Project N°: 262608

ACRONYM: Data without Boundaries

STANDALONE DOCUMENT

Guidelines for Output Checking

WORK PACKAGE 11

(Improved Methodologies for Managing Risks of Access to Detailed OS Data)

Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

This document is extracted from the deliverable "D11.8 - Final reports of synthetic data CTA, ECTA, cell suppression & Guidelines for output checking", part 7.

This document was prepared under WP11 by CBS, Destatis & ONS.

The full report is available here.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608 (DwB - Data without Boundaries).

Guidelines for the checking of output based on microdata research

Steve Bond (ONS), Maurice Brandt (Destatis), Peter-Paul de Wolf (CBS)

Summary: In this document practical guidelines for output checking are given. Output checking is the process of checking the disclosure risk of research results based on microdata files made available in Research Data Centres. Actual rules for different types of output are given. These rules are placed within a framework of two different models for output checking. Also, recommendations and best practices for some organisational and procedural aspects of the output checking process are given.

Keywords: Research Data Centre, disclosure control, output checking, guidelines

Preface

This document builds on a document that was produced within the European project 'ESSnet SDC' in 2009. We are grateful to the authors of that version (i.e., Maurice Brandt (Destatis), Luisa Franconi (ISTAT), Christopher Guerke (Destatis), Anco Hundepool (CBS), Maurizio Lucarelli (ISTAT), Jan Mol (CBS), Felix Ritchie (ONS), Giovanni Seri (ISTAT) and Richard Welpton (ONS)) that we could build upon a solid foundation.

Note that the current guidelines can be considered to be a "live document": additional statistical analyses may be added in the future, insights may change, etc.

Any suggestions from whoever may read this version are very welcome and will be taken into account in future versions of the guidelines. Suggestions can be send to pp.dewolf@cbs.nl and/or maurice.brandt@destatis.de.

Table of contents

1 Introduction and general approach

1.1 Introduction

National Statistical Institutes and Data Archives (henceforth collectively referred to as “data providers”) realise that the full potential of their microdata can never be extracted by just their own staff and in their own official publications. The shift from survey-based to register-based statistics has led to a rapid increase in the amount of available microdata. At the same time, IT developments have made it possible to analyse these very large datasets, but a lot of data providers simply do not have the resources available to make full use of these available datasets. Luckily, a very large number of qualified researchers in the world of academia and policy-making are willing to share in this work. Moreover, data available at data providers’ premises can be a treasure to researchers. Indeed, excellent research can be done using that data, often for the public good. Data providers therefore provide researchers access to their microdata files, for instance in a Research Data Centre (RDC). Data providers must then find the right balance between enabling these researchers to do their work while always ensuring the confidentiality of the data.

Allowing researchers to work on microdata in a controlled environment like an RDC gives them the opportunity to see the microdata and hence individual information. However, researchers want to take results of their analyses out of the controlled environment, e.g., to publish in journals or to discuss with fellow researchers. Then the confidentiality of the data comes into play again.

One common method of protecting the confidentiality of the data is to check all results that researchers want to take out of the controlled environment of the RDC. Throughout this document, this process will be referred to as *output checking*.

This document provides the reader with guidelines for, and best practices of this output checking process. The aim of the document is threefold:

1. to allow experienced data providers to learn from each other by sharing best practice
2. to provide data providers that have little or no experience in RDCs with practical advice on how to set up an efficient and safe output checking process
3. to facilitate harmonisation of output checking methods across data providers.

The third aim is particularly crucial in light of the current focus on the development of European infrastructures for microdata access, as it is crucial that data providers agree on a common method for checking output. As cross-border data access will be most efficient if data provider X can check output that was generated with datasets from country Y as well. Only with common guidelines will data providers delegate the task of output checking to each other.

The remainder of this chapter deals with a general approach to output checking. In chapter 2, possible types of output will be categorised. For each category of output both a simple rule and guidance for more experienced researchers will be outlined. Chapter 3 then deals with a number of practical issues (procedural, organisational), which are necessary for an efficient high-quality output checking process.

At the end of the document some annexes are included. Annex 1 contains some examples of disclaimers that researchers can use when they produce their output based on the data at the RDC. Annex 2 is an example of output description to be used by researchers when producing their output. Annex 3 is a short glossary of some of the terms we used in this document. Finally, in Annex 4 we have for a selected number of analyses some additional information and/or examples.

1.2 General approach

Before continuing with the rest of the document we first need to discuss the general approach to output checking.

The RDC zoo

An RDC is a safe environment where accredited researchers can access the most detailed micro data to undertake virtually any research that they desire (within the restrictions given by the associated data provider, e.g. like “as long as it serves the public good”). This makes output checking in an RDC totally different from disclosure control of official data provider publications. The official publications are of a well-defined form (usually a table) and the intruder scenarios are limited in number. Whereas the output of an RDC can be anything! Researchers twist, transform and link the original data in different and complex new ways. This makes it very difficult to come up with a set of rules that covers every possible output, as one expert vividly states it: designing rules for output checking is like designing cages for a zoo - that will keep the animals both contained *and alive* - without knowing in advance which animals will be kept in the cages[1].

Safe and unsafe classes of output[2]

To bring some order to output checking in an RDC all output can be classified into a limited number of categories (for instance tables, regression etc.). Each class of output is then labelled ‘safe’ or ‘unsafe’. This classification is done solely on the functional form of the output, not on the data itself.

- ‘safe’ outputs are those which the researcher should expect to have cleared with no or minimal further changes; for example, the coefficients estimated from a survival analysis. Analytical outputs and estimated parameters are usually ‘safe’. The exceptions where a ‘safe’ output is not released should be well defined and limited in number.

[1] Ritchie, F (2007) *Statistical Disclosure Control in a Research Environment*. Mimeo: Office for National Statistics

[2] For a more detailed explanation, see Ritchie, F. (2008) “Disclosure detection in research environments in practice”, in Work session on statistical data confidentiality 2007; Eurostat; pp399-406

- ‘unsafe’ outputs will not be cleared unless the researcher can demonstrate, to the output checker’s satisfaction, that the particular context and content of the output makes it non-disclosive. For example, a table will not be released unless it can be demonstrated that there are enough observations, or the data has been transformed enough, so that the publication of that table would not lead to identification of a respondent. Linear aggregations of data are almost always ‘unsafe’.

Note that for safe outputs, the output checker must provide reasons why the output cannot be released against normal expectations; for unsafe outputs, the researcher has to make a case for release. In both situations, the ultimate decision to release an output remains with the output checker.

Output checking is always context specific. It is not possible to say, ex ante, that something will or will not be released. The purpose of the safe/unsafe classification is to give guidelines on the likelihood of an output being released – and, if it is unsafe, to suggest ways of making it safe.

The most important criteria are establishing that no respondent can reasonably be identified, and that no previously unavailable confidential information about groups can be inferred.

‘Reasonably’ is not explicitly defined – all definitions are subject to criticism; moreover, the aim of principles-based output checking is that all definitions are subject to the particular context. However, some factors which would be taken into consideration include situations when identification is

- taking a significant amount of time and effort
- requiring additional knowledge which most individuals would not be expected to have or be able to acquire easily
- needing some technical ability

Not all outputs have yet been classified. The default general classification is ‘unsafe’.

Two types of error

With this in mind, one needs to realise that the optimal way to check output is the one that maximises the use of the datasets while minimising the risk for disclosure. So, phrased differently, a less than perfect output check can lead to two types of errors:

1. Confidentiality errors: releasing disclosive output
2. Inefficiency errors: not releasing non-disclosive output

Rules and guidelines for output checking can prevent both errors. But the trick is to find the right rule.

