WP2 of STACE project

Grant agreement 899218 – 2019-BG-Methodology

Date

August 31, 2024

Task T2.4

Guidelines

Deliverable D2.11

Guidelines for SDC Methods for Census and Demographics Data

Sensitivity

Available to general public

Co-funded by the European Union

Background

These guidelines were written in the framework of the project on Statistical Methods and Tools for Time Series, Seasonal Adjustment and Statistical Disclosure Control (in short: STACE), co-financed by the European Union by means of grant agreement 899218 – 219-BG-Methodology. Work package 2 of this project concerned a Center of Excellence on Statistical Disclosure Control (CoE on SDC) which was formed by the national statistical institutes (NSIs) of The Netherlands, Germany, France, Austria, Iceland, Slovenia, Poland and Bulgaria.

One of the goals of this CoE was to provide three guidelines for applying SDC. The CoE, together with Eurostat and the European Expert Group on Statistical Disclosure Control, decided on three topics for these guidelines: (1) *Guidelines for SDC methods for Census and Demographics Data*, (2) *Guidelines for SDC Methods Applied on Geo-Referenced Data* and (3) *Update of the General SDC Handbook*.

A subgroup of the CoE on SDC was formed by the NSIs of The Netherlands, Germany, Slovenia and Iceland. This subgroup worked on the *Guidelines for SDC methods for Census and Demographics Data*. The current document is the result of this work.

Guidelines for Statistical Disclosure Control Methods for Census and Demographics Data

Contributing authors

All chapters are the results of several contributing authors. Either directly by providing content or indirectly by commenting earlier versions. In the list below, the chapters to which the authors provided content are mentioned. The list is sorted by the first chapter the authors contributed to.

Eric Schulte Nordholt
CBS
e.schultenordholt@cbs.nl
Chapter 1

Reinhard Tent, Sarah Giessing
Destatis
reinhard.tent@destatis.de
sarah.giessing@destatis.de
Chapters 3, 4, 5

Manca Golmajer
SURS
manca.golmajer@gov.si
Chapters 2, 5

Fabian Bach
Eurostat
fabian.bach@ec.europa.eu
Chapter 4

Marieke de Vries
CBS
mm.devries@cbs.nl
Chapters 2, 4, 5

Rob van de Laar
CBS
rwa.vandelaar@cbs.nl
Chapter 4

Peter-Paul de Wolf
CBS
pp.dewolf@cbs.nl
Chapters 2, 3

Nynke Krol
CBS
nc.krol@cbs.nl
Chapter 5

Acknowledgments

The authors would like to express their gratitude to Eurostat, in particular to Aleksandra Bujnowska and Frank Espelage, for their support during the project and for reviewing earlier versions of the guidelines. Finally, the authors would like to thank all the reviewers of the final draft for their valuable comments. We have either taken their comments into account in the current version or listed them for future work to continuously improve the guidelines.

Contents

Chapter 1

Introduction

At least once every ten years all countries in the world are recommended by the UN to conduct a Population and Housing Census. In the European Union all member states have to conduct such a census in years ending on a '1'. Historically, censuses were the only source of information about the number of people residing in a country. A number of countries used these censuses to set up population registers. Over time both the censuses and population registers contain more and more variables.

Many countries still conduct so-called traditional censuses where all information is collected via field work. However, increasingly population registers are used as a backbone for population statistics: the people who are in the population register should be enumerated. Countries with that approach have over time moved from a de facto census to a de jure census. These countries typically conduct also their demographic statistics on their population registers so that consistency between the different population statistics can be achieved. In some countries traditional censuses with field work to enumerate the population do not exist any longer as they are replaced by so-called register-based censuses. These censuses are based on information from population registers combined with information from other administrative sources to which the statistical office has access as well. In other countries combined censuses are conducted. Then some variables are taken from registers and other variables are collected via field work. We can thus conclude that in all the countries that conduct a register-based or combined census population registers play an important role when a population and housing census is conducted. These population registers are then often also used for demographic statistics.

Over time both census and demographic statistics have become more detailed. Among other things this leads to many small cell values and possible disclosures in tabular outputs. Therefore, both kinds of statistics have to be protected against disclosure of individual information. Traditionally, conventional rounding and suppressing cell values are used to protect the information. However, if the aim is to publish detailed information rounding leads to information loss and the traditional protection technique of suppressing unsafe cells leads to high numbers of suppressed cells in tables.

In the European Census 2011 many different techniques were applied by the European Union member states. Although this helped to stay within legal country frameworks, the comparability between country tables was hampered severely. This led to discussions how to improve the situation for users of these tables. Solutions should not only take into account the direct risks of disclosure, but also the risk of disclosure by differencing. Special attention is needed for grid squares (of 1 km x 1 km) tables according to national and European grid definitions. After joint projects of several European countries in a so-called Framework Partnership Agreement (FPA), in the European Statistical System (ESS) it was recommended to make use of Targeted Record Swapping (TRS) and the Cell Key Method (CKM) for Statistical Disclosure Control (SDC) of tabular output of the European Census 2021. Most countries applied at least one of these methods (see [8]).

It is clear that using the same SDC methods will lead to better comparable data. Moreover, with the recommended perturbative methods many more cells will be published than in the previous Census Round. With TRS some large differences may appear, e.g. in case large households of unequal size are swapped. However, the percentage of households to swap is normally low so that this effect is moderate. With CKM there may be small absolute differences between true values and published values in table cells. However, relative differences between true values and published values are quite small for the not too small true values.

Before the use of TRS and/or CKM to protect census tables, the outcomes of the less detailed demographic statistics and the more detailed census statistics were consistent for those countries that used the same sources and same reference dates for both types of statistics. Adding noise to census tables only and not to the demographic tables obviously may lead to small inconsistencies. Especially in case both statistics refer to the same reference day this may be confusing for users. This was indeed observed in some countries with the Census 2021. In the future a solution has to be found for that issue. Note however that, because of the randomness of the methods, using TRS and/or CKM may still lead to small inconsistencies when applied in a non-harmonised way. As at the European level under the planned future ESOP (European Statistics on Population) regulation these statistics will be brought together, it is clear that plans have to be made how to protect the ESOP output. At the national level the ongoing modernisation in many countries leads to the same need for harmonised plans to protect the output of population statistics.

Although the precise ESOP set of tables and their publication frequencies have not yet been decided, it is already known that more frequent publications than in a census context will appear and demographic statistics will become more detailed. Consistent SDC methods have to be included in the future ESOP production processes. The research done in recent years and experiences of the European Census 2021 will be of great help to work further and decide on the SDC methods for ESOP. This document can be considered as a start of this work. Moreover, this work could in the future be of help to protect more European integrated population statistics of varying detail and at different periodicities.

Given the large number of SDC methods, it is difficult to cover them all in the necessary depth in these guidelines. For this reason, we concentrate on what we consider to be the most relevant. However, we can also recommend the 'Guidelines for SDC methods applied on Geo-Referenced Data' [24] to the interested reader, in which methods more targeted at cartographic publications, such as the Quadtree Method or Spatial Smoothing, are dealt with.

The current guidelines are intended to be of help for statisticians in different countries who need to protect detailed census and demographic outputs. In this report an overview of SDC methods for census and demographic tables is given in Chapter 2. A number of current methods is discussed as well as software to apply the methodology. Also a comparison of these SDC methods is made. In Chapter 3 consistency and disclosure risk issues are presented. The difficult and country specific issue of defining parameters when using SDC methods for census and demographic tables is handled in Chapter 4. Finally, communication of SDC methods to data users is the topic of Chapter 5. Communication examples of several countries are also presented in this chapter.

Chapter 2

Overview of SDC Methods for Census and Demographic Tables

Many methods are available to NSIs wanting to protect their demographic and census tables. While the myriad of available methods are all have their own qualities and characteristics, we can broadly divide them on two fronts: the type of noise they introduce, and where in the process these methods are applied. For the first characteristics, the type of noise we introduce with our disclosure control method, we distinguish between perturbative and non-perturbative. Perturbative methods alter the data by changing a cell value or an attribute of a record just enough in order to protect the respondents. Rounding is a simple example of a perturbative method. Non-perturbative methods do not alter the data, other than suppressing values that are deemed sensitive.

