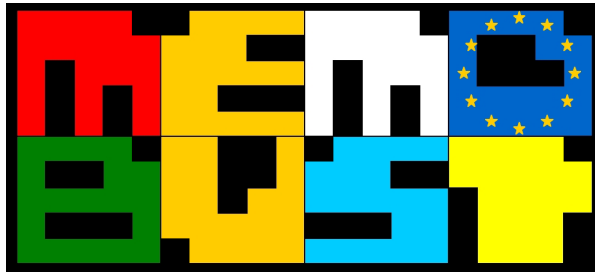
Sample Selection-M-PRN Using Simple Random Sampling

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	03-05-2013	first version	Desislava Nedyalkova	SFSO (Switzerland)
0.2	07-06-2013		Desislava Nedyalkova	SFSO (Switzerland)
0.3	17-06-2013		Desislava Nedyalkova	SFSO (Switzerland)
0.3.1	18-06-2013		Desislava Nedyalkova	SFSO (Switzerland)
0.3.2	18-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:43



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Sample Co-ordination Using Poisson Sampling with Permanent Random Numbers

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
3. Preparatory phase	6
3.1 Updating the sampling frame	6
3.2 Computing inclusion probabilities for the new survey.....	6
3.3 Choosing co-ordination rules between the new survey and previous surveys	7
4. Examples – not tool specific.....	7
4.1 One-occasion business surveys, panels and rotating panels in Switzerland.....	7
5. Examples – tool specific.....	7
6. Glossary.....	7
7. References	7
Specific section.....	9
Interconnections with other modules.....	11
Administrative section.....	13

General section

1. Summary

The method described here allows for the selection of positively or negatively co-ordinated samples, that is to say samples with large or on the contrary small overlap with previous samples. Negative co-ordination is particularly useful in order to spread the response burden evenly on the population by avoiding that the same units get selected in different samples when this is not necessary. Positive co-ordination is desirable when one wants to update a panel sample, or to be able to compare accurately results of a new survey with those of a previous survey. The proposed method is suitable for unequal inclusion probability surveys and dynamic populations with births, deaths, mergers and splits of units.

2. General description of the method

This is an extension of Brewer et al. (1972)'s method for the selection of two positively or negatively co-ordinated samples to the case of an indefinite number of surveys. It belongs to the family of Permanent Random Numbers (PRN) methods for selecting co-ordinated samples. These methods rely on the generation of random numbers for each unit that enters the population and on the principle that these random numbers govern the selection of future samples. In the method presented here, a unit is attached to its random number until it quits the population, so that the random numbers are actually permanent. In other methods, such as in Cotton and Hesse (1992) and Rivière (2001), random numbers are rotated among units after each sample selection, so that random numbers do indeed govern future sample selections, but they are not permanently attached to a given unit of the population. An account on sample co-ordination methods with PRN can be found in Ohlsson (1995),

In the method presented here, each unit of the population is dealt with independently from the others, so that obtained transversal samples, for each sampling occasion, are Poisson samples (see for example Tillé, 2006). An extensive description of the core of the method is available in Qualité (2009), and a shorter one in this document and in Qualité (2011). It can be summarised in the few following statements.

Each unit k of the population receives a permanent random number u_k generated uniformly in $(0,1)$ and independently from the random numbers of other units. If π_k^t denotes the inclusion probability of unit k at a given sampling occasion t , then unit k is selected in the sample if and only if its random number lies in a chosen subset of $(0,1)$ that has a total length equal to π_k^t . This ensures that the probability of selecting unit k in this sample is equal to π_k^t , as long as the choice of this subset is made without information on u_k . The co-ordination between surveys is obtained by a careful choice of the different selection subsets for all sampling occasions. A maximal positive co-ordination between two surveys is obtained when the corresponding selection subsets have the largest possible overlap. A maximal negative co-ordination is obtained when selection subsets are non-overlapping, if possible, or have the smallest possible overlap. The construction of these selection subsets for all sampling occasions is thus the main point of the co-ordination method.

This description, and the method used at the Swiss Federal Statistical Office (SFSO) to construct the selection intervals are better understood recursively and on an example. Since each unit of the

population is dealt with independently from the others, it is sufficient to examine what may happen for a generic unit k .

Suppose that this unit k has inclusion probabilities π_k^1 , at the first sampling occasion, π_k^2 at the second sampling occasion and π_k^3 at the third sampling occasion.

1. At the first sampling occasion, the selection set is naturally defined to be $(0, \pi_k^1)$, for all k (see figure 1). This selection set has a correct length of π_k^1 , and the probability that u_k is in this set, and thus that unit k is selected in the first sample is equal to π_k^1 .