Consider, for instance, a rule that sets a minimum for the number of units in each cell of a table (threshold rule). If the minimum cell count is set too high, this will lead to inefficiency errors: the output will be safe, but will contain less information than could be obtained from the datafile. The researcher might have had to group classes that he was interested in to reach the threshold. On the other hand, if the minimum is set too low, this will lead to unsafe outputs being released.

Finding the correct disclosure control rules for use in an RDC is especially difficult. One has to realise that the exact shape of the output is not known beforehand. The idea of an RDC is to give researchers the maximum amount of freedom in analysing the data files, so they will produce output of all sizes and shapes (tables, models etc.).

To deal with this problem, a two-model approach has been developed. The first model is called the principles-based model. This model minimises both confidentiality and inefficiency errors. The other is called the rule-of-thumb model. For this model, the focus is on preventing confidentiality errors and inefficiency errors are accepted. Both models will be discussed in more detail in the following paragraphs.

Even though the two models could be used independently from each other, the general idea is to 'nest' the two models. I.e., the rule-of-thumb model could be used to make a first selection of safe output. The complement of this selection is not necessarily always unsafe. Then the principle-based model could be used to further investigate and decide on the safety of the dissemination of the remaining output.

1.3 Principles-based model

The principles-based model centres on a good collaboration between researchers and RDC staff. Because this model also aims to prevent inefficiency errors, simple rules for output checking are not appropriate. The reason is that simple rules can never take into account the full complexity of research output. To give maximum flexibility to the researcher, no output is ruled in or out in advance. All output needs to be considered in its entire context before deciding on its safety. For instance, a table that contains very small cell counts (maybe even some cell counts of 1) isn't necessarily unsafe. If, for instance, the table only contains large recognisable groups spread out over different categories of a sensitive variable, it might be impossible to identify the exact individual that belongs to the cell count of 1.

Moreover, when checking output for disclosure, one should take into account the fact that the checked result may be related to another (previously) released output. The combination with other output might hamper the safety of the current results. Indeed, usually researchers ask to check for disclosure of several results at the same time (e.g. a set of related tables).

What is thus needed is a clear understanding of the governing principles behind disclosure control. Therefore, both the researcher and the RDC staff need training in disclosure control.

The principles-based model has the obvious advantage that it leaves a maximum amount of flexibility to the researcher. Data files will therefore be used to their fullest extent. However, the model also has some possible drawbacks:

- The model relies on serious training of the data provider's staff and of researchers. Researchers have to be willing to invest their time and effort on a topic, which is not naturally within their field of interest.
- The model spreads the responsibility for clearing an output. In a rules-based model, the responsibility lies with the people that design the rules. In this principles-based model, the responsibility lies with each individual output checker. There are no strict rules to follow and each checker has to make his own decision on clearing the output, based on his experience and understanding of the underlying principles.

1.4 Rule-of-thumb model

In this model the main focus is on preventing confidentiality errors, and some inefficiency errors are taken for granted. This model typically leads to very strict rules. The chance that an output that passes these rules is non-disclosive is very high. The advantage is that the rules can be applied more or less automatically by both researchers and staff members with only limited knowledge of disclosure control.

It is important to stress the fact that, although the rule of thumb model is very strict, this is not a 100 % guarantee that all output that passes these rules is indeed non-disclosive. There is a very small chance that a disclosive output slips through. This is because the rules are rigid and do not take the full context of the output into consideration.

The rule-of-thumb model is useful for a number of situations:

- Naïve researchers whose output is usually far from the cutting edge of disclosure control (for instance policy makers who just want tabular output with limited detail).
- Inexperienced data providers starting up an RDC. In this case, both users and RDC staff could have too little experience to be able to work with the principles-based model. The rule-of-thumb model provides them with a starting point that ensures maximum safety. In using the rule-of-thumb model, they build up experience along the way. At some point in time they might feel confident enough to set up the principles-based model and open up the way to clearing more complex output.
- Automatic disclosure control for RDCs. This will mainly be useful for more controlled types of data access like remote execution. In remote execution, researchers write their scripts without having access to the real datafile (sometimes dummy datasets are provided for this purpose). They then send the finished script to the RDC, where a staff member (or an automated system) runs it on the full datasets. The results are then returned to the researcher.
- Reducing the burden of output checking. Even for RDCs using the principles-based model, the rule-of-thumb model is usually the starting point when checking any particular output. Using the rule of thumb, attention is quickly focused on the parts of the output that breach these rules.

These parts can then be considered more carefully using the full principles-based model to decide whether they can be released or not.

2 Rules for output checking

2.1 Classification of output

As described before, the RDC zoo can be somewhat structured by classifying all output into a limited number of output types. The table below lists the different types of output. Each type is marked as either generally safe or generally unsafe (see section 1.2 for an explanation of the generally safe-unsafe classification).

Type of Statistics	Type of Output	General Classification
Descriptive statistics	Frequency tables	Unsafe
	Magnitude tables	Unsafe
	Maxima, minima and percentiles (incl. median)	Unsafe
	Mode	Safe
	Means, indices, ratios, indicators	Unsafe
	Concentration ratios	Safe
	Higher moments of distributions, incl. variance, covariance, kurtosis, skewness)	Safe
	Graphs: pictorial representations of actual data	Unsafe
Correlation and regression type analysis	Linear regression coefficients	Safe
	Non-linear regression coefficients	Safe
	Estimation residuals	Unsafe
	Summary and test statistics from estimates (R^2, χ^2 etc.)	Safe
	Correlation coefficients	Safe
	Factor analysis	Safe
	Correspondence analysis	Safe

2.2 The overall rule of thumb

As discussed before, the rule-of-thumb model is based on clear and simple (and strict) rules. Because these rules differ only slightly for different classes of output, an overall rule of thumb can be established.

This overall rule of thumb is presented first; after that, when describing the rules for each class of output, the interpretation of this overall rule of thumb for the specific class of output is given.

The overall rule of thumb has four parts:

1. **10 units**: all tabular and similar output should have at least 10 units (unweighted) underlying any cell or data point presented. A common term for such a rule is a threshold rule (the cell count must exceed a specified threshold).
2. **10 degrees of freedom**: all modelled output should have at least 10 degrees of freedom and at least 10 units should have been used to produce the model.
   Degrees of freedom = (number of observations) -/- (number of parameters)
   -/- (other restrictions of the model)
3. **Group disclosure**: in all tabular and similar output no cell can contain more than 90 % of the total number of units in its row or column to prevent group disclosure. Group disclosure is the situation where some variables in a table (usually spanning variables) define a group of units and other variables in the table divulge information that is valid for each member of the group. Even though no individual unit can be recognised, confidentiality is breached because the information is valid for each member of the group and the group as such is recognisable.
4. **Dominance**: in all tabular and similar data the largest contributor of a cell cannot exceed 50 % of the cell total.

Note that the values of the parameters mentioned above (e.g. the threshold of 10 units) are a convention with the purpose to be able to quickly filter out safe output. Output not passing these rules, is not necessarily disclosive, but needs further attention. See the discussion at the end of section 1.2 as well.

A practical problem: The dominance rule

As simple as it looks, even the overall rule of thumb has one major difficulty. This concerns element n^{o} 4: the dominance rule.

In order to check this rule, the researcher must provide additional information on the value of the largest contributor for each cell. This often burdens the researcher with a lot of extra work. In addition, the extra information obviously has to be removed from the output before release, as releasing the value of the largest contributor of a cell would be very disclosive!

So, although the dominance rule is included in the rule of thumb, the current practice in many countries is that it is not actively checked. Even so, researchers are told that they should take it into consideration when creating their output.

Usually, a data provider decides to actively check it only in certain circumstances, for instance:

- magnitude tables on business data
- output based on very sensitive variables
- variables with a very skewed distribution.