The other characteristic is where in the process we apply our methods: we can choose to apply them before aggregating our data, or we can apply the method to the table that is to be published. When we apply our method to our data before aggregation we call the method pre-tabular, while alterations to the table itself are called post-tabular. Cell suppression is perhaps the most common post-tabular method.

Both axes have their advantages and disadvantages, and will differ on a method by method basis. Despite the abundance of methods available, there is not one perfect Statistical Disclosure Control method; not only because of the advantages and disadvantages for each method, but also because every NSI is dealing with different challenges, laws and demographics.

Despite this, the methods do offer general challenges and advantages, which we will outline in Section 2.3. First, we will give a concise overview of methods in Section 2.1. An important factor in deciding which methods are suitable for protecting demographic tables is the software that can be used to apply these methods; this will be covered in Section 2.2.

2.1 Current Methodology

As per the recommendations given in [6], the common methods advised for the protection of the Census 2021 are the Cell Key Method and Targeted Record Swapping. Both methods rely on data perturbation. While the Cell Key Method is post-tabular, i.e. noise is added at the level of individual tables, Targeted Record Swapping is pre-tabular - its application leads to modifications in the micro-data. In May of 2023, a workshop was organised to allow countries to share their approach to the protection of the Census 2021 and discuss connected issues. Additionally, a survey was sent out to all participants. Out of the 30 responding countries, 10 countries are using (or intend to use) both the Cell Key Method and Targeted Record Swapping. 5 countries are only using Targeted Record Swapping, while 8 only use the Cell Key Method [8].

As mentioned, not all countries decided to apply both CKM and TRS. Some applied only one, occasionally supplemented with other common statistical disclosure control methods. Other countries developed their own methodology. One example is the Multilevel Grids Method developed by INSEE [2, 8]. This method works for multiple levels of geographical grids. It works by aggregation of at-risk squares with other safe or unsafe squares until all groups contain more than a predefined number of households, working from the coarsest to the most detailed level. For more information, see [24, ch.5.1] as well as [20] or [2]. Another example of countries developing their own methodology is the special variant of Small Count Rounding, which is explained in Section 2.1.3 below.

Other methods like rounding or suppression, while applicable for Statistical Disclosure Control for demographic data, have not (yet) been widely applied to the Census 2021 data. These are however viable alternatives, or additions, to the disclosure control methods discussed above. Many more methods are available, as are variations on the methods mentioned. Due to time constraints, we will restrict ourselves to the most popular methods for the protection of demographic data. Thereby, we notably leave out the very commonly used technique of Coarsening. One reason for not including it, is that harmonisation of European demographic statistics and European census statistics leads to prescribed details and formats of tables that have to be submitted to Eurostat. Hence, Coarsening is not a valid SDC method in those situations. On the other hand, for some national demographic statistics that are not harmonized at European level, it could be a valid method.

We would also like to mention that often the application of a single SDC method is not sufficient. Each method implies its own (type of) information loss and each method targets a specific attacker scenario. See e.g. [25] for research on combining SDC methods, be it in a non-census setting.

2.1.1 Cell Key Method

The **Cell Key Method** (CKM) is a method originally designed by ABS [14], implemented in R for the package *cellKey* by [23] and implemented as an additional protection method in τ-ARGUS since version 4.2.0 [9]. The latter two implementations differ slightly from the implementation by ABS, see [14] and [23].

The Cell Key Method works for aggregation of microdata by assigning a uniformly drawn random number to each record in the microdata set, known as the record key. When a table is created by aggregation of the microdata, each cell in the table is assigned a cell key. This is done by aggregation of the record keys of all records belonging to the cell. Said cell key is then used to derive the additional noise from a predefined distribution, which is accessible through the so-called p-table. Therefore, two identical values of cell keys and cell values will give two identical amounts of noise to be added. This method preserves consistency between tables due to the consistency of the record keys: if two cells in two different tables are the aggregation of the same records, the cell key will be made up by the same record keys and therefore have the same value. As a result, the noise is consistent between tables. See [14] or [23] for a more in-depth explanation of the method.

2.1.2 Targeted Record Swapping

Targeted Record Swapping (TRS) is a probabilistic, pre-tabular perturbative method for microdata with geographical components [26]. It is designed as a special case of record swapping, where values of confidential attributes are exchanged between records. Targeted Record Swapping adds a geographical component that is used for deciding what attributes and records are swapped: households that are deemed at risk are swapped geographically with a similar household from a different region.

Which households or individuals are deemed at risk can be defined by the user. For example, households can be identified as at-risk when they contain an individual with a rare characteristic on a sensitive variable, such as religion or place of birth. The swapping method swaps households until both all at risk households are swapped and a predefined swap rate is achieved. If all at risk households are already eliminated while the swap rate is not yet achieved, the algorithm swaps regular households until the threshold is reached. Users of the method can define variables indicating similarity between households. The algorithm will then swap households that are 'similar' according to these variables. See e.g. [26] for more information on this method.

2.1.3 Small Count Rounding

Small Count Rounding is a perturbative method developed by Statistics Norway for frequency tables [22]. This method assumes only small counts are deemed unsafe. It is designed to only round cells which contain small counts, up to a user-defined number, and reduce the amount of rounding necessary.

The basic assumption is that inner cells on a low level are not published and hence the focus lies on the marginals. Now, for those marginals that are smaller than the predefined threshold, the underlying table and their inner cells (which stay unpublished) are considered and only the contributing counts are rounded, according to a specific algorithm. As the values in the underlying table represent frequencies of the unique combinations in the microdata, rounding these means changing the microdata to reflect the new frequencies. Tables created from this customised micro data set are additive and consistent. Further explanation and examples can be found in [22].

2.1.4 Cell Suppression

Cell suppression is a common non-perturbative, post-tabular SDC method that is also frequently used in the protection of census or demographic tables. It reduces the detail in the tables locally by suppressing cells, without changing the data. Sensitive cells are 'primary suppressed' and a number of safe cells are 'secondary suppressed' to make precise recalculation difficult or even impossible. As this is a non-perturbative method, this does not affect safe cells, except those chosen for secondary suppression. Some NSIs chose not to publish entire aggregation levels if they cause issues for statistical disclosure control.

2.1.5 Rounding

Traditional and **controlled rounding** are common perturbative, post-tabular methods applied to magnitude and frequency count tables. While traditional rounding works by rounding each cell to its nearest rounding base and therefore in general sums of inner cell values do not match their corresponding marginals, controlled rounding [13] works by rounding the cells so that additivity is preserved. Traditionally rounded tables may be made additive by adding up rounded cell values to the desired aggregated levels. While controlled rounding might increase noise in individual cells, it decreases noise on the aggregated levels.

2.1.6 Controlled Tabular Adjustment

Controlled Tabular Adjustment (CTA) [3, 7] is another post-tabular, perturbative method that works by generating additive synthetic tables close to the original table, where the adjusted values of confidential cells are outside of an unsafe range. The table that is closest according to a chosen measure will be released instead of the unprotected table. The method by which the synthetic data table is achieved is by solving an optimization problem, usually using linear programming methodology. Various variants are available, such as Restricted controlled tabular adjustment (RCTA) as proposed in [4]. Here adjustments to the true table are only allowed in a subset of all cells. This method allows users to control the noise more precisely. (R)CTA is applied on a per-table basis, and thus does not guarantee consistency between the tables.

2.2 Software Available for Current Methodology

To be able to apply SDC methods to census and demographic tables in an efficient way, software is needed. For all previously mentioned methods such software is indeed available, be it as a method specific kind of software or as part of a general purpose kind of software.

All R-packages mentioned in this section are available on CRAN. Moreover, of all mentioned software except the R-package `SmallCountRounding` the source code and the latest releases can be found on `https://github.com/sdctools`.

For information how to use the software, we refer to the vignettes of the respective R-packages (on CRAN), to the manuals of μ-ARGUS [16] and τ-ARGUS [17] and to the quick references for TRS in μ-ARGUS [10] and CKM in τ-ARGUS [9].

2.2.1 Method Specific SDC Software

Method specific SDC software is software that is designed to apply one single SDC method to data. Often this kind of software is optimized for application of that particular method. A slight drawback can be that, since this kind of software can only apply one particular method, several different software packages are needed when looking for the optimal solution in terms of preserving utility. Method specific software for application of SDC methods to census and demographic tables is listed in Table 2.1.