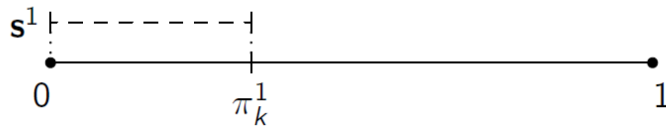


Figure 1. First sampling occasion

2. The second survey may be either positively co-ordinated or negatively co-ordinated with the first survey. We consider only maximal co-ordination. If the desired co-ordination is positive, the selection subset for the second survey is defined as $(0, \pi_k^2)$, in figure 2. If u_k is in $(0, \min(\pi_k^1, \pi_k^2))$, unit k is selected in both samples. If it is in $(\min(\pi_k^1, \pi_k^2), \max(\pi_k^1, \pi_k^2))$, then unit k is selected in one sample but not the other, and if it is in $(\max(\pi_k^1, \pi_k^2), 1)$, then unit k is selected in neither sample. Consequently, the probability of selecting unit k in both samples is maximal and equal to $\min(\pi_k^1, \pi_k^2)$, the minimum of π_k^1 and π_k^2 .

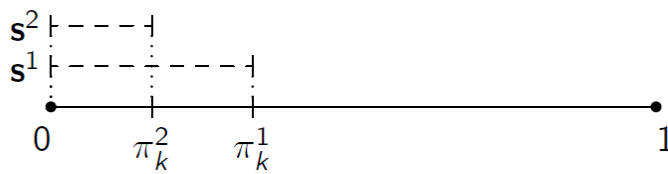


Figure 2. Second sampling occasion, positive co-ordination

If, on the contrary, the desired selection in negative, then two cases may occur:

- a. If $\pi_k^1 + \pi_k^2 \leq 1$, then the selection subset for the second survey is defined as $(\pi_k^1, \pi_k^1 + \pi_k^2)$, in figure 3. Unit k is selected in at most one of the two samples.

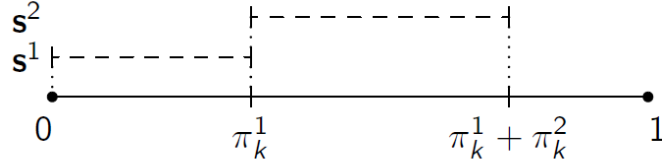


Figure 3. Second sampling occasion, negative co-ordination, first case

- b. If $\pi_k^1 + \pi_k^2 > 1$, then the selection subset for the second survey is defined as $(\pi_k^1, 1) \cup (0, \pi_k^1 + \pi_k^2 - 1)$, in figure 4. Unit k can be selected in both samples, with probability $\pi_k^1 + \pi_k^2 - 1$, which is the theoretical minimal bound in this case.

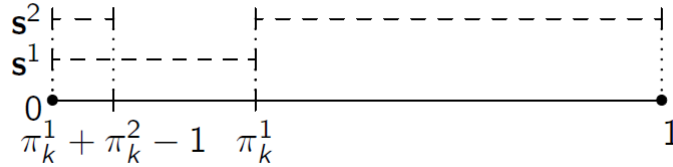


Figure 4. Second sampling occasion, negative co-ordination, second case

3. The third survey may be positively or negatively co-ordinated with the first or the second survey. The method described here allows to mix positive co-ordination with some surveys and negative co-ordination with others, but requires an order of priority for these co-ordinations. Exploring all possibilities for the third sampling occasion would be extremely tedious, but it is enough to investigate a simple example in order to understand the general idea. Suppose that the two first surveys were positively co-ordinated, so that the situation is that of figure 2, and that the third survey must be positively co-ordinated with the second, and, with a lesser priority, negatively co-ordinated with the first. Suppose moreover that π_k^3 is larger than π_k^2 . The first objective is obtained by choosing a selection subset that overlaps the most the selection set of the second survey. As π_k^3 is larger than π_k^2 , the whole subset $(0, \pi_k^2)$ is included into the selection subset of the third survey. Then, it needs to be completed with an additional subset of length $\pi_k^3 - \pi_k^2$ that respects the most the remaining co-ordination rules: this additional subset should not overlap, if possible, the selection subset of the first survey. The solution is thus, in that case, to define the selection subset for the third sample as $(0, \pi_k^2) \cup (\pi_k^1, \pi_k^1 + \pi_k^3 - \pi_k^2)$, in figure 5.

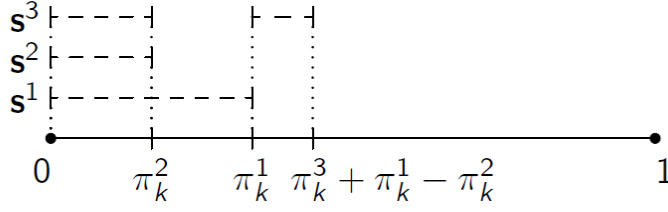


Figure 5. Third sampling occasion

4. Recursively, after t sampling occasions, we want to select a $(t+1)^{th}$ survey co-ordinated negatively with none, some, or all previous surveys and positively with the other surveys. A strict order of priority for these co-ordinations is required. Usually, but not necessarily, the reverse chronological order is used. The interval $(0,1)$ is subdivided into a collection of subsets that are the intersections of all selection subsets of previous surveys. These subsets are then ranked by strict order of compliance with the desired co-ordination rules, so that subsets that respect co-ordinations with higher priority are ranked higher than intervals that do not respect them. The selection subset for the $(t+1)^{th}$ survey is then obtained by including the top ranked subsets and, when necessary, part of one of these subsets, up to a total length of π_k^{t+1} . The progression of the number of subsets to consider is thus linear. For each unit, it is at most equal to $t+1$ after t sampling occasions. This last point is the feature that ensures that the system does not exceed computation capacities too rapidly, which is the main difficulty with multidimensional sampling designs.