Nevertheless, for those wishing to follow only the rule of thumb rather than the principles-based model, checking for dominance should be considered best practice.

The remainder of this chapter will examine each output classified above and discuss the interpretation of the overall rule of thumb and give some more detailed information for the principles-based model.

2.3 Descriptive statistics

Class 1: Frequency tables

Rule of thumb

- Each cell of the table must contain at least 10 units (unweighted).
- The distribution of units over all cells in any row or column is such that no cell contains more than 90 % of the total number of units in that particular row or column (group disclosure)
- No cell in a table is formed from units of whom 90% or more originate from one organisation.

Detailed information for principles-based model

The following issues should be considered:

- Is the data in the frequency table confidential? If not, then counts of 1 could be acceptable. If the data in the table are direct responses from a survey, census or administrative data, then they are confidential and we must ensure that singletons could not be discovered. If the data in the frequency table are derived from confidential data, then the output checker needs to consider whether the original data could be recovered. This will depend on: the means of transformation and whether the algorithm for transformation is provided or could be determined; how many low counts are present in the table.
- In case a singleton in a frequency table can only be discovered using all the information that defines the table, no additional information can be disclosed about that individual. I.e., only identity disclosure is possible but no attribute disclosure. Unless the legislation states otherwise, this is usually not a problem.
- Counts less than 10, but greater than 1, can be released if the likelihood of identification is low enough to be considered negligible. Considerations here include: likelihood of other relevant information being available to assist identification; age of the data; geographical aggregation; sample design; whether the rank ordering of contributors are known. Also to be included is the

question: is this output useful, including reliability of the estimate? If not, then the rule of thumb should be applied.

- Counts of 2 should be treated with caution as one member of the pair could potentially identify the other.
- Whether group disclosure is problematic will depend mainly on whether the information in the table is confidential or not. For example, the number of commuting destinations for the banking sector in central London aggregated by local authority would have a huge majority to the City of London. But this doesn't reveal any information about individual respondents (except that they are all clustered together). An output checker should also consider sample design as for many surveys the fact that a particular respondent is in the sample would be considered confidential.
- For group disclosure at the enterprise level, the first four above apply.

Class 2: Magnitude tables

Rule of thumb

- Each cell of the table contains at least 10 units (unweighted).
- The distribution of units over all cells in any row or column is such that no cell contains more than 90 % of the total number of units in that particular row or column (group disclosure).
- In every cell, the largest contributor cannot exceed 50 % of the cell total.

Detailed information for principles-based model

There may be circumstances in which the cell count rule of ten is inappropriate. For example, a researcher may believe that an output is non-disclosive even though cell counts do not meet the required threshold of ten units. The onus is then on the researcher to explain why this is the case, and the individual responsible for checking output will be required to make a decision as to whether this output can be released.

The following issues should be considered:

- Is the data in the table confidential? If not, then we could permit tables to be published with cell counts of 1. Note that even if RDC data is also in the public domain, is it still classed as confidential *unless* a waiver has been obtained from the owner of the data.
- If the data is confidential, then a count of 1 clearly cannot be released. But what about where the data derive from counts of 2? In this case we would consider this to be unambiguously disclosive as if one of the respondents were to read the research, they would know the data of the other respondent.
- For counts of 3 – 9, the risk of disclosure decreases as it would require groups of respondents to share information in order to reveal information about the remaining respondent. The threshold

of 10 is chosen to be quite high to minimise this possibility. So, when checking tables based on counts less than 10, an output checker needs to be assured that other tables with slightly different aggregations are not also in the public domain, otherwise disclosure by differencing may be possible. This is difficult to do.

- Checks for dominance are usually difficult to make, as the researcher will not provide the value for the largest respondent (because to do so would be to provide confidential information). If the checker is concerned that dominance might be an issue, then the output should be retained until assurances are given. To that end, the researcher may provide additional information about the largest respondent *to the output checker*, as long as is can be assured that the information will only become available to the output checker. Situations where dominance might be an issue are discussed in the appendix.

Class 3: maxima, minima, percentiles

Rule of thumb

Maxima and minima are not released since they usually refer to only one unit.

A **percentile** (or centile) is the value of a variable below which a certain percentage of observations fall. Percentile data, therefore, usually represent the value of the variable for an individual respondent and are not released.

Detailed information for principles-based model

The following issues should be considered:

- Minima and Maxima data may represent responses from a number of individuals. For example, minimum age in a sample (in whole years) would be zero and would simply represent all babies in their first year. As long as there are more than 10 of these, such data would be considered safe. If the researcher provides the count which underlies the minima or maxima it will help make this assessment.
- For maxima, in particular, care needs to be taken to ensure disclosure is not possible, because usually the largest/oldest/wealthiest etc. respondent is known. It may also provide information that a particular respondent is in the sample which may help identify further information from the analysis. Issues of dominance and group disclosure need to be considered in addition to the threshold rule before providing clearance for a maximum value.
- For percentiles, the question to be considered by the output checker is whether the respondent could be identified from the value published. Firstly, is the rank ordering of firms known, or guessable? If so, the percentile cannot be released. What is the variance of values around the percentile? If this is low, or zero, then there is the possibility of group disclosure around the percentile. Is the variance around the percentile very large? If so, the identity of the percentile

respondent might be guessable – for example in a distribution with many high values and many low values with only one central value respondent, the 50th percentile (= median) will provide the returns from the central respondent.

- For minima, maxima and percentiles, the researcher will always have to provide additional information to assure that the risk of disclosure is negligible.

Class 4: Modes

Rule of thumb

The mode is treated the same as a magnitude table.

- A mode must contain at least 10 units (unweighted).
- The distribution of units is such that the mode should contain no more than 90 % of the total number of units in that particular row or column (group disclosure)

Detailed information for principles-based model

As for Class 2.

Class 5: Means, indices, ratios, indicators

Rule of thumb

The same considerations as for magnitude table cells apply; in particular, elements 1 and 3 of the overall rule of thumb apply:

- Each single value should derive from the synthesis of at least 10 units (unweighted).
- For each single value to be released, the largest contributor included in the synthesis cannot exceed 50 % of the total.

Detailed information for principles-based model

The following issues should be considered:

- Is the data in the table confidential? If not, then we could permit tables to be published with cell counts of 1. Note that even if RDC data is also in the public domain, is it still classed as confidential *unless* a waiver has been obtained from the owner of the data.
- If the data are confidential, then a count of 1 clearly cannot be released. But what about where the data derive from counts of 2? In this case we would consider this to be unambiguously disclosive as if one of the respondents were to read the research, they would know the data of the other respondent.
- For counts of 3 – 9, the risk of disclosure decreases as it would require groups of respondents to share information in order to reveal information about the remaining respondent. The threshold

of 10 is chosen to be quite high to minimise this possibility. So, when checking tables based on counts less than 10, an output checker needs to be assured that other tables with slightly different aggregations are not also in the public domain, otherwise disclosure by differencing may be possible. This is difficult to do.

- For counts of 3 – 9, the risk of disclosure decreases as the complexity of the index increases. The appendix provides a discussion on complexity.
- Checks for dominance are usually difficult to make, as the researcher will not provide the value for the largest respondent (because to do so would be to provide confidential information). If the checker is concerned that dominance might be an issue, then the output should be retained until assurances are given. To that end, the researcher may provide additional information about the largest respondent *to the output checker*, as long as is can be assured that the information will only become available to the output checker. Situations where dominance might be an issue are discussed in the appendix.

Class 6: Concentration ratios

Rule of thumb

Concentration ratios are safe as long as they meet elements 1 (threshold), 3 (group disclosure) and 4 (dominance) of the rule of thumb.