Software	Description
`SmallCountRounding`	R-package to apply Small Count Rounding
`ptable`	R-package to produce p-tables for use in Cell Key Method
`cellKey`	R-package to apply Cell Key Method

Table 2.1: Method specific SDC software (R-packages on CRAN)

2.2.2 General-purpose SDC Software

General-purpose SDC software is software that facilitates the application of multiple methods to data. Moreover, it simplifies the process of applying different methods to the same data while comparing the effects on the remaining utility. A slight drawback can be that because of the general applicability, the software might not always yield the fastest implementation. General-purpose software for the application of SDC methods to census and demographic tables is listed in Table 2.2.

2.2.3 An Overview of Software per Method

Table 2.3 gives an overview of which software is available to apply the methods mentioned in Section 2.1. For application of the Cell Key Method, so called p-tables describing the distribution of the noise are needed. They should be specified in a certain format. The R-package `ptable`, available on CRAN, can be used to produce such p-tables for use in the method specific R-package `cellKey` as well as for use in the general purpose software τ-ARGUS.

Software	Included methods for census and demographic tables
τ-ARGUS	Cell Key Method, Controlled Rounding, Controlled Tabular Adjustment, Cell suppression
`sdcTable`	Cell suppression
μ-ARGUS	Targeted Record Swapping
`sdcMicro`	Targeted Record Swapping

Table 2.2: General-purpose SDC software. Names in `this font` are R-packages.

Method	Software
Cell Key Method	`cellKey`, τ-ARGUS
Targeted Record Swapping	`sdcMicro`, μ-ARGUS
Cell suppression	`sdcTable`, τ-ARGUS
Small Count Rounding	`SmallCountRounding`
(Controlled) Rounding	τ-ARGUS
Controlled Tabular Adjustment (CTA)	τ-ARGUS

Table 2.3: Available software per method. Names in `this font` are R-packages.

2.2.4 Embedding τ-ARGUS in the Production Process

For use of τ-ARGUS in a production process, τ-ARGUS has the possibility to be called in batch-mode via commandline: `tauargus.exe foo.arb [log.txt] [temp-dir] [data-dir]`. Parameters in square brackets "`[]`" are optional: `foo.arb` is the name of the batch-file containing commands to be executed by τ-ARGUS, `log.txt` is the name for the log-file, `temp-dir` is the file path for the directory where temporary files will be stored and `data-dir` is the file path where to find data files when not fully specified in the `.arb`-file. For more information see τ-ARGUS [17], section 5.7.

This option can be used to call τ-ARGUS from any production process that is able to run a program commandline. Examples are production processes in R, Python and SAS. For R, some packages are available to facilitate running τ-ARGUS, like `rtaurgus` and `sdcTable`. They can be found at `https://github.com/sdcTools`. `sdcTable` is available through CRAN as well. For SAS some macros are available at `https://github.com/sdcTools/SAS2Argus`. For Python an experimental wrapper can be found at `https://pypi.org/project/piargus`.

2.3 Comparison of the SDC Methods

We have now covered a variety of Statistical Disclosure Control methods. In order to decide which of these methods are suitable for your own statistics, the following criteria might be useful. Some may seem more, some less important, also depending on a reader's perspective and experience.

2.3.1 Decision Criteria

First of all, an SDC method should protect the data against **primary** disclosure, i.e. it must not be possible to directly identify rare or unique frequencies, in particular 1s and 2s, and by this identify respondents as well as their attributes. This may be the case, for example, if a frequency of 1 occurs or if all respondents (of a subgroup) fall in the same category and hence certain **group attributes** can be derived.

Also, some **secondary** protection of confidentiality is needed, which assures that by doing some math (like through differencing and linking) protection of primary unsafe data cannot be undone. To this end the protection of all publications derived from the same data must be **harmonized**. Also, if third parties, like external researchers, are granted access to create their own analyses, it must be ensured, that the protection against disclosure of individual attributes is maintained.

Of course, the Statistical Disclosure Control method of choice also needs to maintain **data quality**, such that the needs of the data users can be satisfied as far as possible. To provide a high accuracy, different SDC methods as well as different parameter choices for those methods should be checked systematically.

A loss of **timeliness** due to the employment of an SDC method should be avoided by a far-sighted scheduling as far as possible. Furthermore, the choice of an SDC method should not complicate the process of combining, comparing and reconciling statistics of same origin.

If an SDC method does not preserve **additivity** or **consistency** a clear communication with the users will be needed, since otherwise they might get confused. The more **complex** a Statistical Disclosure Control method is, the more detailed and transparent the methodology and its effects on the data must be described.

To allow proper analysis of the data, **replicability** is needed, i.e. when using the same data basis and the same methods, analysis results need to be identical. In addition, one has to keep in mind, if users need to generate statistical outputs like multivariate analyses, grid data and time series it should be checked which Statistical Disclosure Control method fits the **scope**. Generally, if it should be possible to protect tabulations created flexibly on user demand, the SDC method should be **flexible** enough to even protect such additional analysis. The method should be easily applicable and ideally be carried out automatedly. This is also of importance if the data should be accessible from a **database** or from the **web**.

Additionally, the method used should also be **cost-effective** and the **cost of implementation** should be taken into account accordingly. This includes the planning phase, in which, e.g. the parameters of the method must be defined, as well as any costs incurred for adapting the existing workflow or possibly programming an automated system. The latter may also result in a **standardised IT solution**, which can have a positive impact on other statistics as well. But you also have to consider the resulting **operating costs**, be it in terms of maintenance costs or the human resources required.

The valuation of the criteria listed above can vary and be weighted differently depending on the prevailing situation and your own needs.

2.3.2 A Short Overview

For the SDC methods already explained, which could be used to protect the tabular data of the census, which consist of a large number of tables and grid tables, in Table 2.4 we show in a very abbreviated form how such a comparison of methods could look like using some of the aspects just mentioned. To simplify presentation, we sometimes only use partial aspects of these criteria, which can be categorised binary as '+' and '−'. These are namely 'simple to execute' (as a partial aspect of operating costs), 'simple to explain' (as a partial aspect of complexity), 'preserves consistency', 'preserves additivity' and also add information about the kind of protection method (perturbative versus non-perturbative) and information loss (if large differences between original and published values are possible).

In Table 2.4, next to the already introduced shorthands CKM, TRS and CTA, we will use SCR for Small Count Rounding, CS for Cell Suppression, Round for Traditional Rounding and CRound for Controlled Rounding.

	CKM	TRS	SCR	CS	Round	CRound	CTA
Simple to execute	+	+	−	−	+	+	−
Simple to explain	−	−	−	+	+	+	−
Data perturbation	+	+	+	−	+	+	+
Preserves consistency	+	+	+/−	−	+	−	−
Preserves additivity	−	+	+	+	−	+	+
Large differences	−	+	−	+/−	−	−	+

Table 2.4: Comparison of the SDC methods. A '+' means that the aspect mentioned applies, a '−' that it does not and a '+/−' that it applies to some extend.

Note that for Cell Suppression, the '+/−' on 'Large differences' means that all *published* cells are exact (no difference), but some cells are not published in which case only an approximating interval can be derived.

2.3.3 A Closer Look at the Methods

Methods which are simple to execute provide solutions for all kinds of tables; methods which are not simple to execute might have feasibility problems in case of large tables, complicated hierarchies, limitations (like, for example, some cells must not be suppressed). Cell Suppression, which can be used in a simple way for separate small tables, becomes a very challenging issue for a large number of linked tables. In practice, it is very difficult or impossible to ensure consistent protection among many large tables that are linked (e.g. Linked Tables - Modular in Tau-Argus can handle tables that have up to four dimensions, but only, if together they must not have more than ten different variables). Non-nested hierarchies for the same variable make the problem even more difficult to solve. Also, there might be a lot of suppressed cells in large tables. CTA can also be difficult to execute, because a feasible solution for a large, complex table might not be found. CKM is simple to use, because it is based on summing the record keys and finding the correct line in the perturbation table (based on the original value of the cell and the cell key). TRS is executed on microdata; after microdata are swapped, all the tables are prepared by aggregation. Traditional Rounding is simple to execute because each cell is rounded independently of the other cell. Controlled Rounding is more complicated because additivity has to be preserved. Small Count Rounding is not easy to execute for a large set of tables. Due to technical limitations of R, it might be difficult to handle large table sets with the Small Count Rounding algorithm. If a set of tables is too big to be protected straightforwardly, some tricks can be used to ensure that the model matrix doesn't exceed the limit, which is the maximum length of data vectors in R (e.g. input rows that can safely remain unchanged are removed).