3. Preparatory phase

Before any new survey is selected, different tasks have to be accomplished. First, if relevant, the sampling frame can be updated using available information in the business register. Second, the inclusion probabilities, or sampling design for the new survey need to be determined. Finally, co-ordination rules must be decided so that the selection subsets defined in section 2 may be computed.

3.1 Updating the sampling frame

When the sampling frame is updated, newborn units are added to the population with an empty selection history, trivial selection subsets and an independently generated PRN. Deceased units are removed from the frame. Parent units from splits and mergers can transmit their history and PRN to child units. However, in order to keep independence between units selection, each history and PRN can only be transmitted to one unit.

3.2 Computing inclusion probabilities for the new survey

With the introduction of this co-ordinated sampling method, all transversal sampling designs have been replaced by Poisson designs. These were previously, at the SFSO, mostly stratified sampling designs. Optimal sample allocation procedures need not necessarily be modified, but special care must be given to small sampling strata, for which Poisson sampling entails the risk of selecting an empty

sample, as is already the case when non-response is possible. When such small strata are present, it may be considered to deviate a bit from the optimal allocation, in order to keep this risk acceptable.

3.3 *Choosing co-ordination rules between the new survey and previous surveys*

For most applications, reverse chronological order is used, but when a panel or a rotating panel is updated, it makes sense to ensure the co-ordination first between the previous panel sample and the new one, and only then with other surveys.

4. **Examples – not tool specific**

4.1 *One-occasion business surveys, panels and rotating panels in Switzerland*

The SFSO has been using a co-ordinated sampling system for business surveys since October 2009. One occasion surveys as well as panels and rotating panels have been selected through this system. Most notably, the survey on value-added, a rotating panel survey with 20% rotation rate, has been selected and thrice updated. The rotation is achieved by considering the survey as a collection of five non-overlapping co-ordinated surveys: the rotation groups. Each year a fresh rotation group is selected, negatively co-ordinated with those surveyed the previous year, and four of the previous rotation groups are updated. This updating is obtained by selecting sample for these four groups anew, with maximal positive co-ordination with the samples they are supposed to replace. Other ongoing business surveys are progressively being integrated in this co-ordinated sampling system, when they undergo planned extensive revision and maintenance.

5. **Examples – tool specific**

6. **Glossary**

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. **References**

- Brewer, K., Early, L., and Joyce, S. (1972), Selecting several samples from a single population. *Australian Journal of Statistics* **3**, 231–239.
- Cotton, F. and Hesse, C. (1992), Coordinated Selection of Stratified Samples. *Proceedings of Statistics Canada Symposium*.
- Ohlsson, E. (1995), Coordination of samples using permanent random numbers. In: *Business Survey Methods* (eds. Cox, B. G., Binder, D. A., Chinnappa, D. N., Christianson, A., Colledge, M. J., and Kott, P. S.), John Wiley and Sons, New York, 153–169.
- Qualité, L. (2011), Developments on Coordinated Poisson Sampling. Presented at the 58th congress of the International Statistical Institute, Dublin.
<http://isi2011.congressplanner.eu/pdfs/950958.pdf>
- Qualité, L. (2009), *Unequal probability sampling and repeated surveys*. Ph.D. thesis, University of Neuchâtel, Switzerland (<http://doc.rero.ch/record/12284>).

- Rivière, P. (2001), Coordination Sample using the Microstrata Methodology. *Proceedings of Statistics Canada Symposium*.
- Tillé, Y. (2006), *Sampling algorithms*. Springer-Verlag, New York.

Specific section

8. Purpose of the method

This method provides a co-ordinated sampling system. Poisson samples can be selected with positive or negative co-ordination with previous samples that were selected within the co-ordinated sampling system. One occasion surveys, panels and rotating panels can be selected while spreading the response burden as evenly as possible on the population.

9. Recommended use of the method

1. This method is recommended for the selection of moderate to large sample surveys, or of a large number of small surveys, when co-ordination between samples is an important feature, and when correlation between selections of different units of the population at a given sampling occasion is not needed.