Detailed information for principles-based model

Concentration ratios are allowable provided the results can be shown to be non-disclosive, which would be the case if

- the sample on which the output is based on contains at least ten units
- the dominance rule (general rule of thumb number 4) is met
- percentages are displayed with no decimal places

Class 7: Higher moments of distributions

Rule of thumb

Higher moments (variance, skewness, kurtosis) are safe as long as part 2 of the rule of thumb is met:

- all modelled output should have at least 10 degrees of freedom and at least 10 units have been used to produce the model.
  Degrees of freedom = (number of observations) -/- (number of parameters) -/- (other restrictions of the model)

Detailed information for principles-based model

The simple rule applies. In discussions with researchers, it may be noted that with less than ten degrees of freedom the statistical value of the output is doubtful anyway.

Class 8: Graphs (descriptive statistics or fitted values)

Rule of thumb

Graphs in themselves are not permitted as output. The underlying (aggregated) data may be submitted as output, which when cleared may be used to reconstruct graphs in the researcher's own environment.

Detailed information for principles-based model

Graphical output is only allowed in cases where it is impossible to reproduce the graph without access to the full microdata, or it should be based upon undisclosive data. E.g., graphical output build from (tabular) output that would be cleared.

Moreover, the graph should meet the following conditions:

- Data points cannot be identified with units. When the graph consists of transformed or fitted data, this is usually not a problem.
- There are no significant outliers.
- The scales used should not be too detailed, to prevent disclosure of (nearly) exact values.
- The graph is submitted as a 'fixed' picture, with no data attached. This means that graphs should be submitted as files with one of the following extensions: .jpg, .jpeg, .bmp or .wmf.

2.4 Correlation and regression type analysis

Class 9: Linear regression coefficients

Rule of thumb

Linear regression coefficients are safe provided at least one estimated coefficient is withheld (e.g. intercept).

Detailed information for principles-based model

Complete regressions can be released as long as

- they have at least 10 degrees of freedom
- they are not based solely on categorical variables
- they are not on one unit (e.g. time series on one company)

Whenever linear regression is applied with only binary explanatory variables, the regression coefficients will reflect table means. Hence, the rules for tabular data apply.

Class 10: Non-linear regression coefficients

Rule of thumb

As with linear regressions, non-linear regression coefficients are safe provided at least one estimated coefficient is withheld (e.g. intercept).

Detailed information for principles-based model

Non-linear regressions differ from linear regressions because of the nature of the dependent variables, which are often discrete. If a regression is estimated, and the same regression is repeated with one additional observation, then by using the means of the variables in conjunction with the regression results it may be possible to infer information about that one particular observation.

However, in practice, this is unlikely. Changes in the number of observations included in a model are the result of

- changing the sample explicitly, or
- changing the specification and so finding observations with inadmissible values (eg missing) dropping out

The first is unlikely to lead to one observation more or less, as the information gain is negligible (unless the researcher is deliberately aiming to circumvent SDC rules). In the second case, different numbers of observations are not relevant because the specification has been changed.

Class 11: Estimation residuals

Rule of thumb

No residuals and no plots of residuals should be released.

Detailed information for principles-based model

As with the rule of thumb, residuals should not be released.

A reasonable request for a plot of residuals may be made by a researcher, for example, in order to demonstrate the robustness of the model. However, plots of residuals should be discouraged. Instead, a researcher should analyse plots within the safe setting, and a written description of the shape of the plot may be released to demonstrate robustness. If there is a need for a plot of residuals to be released, this should be assessed as for graphs (see Class 8).

An exception may be when (standardised) residuals of aggregated figures are considered. Like for instance in the case of a log-linear model describing a frequency count table. Then the residuals of a cell may be released when that cell is based on at least 10 units.

Class 12: Summary and test statistics

Rule of thumb

The following summary statistics can be released provided there are at least 10 degrees of freedom in the model:

- R^2 and variations
- estimated variance
- information criteria (e.g. AIC, BIC)
- individual and group tests and statistics (t, F, chi-square, Wald, Hausman etc.)

Detailed information for principles-based model

No principles in addition to the rule-of-thumb guidelines.

Class 13: Factor analysis

Rule of thumb

Factor scores can usually be released, as they are (usually un-interpretable) combined scores on observed variables. As a rule of thumb, a factor must be related to at least two different observed variables to be safe to release.

Detailed information for principle-based model

Summary information on factor scores (like maximum, minimum, mean, etc.) can be released as long as it relates to more than one observed variable and is not associated with an individual observation. I.e., similar rules as in Class 3 apply.

Class 14: Correlation coefficients

Rule of thumb

Correlation is a measure of linear relationships between variables. The checkers have to ensure that the released outputs cover the first element of the rule-of-thumb, which means a minimum of 10 unweighted units underlying each correlation coefficient and that the correlation coefficient is not equal -1, 0, +1.

Detailed information for principles-based model

A very small number of cases exist where problems could arise (e.g. the correlation matrix includes 0 or 1). Even in these cases, the problem is the publication of correlation coefficients connected with summary statistics.

Correlations of binary variables are treated like linear regressions with binary explanatory variables or full saturated linear regressions. Therefore, the same items have to be taken account of as for linear regression coefficients (see Class 9). Nevertheless, in the majority of analyses, correlation coefficients could be classified as safe.

Class 15: Correspondence analysis

Rule of thumb

The correspondence analysis is mainly based on graphical interpretation of chi-squared-distances of categorical data. The loadings on the dimensions are usually safe if a minimum number of observations is preserved and if the minimum number of variables is two. There is protection because of the categorical structure of the data but also metric variables can be recoded to categorical.

Detailed information for principles-based model

There might occur special cases where the loading of one variable in the first dimension is 1 and in the second dimension is 0. In this special case the values of one observation are known or can be recalculated exactly. But it still depends on the characteristic of the categorical variable if this is a confidential information of a single person or enterprise.

3 Organisational/procedural aspects of output checking

The previous chapter dealt with rules for output checking. An efficient, high-quality output checking process is only possible when effective organisational and procedural practices are implemented in addition to these rules. In this chapter the most important of these will be discussed. The minimum standard can be seen as the best value-for-money measure, while the best practice is the operational excellence every RDC should (ultimately) aim for.

3.1 Legal basis

The aim of this section is to outline the legal framework for output checking. Output (and therefore output checking) is essential to an RDC. Because an RDC provides a researcher access to (highly) confidential data, some of the data provider's responsibility for confidentiality needs to be transferred to the researcher. This responsibility can be underpinned by drawing up contracts that include the dos and don'ts for researchers using an RDC.

Within workpackage 3 of the DwB project, a report is written about legal frameworks in practice within the EU, based on a survey for different types of data. That report can be found at the DwB website as part of deliverable D3.2 "A report on the legal frameworks for data access to Official data".

Minimum standard

A legally enforceable agreement with the researcher needs to be drawn up. As a minimum standard this contract should include the statement that data on identifiable individual persons/households/

companies/organisations etc. can *never* leave the safe environment of the RDC, and that the researcher has a responsibility to ensure this.

Best practice

As a best practice a two-tier approach is suggested. The first level is a contract with the research organisation that the user is affiliated with. This contract makes the organisation as a whole responsible for the research project and for keeping the data safe. The second level is a confidentiality statement signed by each individual researcher, binding him to safeguard data confidentiality. Of course these agreements and statements have to be signed by the data provider as well. The person who signs on behalf of the data provider strongly depends on the national organisation. Several different situations have been identified, including:

- the RDC director: the person in charge of the RDC has the power to give permission;
- the data provider/Institute: President/Director/Legal unit/Board of Directors;
- the Director of the department responsible for the data;
- a Statistical/Confidentiality Committee: this can be internal or external to the data provider/institution.

The first two situations are the most frequently observed.