Cell Suppression is easy to explain, because the values of the suppressed cells are hidden and all the other cells have original values; primary suppressions are calculated using the chosen rules, and secondary suppressions are found by the chosen algorithm which minimises the chosen criterion (e.g. the sum of the suppressed cells) and ensures certain protection requirements. Traditional Rounding and Controlled Rounding are easy to explain because after the rounding all cell values are multiples of base b. On the other hand, CTA is hard to explain because values of sensitive cells are changed by the algorithm and protected cells are not visible at first glance in the published tables. CKM and TRS are even harder to explain, because random choices are included, and the usage of these methods has several consequences. In CKM, record keys are chosen randomly and perturbations are unbiased; consequently cell keys are random, the protection of each cell is independent of the other cells, additivity is lost; the user has to have some knowledge about probability in order to understand CKM. In TRS, the process of household swapping involves randomness; consequently the tables aggregated from the swapped data can have cells with values far from original values. Small Count Rounding is hard to explain because random choices are included and because the protection depends on the list of published cells.

Cell Suppression is the only non-perturbative method in the table above; each cell value is either exact or suppressed (but for each suppressed value an interval on which the suppressed value lies can be calculated). Other methods show every table cell, but the values are perturbed (not exact) which can be a drawback (e.g. a user notices that the published value is not exact, consequently he doesn't trust the published data).

Cell Suppression preserves additivity of the tables, but preserving consistency is much harder or impossible. CTA intrinsically preserves additivity, but the process is done independently for each table, so consistent protection among tables cannot be achieved in general. CKM protects each table cell randomly and independently of each other, so additivity is not preserved, but protection is consistent because random perturbation of a specific cell depends on the units which contribute to the cell (if two cells contain the same units, their perturbation is the same). TRS preserves additivity and consistent protection among tables, because all the tables are prepared by aggregation of perturbed microdata. In Traditional Rounding, each cell is rounded independently, so the method is consistent, but not additive. In Controlled Rounding, additivity is necessary, but each table is rounded independently of the other tables, so consistency is not achieved in general. In Small Count Rounding, marginal cells are obtained by summing inner cells, therefore additivity is obtained. If tables are protected at the same time, consistent protection among the tables can be achieved, but with a lot of effort and for a limited set of tables. It is good to note that protection is only guaranteed for those cells that were considered when creating the adjusted data set. Hence, when generating new tables that were not considered by the algorithm, those might show large deviations compared to their original counterparts or may even contain sensitive cells.

Information loss due to suppression is different from information loss due to perturbation. But one should be aware that even unperturbed values are often actually estimates due to effects during data collection and compilation, and therefore not necessarily correct. In Cell Suppression, sum of suppressed cell values or number of suppressed cells is often minimised in order to minimise information loss. If CKM is used, the absolute amount of perturbation of each cell is limited strictly, controlled by a fixed parameter, which is helpful for analysis. In TRS and CTA the amount of noise can be influenced to some degree by suitable choice of the method's parameters and settings. This way the relative amount of noise for cells at higher aggregation levels can be kept at acceptable rates, but absolute amount of noise in such cells cannot be limited as strictly as with CKM. In Traditional Rounding, all the differences between original values and rounded values are small, because they can't be higher than half of base b. If Controlled Rounding is used, a cell value is not always rounded to the nearest multiple of base b, because additivity has to be preserved. Small Count Rounding tries to keep the protected values close to their original values.

Chapter 3

Consistency and Disclosure Risk Issues

3.1 Introduction

As already explained in Chapter 1, demographic and census statistics are becoming more and more detailed. Moreover, many countries nowadays draw on the same data material when producing these statistics. However, these statistics are often managed by different departments of National Statistical Offices. Additionally, these departments may have different publication schemes, i.e., they may publish the same or similar data at different points in time.

Yet another difference between census and demographic statistics may arise from differences in population definitions. At ESS level there is a lot of effort going on in trying to harmonise the population definitions. However, as long as there is no consensus, these differences may also lead to apparent inconsistencies. Anticipating the success of harmonising population definitions, we will assume the same population definition for census and demographic statistics in the remainder of this chapter.

One thus might end up with published tables that are related content wise, are based on the same data but are constructed using different methodologies. This obviously may lead to inconsistencies in the published figures what in turn may lead to increased disclosure risks. The application of different SDC methods to the different tables also influences the disclosure risk.

We would like to stress at this point that this can indeed lead to serious disclosure problems and we will discuss some of them in greater detail in this chapter. A seemingly easy way out would be to harmonise the disclosure control methods of both statistics. In particular, the possibility of using the same method in both domains should always be examined in order to avoid inconsistencies and minimise the risk of disclosure. This is especially important when the exact same breakdown combinations are used for publications for the same reference date (or reference dates that are close to each other) in both statistical products.

For the Slovenian data, for example, it is the case that the tables for national publication as well as annual demographic tables for Eurostat, partially match the Eurostat hypercubes for the 2021 census. Hence, the Statistical Office of Slovenia (SURS) decided to use the same form of protection for these parts, due to consistency. The only exception are those cells that are suppressed in the national publication and appear also in one hypercube: in this hypercube, they are protected using TRS and CKM.

What makes this harmonisation so challenging is the fact that both statistics serve different needs and have different goals. E.g., demographic data are often used for all kinds of indicators at not-too-low level of geography while census tables aim at detailed information at often detailed spatial level. Consequently, some SDC methods may be more suitable for demographics statistics but not for census data, and vice versa.

3.2 Differences in Used SDC Methods

Even though Coarsening has not been dealt with in our previous considerations, it offers an easy-to-understand way of illustrating the overall problem. One reason for not including Coarsening in Chapter 2 is that harmonisation of European demographic statistics and European census statistics leads to prescribed details and formats of tables that have to be submitted to Eurostat. Hence, Coarsening is not a valid SDC method in those situations.

To illustrate the problem of using different SDC methods in a simple way, let us nonetheless assume that demographic statistics used Coarsening as the only SDC method. Moreover, assume on the other hand that a perturbative disclosure control method is used for the census tables as recommended at ESS expert level, and one or several of the coarsened cells from the demographic statistics are also present in the census hypercubes. The values provided in the demographic statistics then can easily be compared with the values provided in the census statistics.

The protective effect of perturbative methods results from the uncertainty that is generated: none of the published values can be said with absolute certainty to match exactly with the observed value. When a data attacker is at the same time also provided with information about the true values of certain table cells through the (possibly coarsened) demographic publications, this can be used to draw further conclusions about other cells. This means that even if a published unperturbed cell value is completely innocuous on its own, a data attacker can, in the worst case, still use it to disclose sensitive cells. It is comparable to providing a Sudoku puzzle in which always only a few fields are known from the start, but the remaining fields can still be filled in correctly by someone with the appropriate skill. Whereas without the additional information the Sudoku puzzle turns into just an empty grid without any exact numbers given explicitly.

One could argue that mixing a perturbative method and a non-perturbative method is not a good idea in the first place, and that restricting oneself to using non-perturbative methods only might solve the issue. But unfortunately it is not that easy. Non-perturbative SDC methods like Coarsening and Cell Suppression are limited in their possible applications. Coarsening, for example, obviously doesn't allow detailed evaluations, which is in contrast to the goal for very high granularity of the Census publications. Cell Suppression on the other hand requires a high degree of coordination between individual tables to ensure that all logically identical table cells can be correctly identified, linked and protected consistently in all their breakdowns. This task is especially complex if certain cells are subdivided differently in several published tables, using various hierarchical breakdowns and maybe even overlapping subtotals. So, the more different, overlapping tables you want to publish from the same data material, the less information can actually stay unsuppressed. Hence, while non-pertubative SDC methods may be a good choice for the two statistics considered individually, this might not be the case anymore, if you have to consider both statistics together.