10. Possible disadvantages of the method

1. This method provides Poisson transversal designs, with random size. It is usually not a real problem since the sampling-related variability is secondary compared to the variability due to non-response and to its possibly inaccurate anticipation. The implied increase in estimation variance is negated by the use of calibrated estimators. However, precautions must be taken to limit the risk of having empty or too small samples in interest domains with modest planned sample size.
2. The independence between units selection makes it impossible to select samples where a strong dependence is required, such as for face-to-face surveys where geographic proximity is an important cost factor. For the same reason, this method does not provide a global co-ordinated sampling system for businesses and local units surveys, or for households and population surveys.

11. Variants of the method

1. Some dependence between units selection at a given occasion can be introduced by using the co-ordinated sample selection as one phase of a multiphase sampling design. For example, if the system is used to select local units samples, one may not want that more than one or a given number of local units of any business is selected. In the case of population surveys, one may want to avoid multiple selections within households. Inclusion probabilities at each sampling phase must then be computed, and co-ordination between actual surveys on field is necessarily degraded.
2. A rotating panel is obtained by splitting the survey into a collection of smaller surveys corresponding to rotation groups. In order to do so, the rotation rate must be a fraction of one, and kept constant over time.

12. Input data

1. Ds_input1: Sampling frame, essentially a list of identifiers, but other variables are useful in order to produce monitoring indicators;

2. Ds_input2: Inclusion probabilities for all units in the sampling frame, with zeroes for units outside of the target population;
3. Ds_input3: Sign of the desired co-ordination with previous surveys, and an order of priority for these co-ordinations;
4. Ds_input4: History of all units in the sampling frame, with their PRN and collection of selection subsets for all previous surveys.

13. Logical preconditions

1. Missing values
 1. No missing values are allowed.
2. Erroneous values
 1. Erroneous values for auxiliary variables of the sampling frame will not prevent the selection from occurring, but monitoring indicators will be incorrect. Other tables must be clean.
3. Other quality related preconditions
 1. Ds_input1, Ds_input3 and Ds_input4 can be joined with the unique identifier present in all three tables.
4. Other types of preconditions
 - 1.

14. Tuning parameters

1. None.

15. Recommended use of the individual variants of the method

1. When using the co-ordinated sampling as one phase of a multiphase design, organisation of the sampling phases should be such that the inclusion probabilities for the co-ordinated sampling phase are as close as possible to the true final inclusion probabilities. If this is the case, the co-ordination is still efficient. Usually, this can be achieved by using the co-ordinated sampling for the first phase of selection and then tailoring the sample to verify all requirements.
2. When selecting rotating panels, special attention has to be given to the co-ordination priorities in order to ensure that rotation groups at a given sampling occasion do not overlap and that units of the exiting rotation group do not re-enter the panel except when necessary.

16. Output data

1. Ds_output1: updated history of each unit of the population, after the current sampling occasion;
2. Ds_output2: selected sample for the current sampling occasion.

17. Properties of the output data

1. The selected sample is selected with a Poisson sampling design, with specified inclusion probabilities.
2. The co-ordination recorded in Ds_output1 and obtained between the current sample and the previous ones is optimal, in the sense that the current survey is optimally co-ordinated with the one with highest priority. Then, within the remaining wiggle room, it is optimally co-ordinated with the second survey by order of priority, and so on.

18. Unit of input data suitable for the method

Incremental processing of population units is possible since units are treated independently.

19. User interaction - not tool specific

1. Computation of inclusion probabilities.
2. Determination of co-ordination rules (sign and priority).

20. Logging indicators

1. Processing time.
2. Size of the population and expected size of the sample.
3. Number of selections in the population, within the system, as a measure of survey burden.

21. Quality indicators of the output data

1. Comparison between expected and obtained size of the sample.
2. Size of overlaps between the current survey and all previous surveys, number of repeated selections within the system.

22. Actual use of the method

1. Population surveys issued by the SFSO in Switzerland, starting in November 2010.
2. Business surveys issued by the SFSO in Switzerland, starting in October 2009.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Sample Selection – Sample Co-ordination

24. Related methods described in other modules

1. Sample Selection – Sample Co-ordination Using Simple Random Sampling with Permanent Random Numbers

25. Mathematical techniques used by the method described in this module

1. Basic arithmetic

26. GSBPM phases where the method described in this module is used

- 1.

27. Tools that implement the method described in this module

1. Unnamed SAS Macro at the SFSO.

28. Process step performed by the method

Sample selection and co-ordination

Administrative section

29. Module code

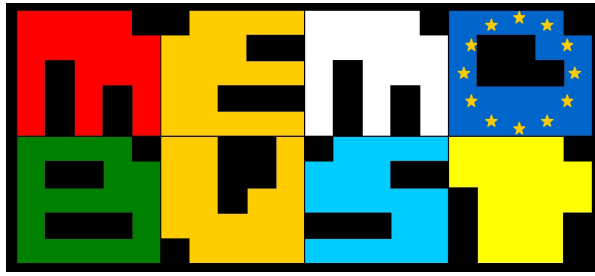
Sample Selection-M-PRN Using Poisson Sampling