3.2 Access requests

An access request is a key component of the organisational and procedural aspect of running an RDC. At the access request stage the following points should be made clear: (i) who is applying to access the RDC (is he/she or his/her institute eligible?); (ii) which results are going to be produced (are they feasible?); (iii) how results will be disseminated (is this consistent with the data providers policy?). Even where these issues do not directly affect output checking, an accurate preliminarily evaluation of the access request can make the checker's job easier (see for instance section 3.6).

Within workpackage 3 of DwB, two reports have been written about researcher accreditation: deliverable D3.1 "Researcher accreditation - current practice, essential features and a future standard" and deliverable D3.4 "Convergence in accreditation, legal framework, and information security". These reports can be downloaded from the DwB website.

Researcher accreditation includes considering issues like

- Qualification and eligibility of the researcher as a "bona fide" researcher.
- Description of the research project.
- Is the requested data available at the RDC and indeed needed for the proposed project?
- Intended dissemination of the results.

3.3 Access to sampling frames and sensitive variables

RDCs provide access to a range of data which are likely to have been generated using different sampling frames: for example, business registers may be used for company surveys, while post/zip codes are often used to generate samples for surveys on households/individuals.

By their nature, sampling frames contain identifiable information, such as names, addresses, or even tax reference numbers. A general principle to minimise the risk of disclosure is that such identifiable sampling information, and any other variables that can be used to directly identify individuals should not be made available to researchers. This leaves the characteristics of variables as the only possible source of disclosure, which is why disclosure control procedures are implemented on all output.

By removing sampling frame data, the risk of accidental disclosure is minimised.

Minimum standard

No direct identifiers should be included in data files that are made available to researchers. Names and addresses of individuals and companies, and other identifiers including administrative codes such as health insurance numbers should be excluded.

Where a unique observation reference number is necessary (for example when constructing a panel), the identifier should be replaced by a unique but meaningless reference number.

Depending on national legislation and organisation, some RDCs may choose to provide access to identifying information in special circumstances, for example, the provision of post/zip codes for spatial research.

In all cases, irrelevant of the variables made available, the usual disclosure control rules should apply to output.

Best practice

The minimum standard applies.

3.4 Responsibility for quality of outputs

The aim of this section is to leave no doubt as to who is responsible for the content of outputs and the conclusions that are based on these outputs. In an RDC-environment the data provider cannot take this responsibility. The aim of an RDC is to allow researchers to undertake analyses for which data providers do not have the remit and/or resources. For a data provider to take responsibility for the output of these analyses it would imply an involvement in the research project and a responsibility for the conclusions and the quality of output. It should always be made very clear that outputs from an RDC are NOT official statistics, even in case data from an NSI are used.

Minimum standard

For this issue there is no minimum standard. All RDCs should comply with the best practice.

Best practice

Researchers are required to include a disclaimer in all publications, papers, presentations etc. that are (in part) based on research performed at the Research Data Centre. The exact wording of the disclaimer can be country-specific but the disclaimer should contain the following elements:

- Data from the data provider has been used in the research, with reference to the exact datasets used
- The presented results are the sole responsibility of the author(s).
- The presented results do not represent official views of the data provider nor constitute official statistics.

Examples of the exact wording in a few countries are included in Annex 1.

3.5 Checking for quality

Almost all RDCs check outputs only for statistical disclosure since quality of output is generally considered the responsibility of the researcher and not of the RDC. Having said that, many RDCs will give researchers guidance if they believe an output has some quality issues. Also, some RDCs might be tempted to judge an output on its possible negative impact on the data provider itself (for instance outputs that show some quality problems in official statistics).

Minimum standard

An RDC should clearly distinguish between comments relating to confidentiality and to quality. Ideally, comments relating to confidentiality should be delivered formally, while comments on quality are by a more informal channel (for example orally in person or by phone).

Best practice

Only check an output for confidentiality, not for quality or possible negative impact on the data provider itself. If there are concerns about output being used to embarrass a data provider, this should ideally be dealt with during the application phase. For added security, a data provider could include a clause in their access agreement which requires users to approach the data provider first when they suspect errors or quality issues in the data used.

3.6 Output documentation

Researchers often produce numerous outputs; to understand the data, for reasons of statistical data editing and to specify models. Therefore over time the amount of output which needs to be checked increases. As the majority of the research projects extend over several years and syntaxes may be sent irregularly, checkers need to frequently reacquaint themselves with the details of individual research projects.

This section aims to outline the information that researchers should provide to ensure that the checkers are able to understand and assess complex output quickly and efficiently. The requirements for the documentation of output should be communicated with researchers as part of the access agreement.

Note that adequate output documentation also benefits the researcher since it will speed up the process of output checking. This in turn means that the output will become available to the researcher outside the controlled environment of the RDC more quickly.

Minimum standard

In order to maximise the efficiency of output checking, as a minimum standard all output should include:

- the researcher's name and institute,
- the name (and number where relevant) of the research project,
- the date on which the syntax is submitted,
- a brief description of the purpose of the output,
- the data files used,
- a description of original and self-constructed variables,
- in the case of sub samples the selection criteria and the size of the sub sample,
- weighted results and unweighted frequencies.

If the output does not correspond to the minimum standard the checkers are advised to reject the output.

Best practice

For best practice an output should include all the information above, and…

- a full record (log-file) of the analysis,
- a self-explanatory documentation / annotation of the steps of analysis,
- full labelling of all variables and all value labels.

3.7 Checking every output or not?

Creating an efficient output checking process means balancing the time spent with the risk avoided. The extremes are obvious for everyone: no checking leads to large risks and zero personnel costs for checking, while checking everything reduces risk, but generates very high personnel costs for checking. There may be some situations where the time saved by checking only a sample of all outputs outweighs the added risk. Obviously, there is a benefit in this for the researchers as well. It means that they receive most of their outputs immediately, without the time delay that would be necessary for checking them.

Minimum standard

As a minimum standard all outputs that leave the controlled environment of the RDC are checked. This is to ensure a maximum data security.

Best practice

As a best practice a data provider should define an output checking strategy. As part of this strategy the relative risk and utility of subsampling output to be checked could be assessed. Some considerations to take into account when developing this strategy are the following:

- New researchers should have all their outputs checked. After a set number of non-disclosive outputs have been submitted, a researcher is labelled ‘experienced’.
- It may be possible for experienced researchers to be checked randomly and by subsample. If disclosive output is submitted, the researcher loses their ‘experienced’ status and falls back to the situation where all his output is checked.
- Outputs based on sensitive variables may not be suitable for subsampling.

3.8 Number or size of outputs

The aim of this section is to provide guidelines to prevent the RDC being buried under a pile of outputs. An RDC will typically be used by numerous researchers at the same time who all want to receive their cleared outputs as quickly as possible. However, output checking capacity is limited, so if researchers submit very many or very large outputs this increases clearing times for RDC users. An incentive should be built in to ensure that outputs are focused and valuable and that no checking-capacity is wasted. Output checking capacity is a resource that is shared by all users and should therefore be claimed in all fairness.

Minimum standard

The RDC should have (and make known) a policy that allows them to reject any output on the grounds of volume or quantity only, irrespective of its content. This policy should be explained to researchers from the start.

Best practice

Provide incentives to reduce the volume of the output to be checked. Possible solutions include:

- Allowing each research project only a limited number of free outputs (in the extreme: zero). Researchers would then pay a fixed fee (based on the average time it takes to clear an output) for each additional output.