Yet that does not mean that restricting oneself to perturbative SDC methods only, is the holy grail that can be used without further precautionary measures. To illustrate this, think of a scenario where both statistics rely on the same data, but are protected with (different) perturbative disclosure control methods like e.g., the Cell Key Method and Targeted Record Swapping.

One important difference between these two methods is that CKM allows to control the amount of deviation per value that is to be published, while TRS perturbation is carried out on the micro data and hence deviations may become very large on higher aggregation levels. Moreover, TRS swaps (similar) households but does not take into account all individual characteristics of all members of the household. Choosing similarity variables and risk variables appropriately, one can influence the added noise to some extend. However, that approach controls the added noise at household level more easily than it does on person level.

Since, from a data quality perspective, large deviations are undesirable, one might tend to configure the TRS algorithm rather weak, such that only very few households get swapped. Since usually providing precise information on the swap rate is avoided, a statistical office can still rely on the fact that it is unclear where a swap actually took place and that any supposed disclosure might in fact be just a perturbed value. However, assume that at higher aggregation levels, for combinations of individual characteristics comparative values are available that were protected with CKM. It then is easy to derive whether or not a lot of micro data have been swapped using TRS: if the values for CKM and TRS are close together then TRS must have been parameterised very weakly. This realisation reduces the intended protective effect of TRS. However, if the values of the two publications differ significantly, this can cause confusion among users and reduce confidence in the data.

3.3 Differences in Reference Dates

At a first glance, differences in reference dates of the census and the demographic statistics seem to solve some of the disclosure control issues. Indeed, tables with overlapping information do not need to have the exact same figures when based on different reference dates. However, it is also problematic if both reference dates are at least close to each other. If only a short time span lies between both reference dates, differences in the original data should be relatively small and therefore the same conclusions can be drawn, with a little more potential for imprecisions. If in addition a potential data attacker is provided with further information, such as statistics on immigration and emigration as well as births and deaths, such temporal differences in the surveys can also be compensated for.

3.4 Concluding Remarks

We have shown that there are several pitfalls when disseminating tables based on (almost) the same data. The relationships between demographic and census based statistics play an important role in determining the most appropriate SDC methods to apply. It is therefore a necessity that the responsible departments for the demographic and the census tables collaborate closely when choosing the SDC methods they want to apply.

At the ESS level, the Working Group on Population and Housing Censuses has already addressed this issue in an ad-hoc solution to an operational problem that arose in the production of the 2021 census compared to the production of demographic data. In an internal paper on the Harmonised protection of census data from May 2022, usage of TRS, CKM and their combination is analysed. It was recommended to use both TRS and CKM and, if external consistency is to be established, as a practical solution, it was recommended to impute the overlapping cell values from the annual demographic tables into the census hypercubes.

Chapter 4

Defining Parameters

In general there are no guidelines or best practices with regard to choosing the parameters for TRS and CKM. There are however multiple ways of coming to a decision on which parameter values to settle on. This chapter is intended to provide readers with recommendations that can be taken into account when selecting parameters.

4.1 Parameter Definition by Trial and Error

One of those methods to decide on which parameter values to settle on is by way of trial and error. Because there are no risk measures yet defined for the *combination* of TRS and CKM, this currently seems to be the best method available to NSIs that use the combination of these methods. This section will delve into some examples which can guide the choice of the parameters.

First we delve into an example from Statistics Netherlands (SN). For TRS, first an assessment was made of how many households were at risk using the risk parameters deemed most relevant by SN. These were variables that were most sensitive if a person in the household had a less common value for it. This gave a minimum swapping rate; the method will have to swap at least this amount of households. Note that even using a lower swapping rate would swap all at-risk households, as the method is designed to continue at least until all at-risk records are swapped. However, the intention was to create more uncertainty by allowing the method to also swap records which were not deemed at-risk. It is also helpful to know the percentage of at-risk households in order to understand the results of the method; if a low swap rate is set but a larger percentage of households are swapped during the procedure, this low swap rate could give a misleading impression. SN chose to compare three values for the swapping parameter, combined with different options for CKM. The comparison of the different settings was done by utility and security experts.

It is important to note that SN used similarity profiles to apply TRS, so as to keep the aggregation of certain variables used in the similarity profiles similar even when households in regions were swapped. For example, using household size as a similarity profile variable can assure that only households of the same size are swapped, thus not affecting the number of people in a certain area. It is also possible to apply multiple similarity profiles, for when no full matching households are available; if no households are found for the largest similarity profile, matching households can be found using less strict criteria. These options can be used to influence the effect TRS has on the aggregates.

For CKM a number of different p-tables were used to evaluate the resulting tables on the sensitivity and utility aspects. The parameters for a p-table are the maximum difference between the original and the resulting cell value, the variance of the added noise and a parameter which can be used to block certain small counts (like 1's and 2') from appearing in the output tables. Several values of the maximum cell difference and of the blocked small counts were tested. The variance was tested at the smallest value possible given the other parameters of the p-table. By that, the perturbed cell values could be as close as possible to the original ones. But also larger values of the variance were tested. For instance, excluding cell values 1 and 2 resulted in seemingly no sensitive cells in the output tables, whereas excluding only value 1 resulted in apparent unsafe cells with value 2. In all cases the resulting cell values, in particular cell values 1 or 2, may or may not be the original cell values, so these 'unsafe' cells are not deducible with certainty to 1 or 2 persons in the population, so they are still protected.

Analyses of Results

Statistic Netherlands applied both methods: first TRS to the microdata and then CKM to the aggregated data for the 2021 census hypercubes and grid cells. For TRS μ-ARGUS was used and for CKM τ-ARGUS with calls from R-scripts. The R-script for the CKM protection uses as input hypercube definitions provided by the user as Excel file. The R-script derives the Argus meta data files and Argus batch files, and automatically subsequently calls τ-ARGUS so that all hypercubes are produced successively in one run. The resulting output tables were analysed and evaluation files were derived in R and saved as a csv-file that could easily be imported in Excel. A large number of evaluation measures were computed, utility and safety measures, besides a number of general table characteristics. During the project several evaluation variables were tested and adjusted if needed using other characteristics to evaluate the quality of the resulting hypercubes. Also the TRS output was analysed, to differentiate its effect on the output tables from the application of the CKM method. All measures were derived for the tables both including and excluding the table marginals, for the original hypercubes (i.e. without TRS and CKM), for the hypercubes after TRS but without CKM, and for the hypercubes after both TRS and CKM. To give an idea of measures, we give the following examples, for each hypercube characterised by its number, number of cells, and variables including hierarchies, e.g. hypercube 3.1, variable names GEO.H, SEX, AGE.M, number of table cells 14784.

- General table characteristics, examples
  - Table number
  - Table variables, including hierarchy level
  - Number of table cells
  - Number of zero cells: original, after TRS, after TRS and CKM
  - Total of all cell values: original, after TRS, after TRS and CKM
  - Minimum and maximum cell value: original, after TRS and CKM
- Safety characteristics, examples
  - Number of unsafe cells: original, after TRS, after TRS and CKM
  - Relative number of unsafe cells: original, after TRS, after TRS and CKM
- Utility characteristics, examples
  - Maximum of all cell values: original, after TRS, after TRS and CKM
  - Mean and median of all table cell values: original, after TRS and CKM
  - Maximum cell value: original, after TRS and CKM
  - Total of all cell values: original, after TRS, after TRS and CKM
  - Relative number of table cell values equal to original value: after TRS, after TRS and CKM

These measures were collected in one Excel sheet for all hypercubes. The utility and risk of each table was evaluated. Other characteristics could easily be added in the R-code script, and then run for the set of hypercubes. Also different p-tables could easily be applied and the script be run again. For 32 hypercubes and 103 tables it took for the Dutch population about two hours to process and evaluate, using a standard Statistics Netherlands virtual Windows 11 computer, Intel Xeon Gold 6146 CPU (3.2 GHz), 3 cores, 12 GB RAM.

4.2 A Deeper Look at the Parameters of the Cell Key Method

The Cell Key Method (CKM) is a perturbative disclosure control method and creates uncertainty about the true cell values to protect them.

To carry out the actual perturbation, the CKM relies on a so-called perturbation table, or p-table for short. In this table, all the necessary information is stored to assign a unique perturbation value for each combination of original value and the eponymous cell key. By this, consistency across tables is guaranteed.