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	12-06-2012	initial version	L. Qualité	SFSO (Switzerland)
0.2	15-05-2013	first revision	L. Qualité	SFSO (Switzerland)
0.2.1	18-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:44



This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Assigning Random Numbers When Co-ordination of Surveys Based on Different Unit Types is Considered

Contents

General section.....	3
1. Summary	3
2. General description of the method	3
2.1 Co-ordination when several unit types in the Business Register are considered.....	3
2.2 Assigning random numbers when several unit types in the BR are considered	4
2.3 Principles for co-ordination.....	5
2.4 Unit types in a Business Register	5
2.5 Top-down or bottom-up approach when assigning random numbers	6
2.6 Assigning PRNs to single-location and single-activity enterprise	6
2.7 Assigning PRNs to multiple-location and/or multiple-activity enterprises	7
3. Preparatory phase	7
4. Examples – not tool specific.....	7
4.1 Example 1	7
4.2 Example 2.....	8
5. Examples – tool specific.....	8
6. Glossary.....	8
7. References	8
Specific section.....	10
Interconnections with other modules.....	11
Administrative section.....	13

General section

1. Summary

Sample co-ordination by the use of Permanent Random Numbers (PRNs) is a common method used to have some control over the overlap (number of businesses in common) between samples for two different surveys or between consecutive samples for the same survey. The basic idea is to associate an independent and unique random number, uniformly distributed over the interval (0,1), with every unit in the Business Register. A BR generally consists of several unit types and unit type for a business survey is chosen on the basis of the statistics to be produced. This means that all unit types must be assigned PRNs. There are various methods for this but the most straightforward way would be to assign PRNs to each unit type separately. This method means that samples based on different unit types are independent but it does not admit co-ordination between such surveys. The fact that business surveys use different unit types implies a need for this kind of co-ordination. Especially the possibility of negative co-ordination between surveys based on different unit types (in order to spread the response burden) is important when it comes to small businesses.

Another approach to assign PRNs would be to use a method implying that the unit types can be co-ordinated through the PRNs. This method has the advantage to admit sample co-ordination between unit types but, as a drawback, brings dependence between samples based on different unit types. Co-ordination through PRNs cannot meet all objectives of sample co-ordination equally strong and different strategies are discussed in more detail below and references are given to other parts of the handbook.

2. General description of the method

2.1 *Co-ordination when several unit types in the Business Register are considered*

Sample co-ordination can be used to have some control over the overlap (number of businesses in common) between samples for two different surveys or between consecutive samples for the same survey. The main objectives of sample co-ordination are to obtain comparable and coherent statistics, high precision in estimates of change over time and to spread the response burden among the businesses¹, see theme module “Sample Selection – Sample Co-ordination” for more information. A common method to obtain sample co-ordination is based on the use of Permanent Random Numbers (PRNs). The basic idea is to associate an independent and unique random number, uniformly distributed over the interval (0,1), with every unit in the Business Register (BR). The method modules “Sample Selection – Sample Co-ordination Using Simple Random Sampling with Permanent Random Numbers” and “Sample Selection – Sample Co-ordination Using Poisson Sampling with Permanent Random Numbers” give different examples of sample co-ordination based on PRNs. The present module discusses assigning PRNs when several unit types in the BR are considered.

The majority of the National Statistical Institutes (NSIs) have not implemented co-ordination of surveys based on different unit types but Australia, France and Sweden are examples of countries using this kind of co-ordination. Australian Bureau of Statistics (ABS) achieves co-ordination between

¹ The word “business” is used as a generic name for all unit types used in business surveys.

samples of different types of units by the way the PRNs are assigned; see Brewer et al. (2000) for more information. The method used at ABS is quite similar to the method used in Sweden.

Institut National de la Statistique et des Études Économiques (INSEE) uses a somewhat different method to co-ordinate samples of different types of units. The co-ordination between unit types is mainly obtained by the way lower level units are connected to their higher level linked unit. See Hesse (1999) for more information.

The methodology described in this module is used in Statistics Sweden's system for co-ordination of frame populations and samples from the Business register (SAMU). For a general description of SAMU see Lindblom (2003).

A BR generally consists of several unit types and each business survey chooses unit type in accordance with the statistics to be produced. For example, institutional statistics is generally based on the enterprise unit, functional statistics is generally based on the kind of activity unit and regional statistics is generally based on the local kind of activity unit. Two types of sample co-ordination are commonly used (discussed in theme module "Sample Selection – Sample Co-ordination"), namely 1) co-ordination over time for one specific survey and 2) co-ordination between surveys based on the *same* unit type. However, there is a third kind of co-ordination to consider, namely co-ordination between surveys based on *different* unit types.