- Charging an hourly rate for the time spent clearing an output that takes an inordinate amount of time (for instance more than 5 times the average time it takes to clear an output)
- Placing large output at the back of the queue. This means that smaller outputs will be checked first even if they were submitted at a later point in time. Without such a measure, the ‘cooperative researchers’ will be punished because they have to wait longer for their output, since the large (and therefore time consuming) output needs to be cleared first.
- Provide facilities to save intermediate results within the controlled environment and to share intermediate results between researchers on the same project. Moreover, provide tools to write reports within the controlled environment, so only the “final” output/report needs to be checked for disclosure.

It should be realised that if a financial barrier is implemented, researchers could start bundling outputs together to form one large output, so they will only pay a fee once. Setting a sensible pricing level (i.e. not too high), could go a long way in preventing this.

3.9 Output clearance times

Most organisations give guidelines but do not make a strict commitment to RDC-users on the time taken to check and return their output. Experience seems to suggest that the majority of output can be cleared in five working days. However, this depends both on the type and the size of the results submitted, as well as RDC staff resources. Defining a fixed timeframe in which researchers may expect to receive cleared output may be seen as a pressure on checkers, but the aim has to be to avoid uncertainty among users over how long they can expect to wait for output.

Within workpackage 4 of DwB, several surveys and workshops with researchers (as users of the data) showed that shortening output clearance times is ranked high on their wish list.

Minimum standard

Provide users with an indication of the average time needed to check the output with reference to the type and the volume of the submitted output, without making a commitment when this specific output will be checked.

Best practice

Response times should be monitored, and exceptions should be documented.

A commitment should be made on the maximum time a user can expect to wait for a response, either in the form of a released output or a request for further information.

It is worth noting that response time can be influenced by the way the RDC is funded. If a fee is charged for output checking, it limits the amount of output submitted and makes the users more aware of what they really need as output but, on the other side, it places an obligation on the data provider to reply in a specified timeframe.

3.10 Number of checkers

The aim of this guideline is to protect the RDC against mistakes and to ensure confidentiality of outputs. The guideline also ensures that the output checks are consistent across all checkers.

Minimum standard

As a minimum standard the output should be checked by one employee. The RDC has to ensure that the output corresponds to the legal rules for confidential data.

Best practice

The best practice is the 'four eyes principle'. Two options exist for implementing the four eyes principle. First, the two checkers are both staff members of the RDC, and second, the first check is accomplished by one RDC employee (expert in statistical analyses) and the second check is done by the subject-matter department (expert in data). This last procedure lowers the risk of disclosive output being released, because the two checkers each bring their own expertise and in that way complement each other.

3.11 Skill of checkers

Researchers often use complex statistical methods for their analyses, which generate a wide range of statistical and econometric outputs. Because output checking is time consuming and costly it is vital for efficiency that checkers fulfil a minimum standard of skill. The aim of this guideline is to ensure that checking is accurate, consistent and that checkers do not waste time understanding data and statistical methods.

Minimum standard

The RDC has to ensure that the employees fulfil the following requirements

- the technical expertise to understand most of the output,
- a working knowledge of the data held in the RDC,
- periodic training in new statistical methods combined with regular reviews.

Best practice

In addition to the requirements above, for best practice

- at least one checker should have active research experience,
- output checking should form the main part, but not all of the checker's job. This enables RDCs to employ qualified staff who might otherwise be uninterested in a job that consisted only of checking other people's work.
- To facilitate the training and review process mentioned above, a copy of all outputs and discussions relating to those outputs should be archived.

- To improve consistency as well as skills, periodic meetings to exchange experiences are helpful. Ideally this may even be done on an international level, e.g. by having a “default” session at EDAF on output checking experiences.

3.12 Researcher training

Training for RDC-users (use of the facility, legal aspects, disclosure control, etc.) is widely considered a good practice. Usually, a face-to-face contact is used to train the users, and this helps in building trust and increase security.

Minimum standard

For the minimum standard face to face training is not necessary, however documentation should be provided covering:

- the researcher’s legal responsibilities
- instructions on how to use the facility
- disclosure control principles to ensure researcher understand the issues

Best practice

For best practice a standard face-to-face training module should be developed covering the issues outlined above. It may be designed for group or individual presentation, but the criteria for such a module would be that:

- it should give researchers basic tools to use the facility without assuming that they read the documentation
- it can be delivered by more than one person without significant loss of consistency

The mentioned documentation and training should include the following topics, where the exact format can be tailored to a country’s needs:

Part 1: Process

- Legal and ethical background
    - Laws and regulations governing RDC
    - Code of ethics
- The Research Data Centre (RDC): role and purpose
    - data provider has a duty to support research
    - concerns about confidentiality risks

- RDC security model: principles

valid statistical purpose	safe projects
trusted researchers	+ safe people
anonymisation of data	+ safe data
technical controls around data	+ safe settings
disclosure control of results	+ safe outputs
	⇒ safe use

- Who can access the RDC
- RDC use:
  - description of interactions between user, data, staff, output, …
- What users cannot do:
  - attempting to remove data (including writing down data from the screen)
  - attempting to identify individuals, households, or firms
  - using data which they are not allowed to
  - using data for anything other than the proposed project
- Penalties and disciplinary actions:
  - In violation of laws
  - In violation of RDC rules / agreement
- RDC house rules
  - booking
  - charges and payment
  - user’s behaviour (keep workspace clean, do not misuse resources,…)
  - output description
  - output releasing
- Applying for access:
  - Research project description (including preliminary output description)
  - Details of people applying for access
  - Legal agreement with the institution and the researcher

    - How to submit applications / time frame
- Role of the RDC team

Part 2: Disclosure control

- General principle of disclosure control
    - finding out confidential information
    - associating that information with its source
- Key concepts
    - Two approaches: rule of thumb and principles–based model
    - Safe/unsafe classes of output
    - Primary disclosure
    - Secondary disclosure
- Rules
    - Rule of thumb for each class of output (with examples)
    - Extra attention on tabular output
        - Primary/secondary disclosure
    - Recommendations and hints for safer output
- Practicalities of disclosure control at RDC
    - Expectations on clearing times
    - Output description
    - Number and size of outputs

Annex 1. Examples of disclaimers

Researchers are required to include a disclaimer in all publications, papers, presentations etc. that are (partly) based on research performed at the Research Data Centre. The exact wording of the disclaimer can be country-specific but the disclaimer should contain the following elements:

- Data from the data provider has been used in the research
- The presented results are the sole responsibility of the author.
- The presented results do not represent official views of the data provider nor constitute official statistics.

Examples of the exact wording in a few countries are included below.

UK

This work contains statistical data from ONS which is Crown copyright and reproduced with the permission of the controller of HMSO and Queen's Printer for Scotland. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Italy

The data used in the present work stem from the Italian National Statistical Institute (Istat) – survey xxxx. Data has been processed at Istat Research Data Centre (Laboratorio ADELE).

Results shown are and remain entirely responsibility of the author; they neither represent Istat views nor constitute official statistics.