It is important to choose the p-table such that the uncertainty it creates on the one hand is large enough to protect sensitive cells and on the other hand is small enough to maintain reliable data. Hence a balance must be found between information loss and disclosure risk, and the evaluation of these two aspects may vary from case to case.

To create a suitable perturbation the freely accessible R-package *ptable* can be used, which is based on a maximum entropy approach as described for example in [15]. In order to adapt the perturbation table to our needs, parameters such as the maximum deviation and the variance of the perturbation value must be set by the user. The choice made should always be reviewed for these two aspects by means of appropriate tests before CKM is used in practice.

Including High-Risk Scenarios in Considerations

Defining the parameters for the Cell Key Method always is a process that needs review and rectification. One of the main risks when it comes to the Cell Key Method is disclosure by differencing, as with any non-additive SDC method.

Since the noise of each cell is added independently, the deviation between original values and perturbed values can be controlled very strictly by the maximum deviation parameter D. This enhances the quality of the perturbed data, compared to mechanisms that modify values on the lowest level and then sum up all these deviations with each aggregation step. But this also means the results are not additive anymore. So by adding two perturbed values in most cases we will get a result that differs from the perturbed sum of the original values and an attacker can use this to their advantage in rarely occurring cases and sparse tables.

A typical example of a disclosable constellation is the following: Suppose an intruder is aware that the maximum deviation used is $D = 2$, and they also know that zero counts are never changed to non-zero counts. Now imagine a table row, which consists of two original counts of 1 each (resulting in an original margin of 2). Suppose both 1s remain unchanged, and the margin 2 is perturbed to 0, as shown in table 4.1.

	Inner Cell 1	Inner Cell 2	Margin
Original	1	1	2
Perturbed	1	1	0

Table 4.1: Example for a table row before and after perturbation.

If an attacker now knows the maximum deviation D, they can identify a range for each published value from which the corresponding original value must originate. So in our example they can conclude the margin must be at most 2 and the original interior values must both be 1. Since otherwise this would result in a margin greater than 2.

Now during the conception phase the probability that for a combination of original values as depicted above we end up with such a disclosable result can be computed easily. One just needs to look up the corresponding probabilities for each contributing cell of such an event in the p-table and multiply those to get a - for this purpose usually sufficiently precise - upper bound for the event to occur[1]. By doing this for multiple risky cells value combinations, we achieve a risk indicator for a given set of parameters. This allows us to compare multiple parameter sets.

If we want to reduce the likelihood of the above described scenario, we must be aware of the following: The probabilities of an original 2 turning into a zero and of an original 1 staying a 1 need to be small. Hence it is advantageous to raise the maximum deviation and to construct the perturbation table in a way such that the probability of an original frequency to remain unperturbed is rather low. In addition, lowering the variance makes 'extreme' deviations (like adding or subtracting the maximum possible deviation) also less likely. Yet, if both the variance and the maximum deviation are very low, very little perturbation will be carried out and hence the protective effect is low as well. So, the choice of the maximum deviation and of the variance need to be harmonised. Here we need to keep in mind that a higher variance should be accompanied by a higher maximum deviation, otherwise the value of maximum deviation will be reached too often. In any case, it is highly relevant not to set the maximum deviation D too low. In addition, this example shows, that the exact value D should never be published, since it is a valuable information for any attacker.

It is possible to obtain a more elaborate measure, if we also include the probability that a risky combination of values like in the above example occurs at all. For further insights we would like to refer the interested reader to [12].

While the previous example illustrates a risk related to the prevalence of specific value constellations in the table set, there are more generic properties of CKM protected table sets (i.e. not depending on specific constellations) that may be exploited in principle to try to disclose D and ultimately reconstruct original table values. For instance, Section 3 of [1] shows how attack methods that rely on generalised margin exploits or on the redundancy of information in large table sets ('massive averaging') can be used for a systematic, quantitative assessment of CKM parameter values. The analysis of [1] results in interdependent risk constraints on the CKM parameters D and V (variance) tailored to the specific table set to be released.

Alternatively, computation of feasibility intervals (cf. [19] , 4.3.1) for the original counts is suggested, e.g. in [11]. By using methods like linear programming, these intervals can be determined for each cell of interest. To do this, a variable representing the value of the cell is minimised and maximised, subject to a series of constraints representing the "logical" table relationships as well as some a priori upper and lower bounds for unknown cell values. This approach can be used to assess disclosure risks resulting from this kind of attack for a given parameter set. By executing the "attacks" with different values for the parameter sets we can compare the suitability of the corresponding p-tables with

[1] due to dependencies between its factors, the product of will be an upper bound of the event probability only

respect to this aspect of disclosure risk. The final parameter set can then be determined, taking into account the loss of data utility. The latter could be evaluated by comparing the original and the perturbed table, using information loss metrics based on, e.g. average distances or Hellinger's distances [18, ch. 4.7.2 and 5.8].

Chapter 5

Communication of SDC Methods to Data Users

The communication of statistical disclosure control (SDC) methods is not new, but has for a long time been primarily focused on non-perturbative methods such as suppression. It is however undeniable that the popularity of perturbative methods is growing. Because of the nature of these methods, communication of perturbative SDC methods is slightly different in its challenges. This chapter is thus applicable for all communication surrounding population statistics, but will focus especially on perturbative methods.

5.1 Introduction

Communication of SDC techniques and considerations to researchers and the public in general is an increasing priority in official statistics. Multiple challenges have been identified in communication surrounding SDC principles applied in the census 2021. First, it is vital that the value to preserve privacy is understood and shared within the NSI, between NSIs, and by the general public. Furthermore, SDC techniques can be perceived as (increasingly) advanced and obscure, which poses a challenge in the understanding what has been done and why. Also, there are some specific challenges that arise around communicating uncertainty in general which are also applicable to population statistics and specifically the census.

Epistemic uncertainty describes uncertainty about facts and numbers that could theoretically be known. Ideally, it can be quantified as some bandwidth in which we are fairly certain a number falls. This is the case for census numbers, as exact precision is generally not achieved, for example due to survey design, administrative errors, missing data, rounding errors etc. Taking in account SDC principles, another dimension is added in that it is not only impossible to eliminate uncertainty, but it may also not be fully desirable, as some minor imprecision might help protecting privacy of people. Uncertainty can even deliberately be added to data to obscure information and thereby preserve privacy. And the same confusing inconsistencies as rounding errors can arise when SDC techniques such as CKM are applied.

It is in all cases important to prepare communication of SDC principles in the census at different levels of detail for the intended audience. A useful and common practice is to identify persons to be able to adjust the message to the channel and targeted audience. For example, on social media a person might watch a short animation on the census and SDC techniques used, but in that context limited prior knowledge and a limited attention span must be assumed. In contrast, within an NSI, between NSIs, and between NSIs and researchers, a higher level of shared knowledge and focus is to be expected. It is essential to consider a communication strategy and keep your communication goals and targeted audience in mind when corresponding about census data and SDC techniques.

5.2 Examples from Statistics Slovenia

5.2.1 Communicating Usage of CKM

The Statistical Office of the Republic of Slovenia (SURS) has used the Cell Key Method (CKM) for protection of census grids. Thirteen statistics on 1 km^2 grids are prepared for European statistics. Additionally, various statistics for grids with 1 km, 500 m, and 100 m sides are prepared for national publication (total population, men, women, big age groups, 5-year age groups, education ...) in a special application called STAGE: `https://gis.stat.si/#lang=en`.

In STAGE, view and download enable adding explanations, but the policy is to include only essential explanations. When census grids were published, the decision was not to include explanation about the usage of the Cell Key Method. But later this decision was changed and explanation was included, because inconsistency and/or non-additivity effects of the usage of the Cell Key Method became obvious in the data, and had to be explained. Link to the general methodological explanation on Statistical Disclosure Control for tabular data was included (`https://www.stat.si/StatWeb/File/DocSysFile/9659/General_ME_Statistical_disclosure_control_for_tabular_data.pdf`).

For example, there was feedback from a user who has noticed that total population differs from the sum of men and women in many grids in STAGE. The user asked if it is better to use protected total population or the sum of protected men and protected women. The answer is that, in general, it is better to use the protected total population, because maximum absolute difference between the original total population and the protected total population is D, while maximum absolute difference between the original total population and the sum of protected men and protected women is 2D.