2.2 *Assigning random numbers when several unit types in the BR are considered*

The fact that business surveys use different kind of units in the BR means that all unit types must be assigned PRNs. There are several methods but the most straight-forward method would be to assign PRNs to each unit type separately meaning that the set of PRNs assigned to one unit type is completely independent of the set of PRNs assigned to another unit type. This method is simple and has the advantage that samples based on different unit types are independent of each other. However, it does not admit sample co-ordination between surveys based on different unit types. This drawback affects especially small businesses where the possibility to co-ordinate negatively (to spread the response burden) between surveys based on different unit types is very important.

Another approach to assign PRNs would be to use a method implying that the unit types can be co-ordinated through the PRNs. This method has the advantage to admit sample co-ordination between unit types but, as a drawback, brings dependence between samples based on different unit types. In the simple case with single-location and single-activity businesses this method means to assign the same random number to all units within a business. And, the majority of the small businesses consist of single-location and single-activity businesses which means that the proposed method for co-ordination between unit types works very well in this case. For the multiple-location and/or multiple-activity businesses this kind of co-ordination is less efficient because it is only possible to co-ordinate a multiple-location and/or multiple-activity enterprise with *one* of its lower level linked units. However, the majority of these businesses are large and large businesses are almost always included in samples so there are limited opportunities for spreading the response burden among them.

2.3 Principles for co-ordination

Co-ordination through PRNs offers a simple way to obtain co-ordination between unit types even though this method cannot meet all three objectives of co-ordination equally strongly. The reason is that the strategy to obtain the different objectives of co-ordination is somewhat contradicting:

- co-ordination over time for one specific survey and co-ordination between surveys based on the same unit type requires PRNs as permanent as possible
- co-ordination between surveys based on different kind of unit types would require PRNs that are, to some extent, updated

Strongest co-ordination, for one specific survey over time and between surveys based on the same unit type, is obtained by keeping the initially assigned PRN as permanent as possible. On the contrary, to maintain a strong co-ordination over time between unit types means that the PRNs needs to be somehow updated in order to follow changes in the business population in terms of registrations, de-registrations, mergers, split-offs, breakups and take-overs. An initially perfect co-ordination between unit types will otherwise gradually degenerate.

Updating PRNs contradicts the requirement from the two other types of co-ordination, namely keeping the PRNs as permanent as possible. To conclude, main objectives of co-ordination must be considered prior to the introduction of a system for co-ordination of surveys by the use of PRNs. Focus only on co-ordination over time for one specific survey and co-ordination between surveys based on the same unit type means that the best method is to assign PRNs to each unit type separately. Focus also on co-ordination between surveys based on different unit types means additional demands on the method for assigning PRNs.

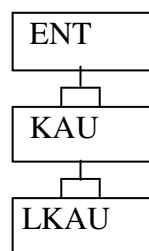
2.4 Unit types in a Business Register

A BR includes several unit types, generally at least the following:

- Enterprise Unit (ENT)
- Kind of Activity Unit (KAU)
- Local Kind of Activity Unit (LKAU)

A BR often includes more unit types compared to the above mentioned but principles for co-ordination of surveys based on different kind of units can easily be applied to a BR-structure including more unit types.

The relationship between the above mentioned unit types are showed in the figure below:



The unit types in the BR are linked together in a hierarchical way. In this example the LKAU is the smallest building brick in the BR. Each LKAU is linked to *one* upper level KAU and several LKAUs can be linked to the same upper level KAU. In the same way, each KAU is linked to *one* upper level ENT and several KAUs can be linked to the same upper level ENT.

2.5 *Top-down or bottom-up approach when assigning random numbers*

PRNs are assigned to all new units in the BR and a vital question is whether the assignment should be done by a “top-down” or a “bottom-up” approach. A top-down approach would mean to start the assignment of PRNs on the enterprise level and then go further down to the lower level linked units within the enterprise. And consequently, a “bottom-up” approach would mean to start the assignment of PRNs on the LKAU level and then go further up to the higher level linked units within the enterprise. The top-down approach means that a new enterprise is assigned a new random number and that the lower level linked KAU is assigned the same random number. If an enterprise has several lower level linked KAUs, one of them is assigned the same random number as the enterprise. Remaining new KAUs are assigned new random numbers. LKAUs are assigned random numbers according to the same method. A disadvantage with the top-down approach arises when a *new* enterprise is founded by one or more existing lower level linked units. As mentioned earlier, in order to co-ordinate between unit types one lower level linked unit should have the same random as the enterprise. However, one (or more) of the lower level linked units already have a random number and therefore run the risks of being forced to change from the existing random number to the new random number assigned to the enterprise.

The bottom-up approach means that a new LKAU is assigned a new random number and that a *new* higher level linked KAU is assigned the same random number. If a new KAU has several lower level linked LKAUs, the new KAU is assigned one of the LKAUs random number. And accordingly, a new enterprise is assigned the random number from one of its lower lever linked LKAUs. Note that another method (within the bottom-up approach) would be to assign a new enterprise the random number from one of its lower level linked KAUs. Examples 4.1 and 4.2 illustrate the difference between those two methods (or strategies).