Annex 2. Output description

Each output should be described by the researcher. This description should contain at least the following items:

- Researcher's name
- The date on which the output is submitted
- The research project that the output belongs to
- Brief description of the purpose of the output
- The datafiles used

The following items could be added if need be:

- Any relation to earlier outputs (for instance if it is a small adaptation of a previous output)
- Name and location of the output file
- Email address to send the output to after clearing it
- The company that the researcher works for

Demands on the output itself

- Full labelling of all variables (including self-constructed ones)
- A description of self-constructed variables (specifying the analytical transformation or recoding applied)
- Use of subsamples/subpopulation: specify the selection criteria and size of the subsample/subpopulation
- Use of weights: show weighted and unweighted results
- For magnitude tables, report the corresponding frequency table
- For graphs, report the underlying data (or better yet, just the underlying data without the graph)
- For indices, report the formula and each single factor composing the index
- No hidden items (for instance hidden columns in excel, hidden data behind an excel graph, hidden tables in Access etc.)
- Analytical results separated from descriptive tables

Annex 3. Glossary

(most definitions were taken from http://neon.vb.cbs.nl/casc)

NSI	National Statistical Institute
Microdata	A microdataset consists of a set of records containing information on individual respondents or on economic entities.
Data provider	Any institute or body that provides (access to) microdata to researchers from outside that institute.
Research Data Centre	A common term for the part of an organisation that is responsible for providing access to their microdata for research purposes.
On Site	A facility that has been established on the premises of a data provider. 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.
Remote Access	A facility very similar to On Site. The only difference is in the access itself. For On Site, users have to come to the premises of the data provider, while for Remote Access, researchers can access the facility over a secure internet connection. Sometimes tokens or biometric devices are used for added security when logging on.
Remote Execution	Submitting scripts for execution on microdata stored within an institute's protected network. The submitter writes the scripts using the metadata of the original microdata file or with the use of a dummy dataset with the same structure as the original microdata file. Results of the script are checked against disclosure.
Output checking	The process where research results (tables, models, estimations etc.) that researchers have created and want to take out of the controlled environment of the RDC are checked for possible disclosure. Only when found non-disclosive, they are then sent to the researchers.

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. Identification: Identification is the association of a particular record within a set of data with a particular population unit. Attribution: Attribution is the association or disassociation of a particular attribute with a particular population unit.
Statistical disclosure control (SDC)	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.
Threshold rule	Usually, with the threshold rule, a cell in a frequency or magnitude table is defined to be sensitive if the number of contributors to the cell is less than some specified number. 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. The threshold rule can be set on weighted or unweighted cell counts.
Group disclosure	Group disclosure is the situation where some variables in a table (usually spanning variables) define a group of units and other variables in the table divulge information that is valid for each member of the group. Even though no individual unit can be recognised, confidentiality is breached because the information is valid for each member of the group and the group as such is recognizable

Annex 4. Explanations of methods and examples

Class 1: Frequency tables

Definition

A frequency table is made up of a number of categories across one, two, or more dimensions, with the content of the table being the number of unweighted responses that fit within the joint categories defined by each cell.

Example two dimensional table:

	Cat 1	Cat2	Total
Cat 3	X1	X2	X1+x2
Cat 4	X3	X4	X3+x4
Cat5	X5	X6	X5+x6
Total	X1+x3+x5	X2+x4+x6	X1+x2+x3+x4+x5+x6

Each X is the number of respondents in the joint category.

For example, if columns are gender, and rows three age bands, then X1 is the unweighted count of respondents who are young males.

Issues with a frequency table are:

- A count of 1 in any cell could directly reveal a respondent's information. Depending on the nature of the data and categories, it may be possible for someone to identify that individual based on the singleton response. For sure, the respondent could recognise themselves.
- A count of 1 in any cell could reveal information about the participation of an individual in a survey which may be problematic. For example if we know that our competitor is in the survey we could use that information later to determine some confidential information about that company.
- Low counts are considered risky as they could potentially lead to counts of 1 by combination with external sources. For example, if X1 is a count of three, and I know two of the respondents, X1 has revealed the existence of a third. As the count increases, the risk of this happening decreases, but never reaches zero. We select a rule of thumb threshold of 10 to account for this.
- If one cell in a table accounts for the majority of respondents in the row or column, then this potentially could reveal information. This is called group disclosure. For example if Cat 1 is 'has committed an offence', and Cat 2 is 'has not committed an offence', and X1 is 1000 and X2 is 0, then we know that all people in the survey have committed an offence. As long as this isn't a survey of offenders, it is providing information about all the respondents. If the numbers are 990

and 10, then in all likelihood, any respondent is an offender. Obviously, whether or not this is problematic depends on the categories.

- If respondents are grouped hierarchically, then it is possible that frequency tables can provide information about an entity at a higher level in the hierarchy (whilst passing all the issues above). For example, consider frequency tables of high street shops, where all counts are in excess of 10 and no group disclosure exists. If all of the shops are owned by the same enterprise, then we are revealing information about that one enterprise. Children in a school, patients in a hospital, interns in a prison, all have the potential for such disclosure.

Back to the rules of Class 1.

Class 2: Magnitude tables

Definition

A magnitude table has a similar construction to a frequency table, except that the contents of the table (the X's in the example of A4.1) comprise some numerical characteristic of the respondents – for example, income. Underlying each magnitude table is an associated frequency table that provides the number of respondents whose characteristics have contributed to the magnitude table.

Issues with a magnitude table are:

- A magnitude table can be constructed from individual respondents, hence revealing their confidential information. When provided for clearance, a magnitude table should always have the associated frequency table provided so that low counts can be checked for.
- Magnitude tables suffer exactly the same issues as for frequency tables above, particularly where the same magnitude information is known through other sources. For example business statistics provided to the RDC may also be provided to tax authorities, or the stock market.
- Dominance is a particular issue with magnitude tables. Dominance occurs when one respondent is much larger (in terms of the characteristic being reported) than the rest. Adding them up provides virtually no disguise for the largest contributor's information. We include a threshold of 50% for the largest contributor to ensure that the second largest cannot ascertain the information from the largest using knowledge of their own contribution. 50% provides an ambiguity of at least 10%.
- Dominance is particularly problematic as researchers will not provide checkers with sufficient information to check for dominance. To do so, the researcher would have to provide information on individual respondents. Situations where dominance is possible, and hence when researchers have to carry out additional checks include:
  a. Detailed industry classifications, coupled with small geographies often have a dominant business.
  b. Times or locations of very large capital investment projects can lead to dominance situations
  c. Major mergers and acquisitions can, for the year when this occurs, lead to statistics that are dominated by this business

Back to the rules of Class 2.

Class 3: maxima, minima, percentiles

Definition

A maximum is the value of a particular variable that is the largest within the sample.

A minimum is the value of a particular variable that is the smallest within the sample.

A percentile (or centile) is the value of a variable below which a certain percentage of observations fall

Thus, each of these statistics could relate to an individual record and hence be disclosive.

Issues for maxima include:

- The largest contributor in a particular sector is usually well known (e.g. the largest telecommunications provider or the oldest resident). Thus providing the data for the largest contributor is usually immediately disclosive.
- In some cases, many contributors will share the largest value. This may be disclosive if the class of people is identifiable. Imagine, for example, that all generals are paid the same salary, and we published the maximum salary within the army. This would relate to many people, but would still provide confidential returns for anyone that we know is a general.
- In contrast, there are situations where the maximum could be released if it relates to a number of unrelated respondents. For example, the maximum percentage of employees in a company who have a PC will be 100% for many companies, but wouldn’t provide any way of identifying such companies. Other information published alongside this might, of course, negate this (for example if we published the information that no companies in our sample had between 50 and 99% of employees with a PC, then if we knew that a particular company which had lot of PCs did, in fact, have 100%).

Issues for minima include:

- Generally minima are less disclosive than maxima, as usually larger things are more noteworthy, and minima tend to be zero in a lot of cases.
- The same considerations, though, must apply to ensure that we do not release confidential information.
- Where the minimum is not zero can be problematic. For example, if we released information from our sample that the minimum number of police convictions is 2, then it would reveal that all respondents had a conviction (which information was given to us in confidence).

Issues for percentiles include:

- Generally percentiles are less disclosive than either minima or maxima as their location is within the body of responses which provides some disguise. Notwithstanding this disguise, a percentile is still a piece of information given in confidence by a single respondent, and therefore caution is needed before permitting it to be released.
- For smooth distributions with a large number of respondents, percentiles will be generally safe.