5.2.2 Communicating Usage of Cell Suppression

SURS has suppressed some too detailed parts of tables for national publication (e.g. if the number of inhabitants in a settlement is lower than a threshold t, information about their gender and age group is suppressed; but contrary, if the number of inhabitants in a settlement is equal to t or higher, information about their gender and age group is given regardless of the distribution; secondary protection is also made). This suppression is different from the classical cell suppression method implemented in τ-ARGUS and sdcTable. Suppressed cells are marked with z, which is a common sign for confidential cells in all tables for national publication. No additional explanation was made.

In general, there are few questions about confidential cells asked by the users. Each table has explanations for different statistical signs, among which there is z with its explanation *confidential*. Some users who seek data online and find confidential cells instead of numbers, write an email to the statistical office, claiming that the data are not available online and asking for the numbers. In such cases, suppression is explained to the users, and sometimes useful data are prepared (e.g. data on higher level that are not confidential). Some users that regularly use a specific table published online are unhappy if the table changes (e.g. until year 2020 there are no confidential cells, but when year 2021 is added there are some confidential cells for year 2021). In such cases, suppression is explained to the users, but it is very hard to persuade them that protection is indeed necessary.

5.3 Examples from Statistics Netherlands

Statistics Netherlands (CBS) has decided to accompany their publication of the 2021 EU census tables with three separate but connected publications. The first is a technical article covering the background of the census, the new Statistical Disclosure Control methods and some results [21]. This article is published on the CBS website for their series called "Statistical Trends". This series consists of long reads written by the researchers responsible for a statistic to give more information on the research process or provide more background and depth to the results. The target audience for these articles are users of the data and people with a more technical background who are interested in reading more about research conducted by CBS. The long read gives an overview of the two new methods used for the 2021 census (namely Targeted Record Swapping and the Cell Key Method) and discusses the impact the combination of these methods has on the published data. It does not mention parameters or other choices made by Statistics Netherlands in the protection process.

The second publication is an online article written for the corporate website of Statistics Netherlands [30]. This article is an interview with some of the researchers involved with the Census 2021 and covers some noteworthy aspects of the current census. This article includes two paragraphs on the protection of the census, but does not mention specific methodologies. It does refer to the previous article for more information on the method. The target audience is the general public.

The third publication is a video, which accompanies the aforementioned corporate article. This video has also been shared on the YouTube channel of Statistics Netherlands [5]. The target audience for this video is the general public, and the video is written in a way that is easy to understand for most Dutch speakers. The video is roughly 3 minutes and covers a number of topics: the history and necessity of the census, how the census is conducted virtually, the advantage of publishing in grids and how CBS keeps citizens' data safe. The topic of Statistical Disclosure Control is however covered only briefly in this video, and is kept without methodological details: there is only a brief mention of adding noise to the data in such a way that the results do not change too much.

5.4 Examples from Official Statistics in Germany

The Federal and the State Statistical Offices of Germany use the Cell Key Method (CKM) to protect all of their publications for the 2021 Census. To this end a webservice was implemented to allow perturbation of data retrieved by users online and it was decided to also republish the data from the 2011 Census using the same webservice and hence changing the SDC method for those data from the method "SAFE" used so far for SDC of the 2011 Census data to CKM.

To inform the users about this new SDC method and to also explain, why the old data got republished, on the official website of the German Census an information text is provided. The main point here is to explain the advantages of CKM to the prior SDC method "SAFE" and that republishing the old data with the new SDC method enhances comparability of the results of the two censuses.

In addition, when data is analysed by Destatis for external parties, the results contain the following information text:

> *In order to maintain confidentiality in accordance with article 16 of the Federal Statistics Act (BStatG), a stochastic perturbative method, the Cell Key Method (CKM), is used for analyses based exclusively on demographic data, building and housing data, household data and family data. For quality reasons, the individual values of a table row or column do not necessarily add up to the total shown. Results are only suppressed (represented in the tables with a dot) if the table - or parts of the table – comes with a too high risk of disclosure and/or too high loss of information.*

The text on the official website explains to the users that although both "SAFE" and the Cell Key Method slightly change some of the frequencies shown in the published tables, compared to their original values, for CKM this deviation is comparatively low, on average. Furthermore, users get assured that the maximum absolute deviation is also low with this method, although they don't get informed that this is due to the fact, that this value can be controlled. Additionally, original values of zero always remain unchanged, such that no implausible populations are created out of nothing. But of course, that also means that the results available in the 2011 census database may differ slightly from previous publications on the 2011 census due to the change in procedure. However, neither the quality of previously available publications nor the quality of the results available in this database are affected by the change in the confidentiality procedure.

The readers also get informed that the Cell Key Method is applied only to results based exclusively on demographic data, building and housing data, household data and family data. Results obtained from a sample survey, e.g. on employment, education and training, do not require further protection, since extrapolation and subsequent rounding to a multiple of 10 immediately prevents any conclusions being drawn about individuals.

As an exception, the official population figures for all administrative territorial units (i.e. total inhabitants) are shown with the unchanged original value, since these have legal consequences. All other figures are considered confidential, and are only published after perturbation with CKM.

A special feature of the way the data are presented is that due to the use of the Cell Key Method, the individual values of a table row or column do not necessarily add up to the total presented. This is illustrated using a simple example in which a subgroup of the population is categorised by age and gender:

The total value across all age groups for the characteristic "Male" presented in the first column of Table 5.1 is 175. However, if the corresponding table cells in the first row are added separately, their sum is 173 ($= 20 + 31 + 32 + 40 + 50$*). The total value across the entire table is shown as 371. However, if the values for "Male" and "Female" are added separately, they add up to 372 (*$= 175 + 197$*). The addition of the total values for all age groups (*$47 + 56 + 71 + 86 + 109 = 369$*) and the addition of all individual values in the table (*$20 + 31 + 32 + 40 + 50 + 25 + 25 + 40 + 45 + 60 = 368$*) also result in slightly different sums than the total value presented.*

Sex	Total	Age				
		<18	18 to 29	30 to 49	50 to 64	≥65
Male	175	20	31	32	40	50
Female	197	25	25	40	45	60
Total	371	47	56	71	86	109

Table 5.1: Example: Frequency by age and gender.

It is brought to the readers' attention that this effect is a direct consequence of the Cell Key Method and ensures in addition to the confidentiality of each individual's information the highest possible data quality. It is pointed out to the users that, whenever they need high accuracy in the aggregated data and in order to avoid small deviations to the official figures (the "official figures" are in fact the perturbed figures after application of CKM, published on the official census data base web-site of Destatis) they should use the census data base of Destatis since generating sums or differences from already perturbed results may possibly lead to larger deviations from the original results.

Another topic is the impact of CKM on statistical indicators like proportions.

> *Many interesting indicators for statistics are based on frequency counts. These include, as is now available for the 2011 census, proportions. These values are calculated using the slightly perturbed frequencies, which means that the respective results may deviate slightly from the corresponding original values. The procedure prevents conclusions about individual data and at the same time ensures the highest possible quality of results. A correction mechanism also prevents implausible results (like proportions larger than 100%).*

In addition, when using the census database, users get warned, when statistical indicators are of low reliability. This may happen when the underlying frequencies are very small, since in that case even small absolute deviations from the original values may lead to large relative deviations. So, if certain figures have a low reliability, they are shown in brackets. Hence, again, users are advised instead of performing own calculations on the noisy data, to use the census data base of Destatis for obtaining the data, to ensure a higher accuracy, to prevent implausible results and to receive such additional information on data quality.

Furthermore, for any interested readers a link to an article about disclosure control in German university statistics is given, where the Cell Key Method is described in greater detail.

5.5 Examples from Statistics Austria

Statistics Austria uses Targeted Record Swapping (TRS) to protect tables of the 2021 Census. Since this method was already used for their publications of the 2011 Census, there are now several informative texts on this topic. These texts provide information on the advantages of this method, such as consistency and maintained additivity, as well as the universal applicability of this method, including for special analyses and, of course, the comprehensibility of the procedure for users.