The situation where a new enterprise is founded by existing lower level linked units causes no problem when using the bottom-up approach. Although, a disadvantage is that it can cause random number duplicates on the enterprise (and KAU) level due to changes in the business population in terms of mergers, split-offs, breakups and take-overs. However, the problem with random number duplicates can be solved quite easily.

2.6 *Assigning PRNs to single-location and single-activity enterprise*

In the simple case (single-location and single-activity enterprises) co-ordination of unit types through PRNs means to assign the same random number to all units within the enterprise. Note that this simple case applies to the absolute majority of the enterprises in the BR. In other words; the kind of activity unit and the enterprise unit are assigned the same random number as the local kind of activity unit. Bear in mind that a single-location and single-activity enterprise can change into another more complex structure and to maintain the co-ordination requires well considered continuity rules for the PRNs.

2.7 Assigning PRNs to multiple-location and/or multiple-activity enterprises

The assignment of random numbers to a multiple-location, or multiple-activity, enterprise is more complicated when co-ordination between unit types is considered. There are several possibilities to assign PRNs in this case and, compared to a single location and/or activity enterprise, the co-ordination for a multiple-location and/or multiple-activity enterprise will of course be less efficient. This is due to the fact that it is only possible to co-ordinate a multiple-location and/or multiple-activity enterprise with *one* of its lower level linked units. But the objectives to obtain comparable and coherent statistics imply that the method should facilitate co-ordination of the most important “enterprise-like” units within an enterprise unit. In several countries serves functional statistics as the most important input to the National Accounts. In addition, many other users of economic statistics want to follow different economic activities over time. A way of meeting this requirement would be to give the largest unit (from each unit type) classified into the same industry as the enterprise the same random number. Number of employees/persons employed is, in general, auxiliary information known at each unit type and therefore recommended to use as the size measure. Another approach would be to give the largest unit (from each unit type) classified into the same region as the enterprise the same random number. The chosen method for co-ordination of units within a multiple location/multiple activity enterprise must be decided after taking different demands on co-ordination into account.

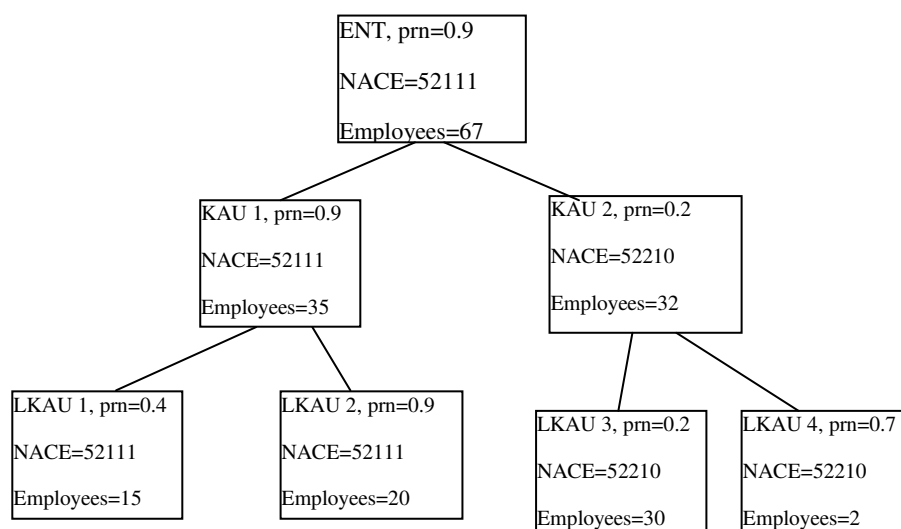
3. Preparatory phase

4. Examples – not tool specific

Different strategies can be used when assigning PRNs according to the “bottom-up” approach:

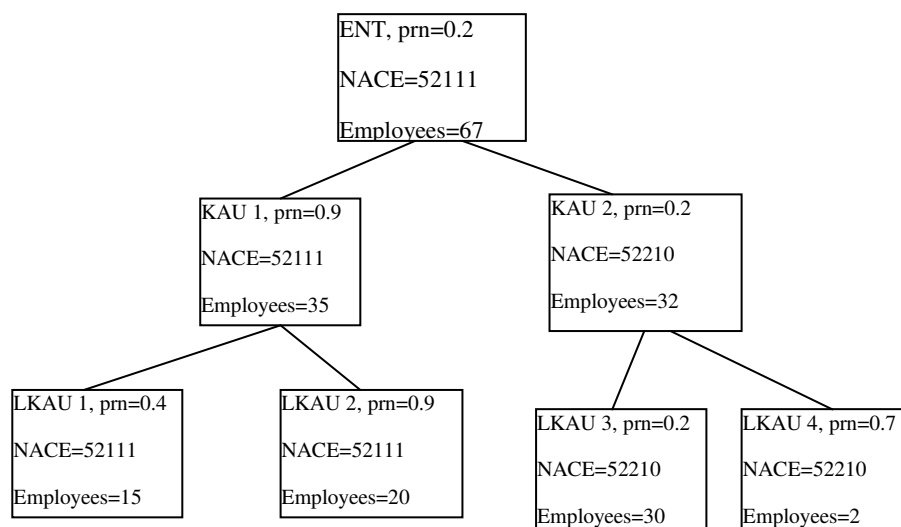
- Strategy A means to select a main unit on each level in respect to the closest upper level linked unit, see example 1 below.
- Strategy B means to select a main unit on each level in respect to the enterprise unit, see example 2 below.