- If the number of respondents is below 10, then a percentile will be an individual record (or a linear combination of two records) and in all likelihood can be identified by someone who knows about the rank ordering of the few respondents. These should not be released.
- If the distribution is bi-modal, or very highly skewed, then it is possible that a percentile could provide information about a respondent that could be identified.

A median is the 50th percentile, hence the issues of the median include those of percentiles.

Back to the rules of Class 3.

Class 4: Modes

Definition

The mode is the value that appears most often in a set of data. As it is the most frequent value, it is also likely to be the least disclosive one.

Issues for Mode include:

- If all the data points have the same value, then releasing the mode could results in a class disclosure. For example, in an organisation that employs a large pool of telephone operators all paid the same, releasing the mode would provide confidential information about the employer's pay.
- If the mode is based on less than 10 observations, then it could be disclosive in the same way that frequency data are disclosive.

Back to the rules of Class 4.

Class 5: Means, indices, ratios, indicators

Definition

Indicators are statistics derived from the data. As such, they are not the confidential data provided by respondents, but potentially such confidential data could be derived from them. Means, indices and ratios can be considered special cases of indicators.

Issues for means, indices, ratios, indicators include:

First of all, index formula should be considered.

In general, a simple index summarises the individual variable values for the statistical units in a given population:

$$I = f(X, n)$$

For the output evaluation, the index formula f and the population size n must be specified.

Sometimes the index formula is very complex, involving a lot of attributes for each unit and combining values in such a way that reverse calculation of single values is unrealistic.

As an example, consider the complexity of the Fisher price Index:

$$P_F = \sqrt{P_L P_P} = \sqrt{\frac{\sum_{j=1}^{m} p_{1,j} q_{0,j} \sum_{j=1}^{m} p_{1,j} q_{1,j}}{\sum_{j=1}^{m} p_{0,j} q_{0,j} \sum_{j=1}^{m} p_{0,j} q_{1,j}}}$$

Therefore, assuming that the index is not calculated upon very few units, complexity of data transformation is itself a reasonable protection against the disclosure of individual information from the value of the index. In the same way, complex indices (i.e. a combination of several simple indices, including ratios) are in general less disclosive than simple ones.

Nevertheless, index formula can also be very simple, e.g.:

$$a)\ \ I = \frac{\sum_{i=1}^{n} X_i}{n} \quad , \quad b)\ \ I = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{N} Y_i} \quad , \quad c)\ \ Range = X_{max} - X_{min}$$

Furthermore, the value of one or more arguments of the index formula could be easily publicly available. Note that a) corresponds to the arithmetic mean.

For instance, in cases a) and b) the denominator could be known: in such cases, the problem is reduced to the evaluation of the numerator ($\sum X_i$). This matches the issue of checking a cell of a magnitude table (if X is quantitative) or a frequency table (if X is dichotomous).

Only in this last particular case, where $X = \{0,1\}$, coherently with the case of frequency tables, the applied rule should be: $\sum X_i \geq 10$, and, since the frequency count of '0' values can also be derived, also: $n - \sum X_i \geq 10$.

Example: with X dichotomous, a mean value $\bar{X} = 0.7$ with a population size of $n = 20$ corresponds to the frequency table cells:

X = 0	X = 1	Total
6	14	20

In this example, even if the mean of X is calculated upon 20 units (more than 10) and the corresponding frequency of values '1' is 14 > 10 (rule for frequency cell counts), one can derive that there are only 6 (<10!) units having $X = 0$. However, it has to be considered whether, in this kind of situation, there is an effective risk of disclosing information upon individuals or not (see the section on frequency tables, Class 1).

About case c), with Range being a linear function of maximum and minimum, it should be released only if its components (max and min) could be released (see Class 3). In the same way, as stated before, when evaluating other comparison or dispersion indices, the formula (whether it is linear or not) and its arguments should be considered.

Nevertheless, also publicity of involved variables should be considered.

The 'Indices' category comprises a wide range of statistical results (including means, totals, etc.). From the output checking point of view indices pose an SDC problem practically quite easy to be dealt with, but theoretically analogous to that of tabular data. Therefore, the same items have to be taken into account (see Class 1 and 2), particularly the type of variables involved (if they are publicly available as often is the case for social indices) and the characteristics of the sub-population of n units involved in the index computation (if they are selected according to sensitive or identifying spanning variables).

Back to the rules of Class 5.

Class 6: Concentration ratios

Concentration ratios can be absolute and relative, they can be applied for continuous variables and categorical variables as well. Typical examples for concentration ratios are income of persons or households for social statistics. It can be turnover, import, export, investments or profit for business statistics. Concentration ratios are necessary to investigate i.e. monopolistic behaviour of enterprises or to compare concentrations in different countries. They give an overview of the distribution of income and wealth in the population, e.g. 20% of individuals hold 80% of the capital.

Especially for enterprise statistics the concentration rations can be used for re-identification, because in certain regions, size classes or NACE categories are often comprised of dominant enterprises.

Concentration ratios like GINI coefficient or Lorenz curve are usually safe as long as they meet certain requirements. These are described in the rule of thumb and principles based rule.

Back to the rules of Class 6.

Class 7: Higher moments of distributions

Higher moments like variance, skewness and kurtosis are natural phenomena in empirical data. Most of the statistical assumptions are based on normal distribution of the data. Very often this is not true and data are skewed. Those skewed distributions or even outliers could be used for re-identification, because there are not many observations around a single value, which is therefore not protected.

Back to the rules of Class 7.

Class 8: Graphs (descriptive statistics or fitted values)

Graphs are very often used for showing and visualising developments. They can be used for graphical interpretation or to illustrate coefficients in a statistical analysis. If a graph is considered safe depends on the underlying data, and whether it is possible to re-identify single values from this graph.

The amount of detail in the scales of a graph should also be considered in relation to the resolution of the graph. High detail and high resolution could lead to precise values for single contributors.

When graphical output is pasted into, e.g., a Word document, it is possible that a link is maintained between the graph and the underlying data (i.e., the underlying data are ‘embedded’). This should be taken into account when checking graphical output. The risk of embedding data in a picture is reduced by allowing only bitmap-type file formats.

Back to the rules of Class 8.

Class 9: Linear regression coefficients

Linear regression is an often used technique. In general, the release of regression coefficients is non disclosive. However, in certain (rare) situations releasing regression coefficients together with some standard summary statistics might compromise confidentiality. In Ritchie (2011), several exceptional cases are described where exact or approximate disclosure is possible.

As an example we mention the situation where on a single dataset, a regression is performed twice. The two regressions differ only in the deletion of one observation. This is for example the case when an outlier is removed from the analysis. It can then be shown that exact disclosure is possible by differencing the two sets of estimated regression coefficients and the explanatory variables are all categorical variables.

Reference:

Ritchie, F. (2011). Disclosure control for regression outputs, WISERD data resources, WISERD/WDR/005.

Back to the rules of Class 9.

Class 11: Estimation residuals

Estimation residuals can be seen as the difference between the original observations and the estimated observations based on an estimation model. As illustration: in case of linear regression the original observations are used to estimate the regression model parameters. Using that model and the observed explanatory variables, the expected dependent variable can be calculated. To assess the quality of the regression model, one often looks at the residuals $\hat{y}_i - y_i$ where $\hat{y}_i$ is the value derived from the regression model and y_i is the observed value. Obviously, releasing the residuals along with the model will lead to disclosure of the original value whenever the explanatory variables of the unit of interest is known.

Back to the rules of Class 11.

Class 13: Factor analysis

In factor analysis the aim is to describe the variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. Often, summary statistics of the factors are the output that needs to be assessed for disclosure. For example, the maximum and minimum score on a factor is often given.

In certain situations, a factor score may be equal to the score on a single observed variable. In that case the usual conditions hold (see e.g. Class 3).

Back to the rules of Class 13.