For the 2011 Census Statistics Austria produced a document to inform about the methodology of TRS [28]. It explains to users that so-called "risky records" are searched for first, i.e. data records that represent a rare combination in the data set due to their combination of characteristics and hence can be identified by a possible data attacker easily. Then for each of these risky records a suitable partner (B) is searched for in the remaining data, with which individual characteristics are swapped. So, instead of making two respondents change places, only some of their characteristic features get swapped. Of course, this may lead to some issues, but the reader gets assured that Statistics Austria is aware of those and took sufficient care of it. Hence it is pointed out, that it is of high importance to identify in advance which characteristics can be replaced arbitrarily and which cannot. An example given is the age which cannot be exchanged independently from the employment status without generating implausible data. Hence age categories have been defined and swapping partners are only searched within the same age groups. A simple example is used to illustrate to readers what TRS does with two selected swapping partners, but, for reasons of confidentiality, it is not stated explicitly which characteristics are chosen to be swapped in the real use case and which parameters are used. However, it is explained that, depending on the choice of characteristics to be swapped, inconsistencies/biases may occur in the data in the course of Targeted Record Swapping. For this reason, all changes are checked with the help of extensive analyses (explained within the document) in order to adjust or change the characteristics to be swapped if necessary.

For the 2021 Census further online publications have been produced, in which users are also (but not only) informed about TRS: One shows population results from the register census [29] and the other is a documentation on definitions, explanations, methods and quality [27].

Here the sections about TRS are rather short, giving an easy to understand overview for the broad readership instead of focussing too much on technical details. The main focus is on reassuring users that data protection is taken seriously, while at the same time data quality is maintained through appropriate measures. Hence Statistics Austria emphasises that the protection of personal data is one of their central concerns. It is explained in general terms that by TRS the data gets partially "soiled" to prevent the identification of individual units, so in case of smaller cell populations the data should be interpreted with caution. Yet users are assured that particular attention was paid to the quality assessment of the generated results and that care is taken to ensure that the most important key figures are unbiased. It is emphasised that important advantages of TRS over other data protection measures are that additivity and consistency of the tables is maintained, and that the possibility of subsequent analyses is given at any time without the need for further measures such as the suppression of individual values, and without the risk of the results contradicting each other. This makes further usage by users easy.

Chapter 6

User Feedback

More information will be added to this chapter once the census 2021 data of different countries have been published and feedback from the users on used SDC methods was given, compiled and communicated to Eurostat.

Bibliography

[1] Bach, F. Differential Privacy and Noisy Confidentiality Concepts for European Population Statistics. *Journal of Survey Statistics and Methodology 10*, 3 (12 2021), 642–687.

[2] Branchu, M., Costemalle, V., and Fontaine, M. Données carroyées et confidentialité.

[3] Castro, J. Minimum-distance controlled perturbation methods for large-scale tabular data protection. *European Journal of Operational Research 171*, 1 (2006), 39–52.

[4] Castro, J., and Giessing, S. Testing variants of minimum distance controlled tabular adjustment. *Monographs of Official Statistics* (2006), 333–343.

[5] Centraal Bureau voor de Statistiek. Volkstelling. `https://www.youtube.com/watch?v=BBqvqsa0_-E`, Jul 2023.

[6] CROS - European Commission. Recommendations for the protection of census data. `https://wayback.archive-it.org/12090/20231229103253/https://cros-legacy.ec.europa.eu/content/recommendations-protection-census-data_en`, Nov 2017.

[7] Dandekar, R. A., and Cox, L. H. Synthetic tabular data: an alternative to complementary cell suppression. *Manuscript, Energy Information Administration, US* (2002), 4.

[8] de Vries, M., Golmajer, M., Tent, R., Giessing, S., and de Wolf, P.-P. An overview of used methods to protect the european census 2021 tables. *UNECE Expert Meeting on Statistical Data Confidentiality* (2023).

[9] de Wolf, P.-P. Quick reference for CKM in τ-argus 4.2.x. `https://github.com/sdcTools/tauargus/releases/download/v4.2.4.2/QuickReferenceCKM4.2.x.pdf`.

[10] de Wolf, P.-P. Quick reference for TRS in μ-argus 5.1.7. `https://github.com/sdcTools/muargus/releases/download/v5.1.7b3/QuickReferenceTRS5.1.7.pdf`.

[11] Enderle, T., Giessing, S., and Tent, R. Taking into account disclosure risk in the selection ofparameters for the cell key method. *Privacy in Statistical Databases* (2018).

[12] Enderle, T., Giessing, S., and Tent, R. Calculation of risk probabilities for the cell key method. *Privacy in Statistical Databases* (2020).

[13] Fischetti, M., and Salazar, J. J. Computational experience with the controlled rounding problem in statistical disclosure control. *Journal of Official Statistics 14*, 4 (1998), 553–565.

[14] Fraser, B., and Wooton, J. *A proposed method for confidentialising tabular output to protect against differencing.* Eurostat Office for Official Publications of the European Communities, 2005, p. 299–302.

[15] Giessing, S. Computational issues in the design of transition probabilities and disclosure risk estimation for additive noise. *Privacy in Statistical Databases* (2016).

[16] Hundepool, A., de Wolf, P.-P., Bakker, J., Reedijk, A., Franconi, L., Polettini, S., Capobianchi, A., and Domingo, J. μ-argus version 5.1 user's manual, 2014.

[17] Hundepool, A., de Wolf, P.-P., Giessing, S., Salazar, J.-J., and Castro, J. τ-argus version 4.1 user's manual, 2014.

[18] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte Nordholt, E., Seri, G., de Wolf, P.-P., Tent, R., Młodak, A., Gussenbauer, J., and Wilak, K. Handbook on statistical disclosure control, second edition. STACE project WP2 Deliverable D2.10, 2024. https://sdctools.github.io/HandbookSDC.

[19] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and de Wolf, P.-P. *Statistical Disclosure Control.* Wiley, Chichester, United Kingdom, 2012.

[20] INSEE FR Lab. Gridy. https://github.com/InseeFrLab/gridy, 2024.

[21] Kuin, I., Schulte Nordholt, E., and de Vries, M. Volkstelling 2021: Nederland in vierkanten. https://www.cbs.nl/nl-nl/longread/statistische-trends/2023/volkstelling-2021-nederland-in-vierkanten. (Only available in Dutch).

[22] Langsrud, Ø., and Heldal, J. An algorithm for small count rounding of tabular data. *Privacy in Statistical Databases* (2018).

[23] Meindl, B., and Enderle, T. cellkey - Consistent perturbation of statistical tables. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, Netherlands* (2019). https://unece.or

g/sites/default/files/datastore/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S7_Austria_and_Germany_cellKey_Meindl_AD.pdf.

[24] Möhler, M., Jamme, J., de Jonge, E., Młodak, A., Gussenbauer, J., and de Wolf, P.-P. Guidelines for Statistical Disclosure Control methods applied on geo-referenced data. STACE project WP2 Deliverable D2.9, 2024. https://github.com/sdcTools/GeoSpatialGuidelinesSources/releases.

[25] Pietrzak, M., Józefowski, T., Klimanek, T., and Młodak, A. An optimized selection of statistical disclosure control methods—a case study involving microdata from the polish survey of accidents at work. In *Modern Classification and Data Analysis* (Cham, 2022), K. Jajuga, G. Dehnel, and M. Walesiak, Eds., Springer International Publishing, pp. 63–78.

[26] Shlomo, N., Tudor, C., and Groom, P. Data swapping for protecting census tables. In *Privacy in Statistical Databases* (2010), J. Domingo-Ferrer and E. Magkos, Eds., Springer Berlin Heidelberg, pp. 41–51.

[27] Statistik Austria. Standard-dokumentation metainformationen zur abgestimmten erwerbsstatistik und erwerbsstatistik der registerzählungen. https://www.statistik.at/fileadmin/shared/QM/Standarddokumentationen/B_1/std_b_abgestimmte_erwerbsstatistik.pdf.

[28] Statistik Austria. Target record swapping. https://www.statistik.at/fileadmin/pages/402/record_swapping.pdf.

[29] Statistik Austria. Zensus 2021 volkszählung – ergebnisse zur bevölkerung aus der registerzählung. https://www.statistik.at/fileadmin/user_upload/Zensus-VZ-2021.pdf.

[30] van der Sangen, M. Dutch census has been digital for 40 years. https://www.cbs.nl/en-gb/corporate/2023/27/dutch-census-has-been-digital-for-40-years.