4.1 Example 1



In the first example PRNs are assigned to each LKAU. Applying strategy A in this example means that, according to the earlier mentioned rules, KAU 1 is assigned the same PRN as LKAU 2 because this LKAU is the largest LKAU classified into the same two-digit industry as the KAU 1. In the same way, KAU 2 is assigned the same PRN as LKAU 3. Keeping to strategy A when assigning a PRN to the enterprise means selecting the main KAU and assign this PRN to the enterprise. This is KAU 1 because this is the largest KAU within the same industry (two digit-level) as the enterprise.

4.2 Example 2



Applying strategy B instead of strategy A gives example 2. As in example 1, KAU 1 is assigned the same PRN as LKAU 2 and KAU 2 the same PRN as LKAU 3. But, when assigning a PRN to the enterprise, strategy B means to select the main LKAU in the same industry (two digit-level) as the enterprise. LKAU 3 is the main unit and following strategy B means to directly assign this PRN to the enterprise (and not go via KAUs).

To conclude, strategies A and B give a different PRN to the enterprise level. In the first example LKAU 2 and KAU 1 are co-ordinated with the enterprise. In the second example LKAU 3 and KAU 2 are co-ordinated with the enterprise.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

- Brewer, K. R. W., Gross, W. F., and Lee, G. F. (2000), PRN Sampling: The Australian Experience. *ISI Proceedings: Invited Papers, IASS Topics, Helsinki August 10-18, 1999*, 155–163.
- Hesse, C. (1999), Sampling co-ordination: A review by country. Technical Report E9908, Direction des Statistique d’Entreprises, INSEE, Paris.

Lindblom, A. (2003), SAMU - The system for coordination of frame populations and samples from the Business Register at Statistics Sweden. Background Facts on Economic Statistics 2003:3, Statistics Sweden.

Specific section

8. Purpose of the method

Co-ordination of surveys based on different unit types

9. Recommended use of the method

1. Co-ordination through Permanent Random Numbers (PRNs) offers a simple way to obtain co-ordination of surveys based on different unit types.
2. Negative co-ordination is a very effective tool to spread the response burden among small businesses. Using this method means that negative co-ordination between surveys based on different unit types works very well for small single location businesses.

10. Possible disadvantages of the method

1. The method used to assign PRNs is more complicated.
2. PRNs on different unit types become dependent by using this method (and samples drawn based on different unit types).

11. Variants of the method

1. Top-Down approach when assigning PRNs in multiple-location and/or multiple-activity businesses.
2. Bottom-Up approach when assigning PRNs in multiple-location and/or multiple-activity businesses.

12. Input data

1. A Business Register

13. Logical preconditions

1. Missing values
 - 1.
2. Erroneous values
 - 1.
3. Other quality related preconditions
 - 1.
4. Other types of preconditions
 - 1.

14. Tuning parameters

- 1.

15. Recommended use of the individual variants of the method

1.

16. Output data

1.

17. Properties of the output data

1.

18. Unit of input data suitable for the method

19. User interaction - not tool specific

1.

20. Logging indicators

1.

21. Quality indicators of the output data

1.

22. Actual use of the method

1. This method is implemented in Statistics Sweden's system for co-ordination of frame populations and samples from the Business register (SAMU).

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Sample Selection – Sample Co-ordination

24. Related methods described in other modules

1. Sample Selection – Sample Co-ordination Using Simple Random Sampling with Permanent Random Numbers
2. Sample Selection – Sample Co-ordination Using Poisson Sampling with Permanent Random Numbers

25. Mathematical techniques used by the method described in this module

1.

26. GSBPM phases where the method described in this module is used

1. Design phase

2. Data collection phase for frame creation and sampling

27. Tools that implement the method described in this module

1.

28. Process step performed by the method

Administrative section

29. Module code

Sample Selection-M-PRN with Different Unit Types

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	01-04-2013	first version	Annika Lindblom	Statistics Sweden
0.2	16-05-2013	improvements based on the Norwegian and Swiss reviews	Annika Lindblom	Statistics Sweden
0.3	29-05-2013	improvements based on the Norwegian and Swiss reviews	Annika Lindblom	Statistics Sweden
0.4	15-08-2013	improvements due to new information on useful references	Annika Lindblom	Statistics Sweden
0.4.1	18-09-2013	preliminary release		
0.5	27-09-2013	improvements due to the EB-reviews	Annika Lindblom	Statistics Sweden
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:44