Center of Excellence SDC

# Handbook on Statistical Disclosure Control

**Second Edition**

August, 2024

*Authors*
Anco Hundepool
Josep Domingo-Ferrer
Luisa Franconi
Sarah Giessing
Rainer Lenz
Jane Naylor
Eric Schulte Nordholt
Giovanni Seri
Peter-Paul De Wolf
Reinhard Tent
Andrzej Młodak
Johannes Gussenbauer
Kamil Wilak

Co-funded by the European Union

## Background

This handbook was 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 (NISs) 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 Germany, Austria, France, The Netherlands, Poland and Slovenia. This subgroup worked on the *Update of the General SDC Handbook*. The current document is the result of this work.

# Table of contents

**Preface to the second edition** 1

**1 Introduction** 3
- 1.1 Concepts and Definitions . . . 4
- 1.2 An approach to Statistical Disclosure Control . . . 7
- 1.3 The chapters of the handbook . . . 9
- 1.4 References . . . 9

**2 Regulations, a general Background** 10
- 2.1 Introduction . . . 10
- 2.2 Ethical codes . . . 11
- 2.3 Laws . . . 15
- 2.4 References . . . 16

**3 Microdata** 18
- 3.1 Introduction . . . 18
- 3.2 A roadmap to the release of a microdata file . . . 19
  - 3.2.1 Need of confidentiality protection . . . 21
  - 3.2.2 Characteristics and uses of microdata . . . 21
  - 3.2.3 Disclosure risk (ex ante) . . . 23
  - 3.2.4 SDC-methods . . . 25
  - 3.2.5 Implementation . . . 27
  - 3.2.6 Ex post assessment of disclosure risk and information loss . . . 27
- 3.3 Risk assessment . . . 28
  - 3.3.1 Overview . . . 28
  - 3.3.2 Disclosure risk scenarios . . . 30
  - 3.3.3 Concepts and notation . . . 31
  - 3.3.4 ARGUS threshold rule . . . 34
  - 3.3.5 ARGUS individual risk methodology . . . 34
  - 3.3.6 The Poisson model with log-linear modelling . . . 38
  - 3.3.7 SUDA . . . 40
  - 3.3.8 Record Linkage . . . 41
  - 3.3.9 References . . . 43
- 3.4 Microdata protection methods . . . 46
  - 3.4.1 Overview of concepts and methods . . . 46
  - 3.4.2 Perturbative masking . . . 48

3.4.3 Non-perturbative masking . . . 55
3.4.4 Noise addition . . . 61
3.4.5 Microaggregation: further details . . . 65
3.4.6 PRAM . . . 73
3.4.7 Synthetic microdata . . . 77
3.5 Measurement of information loss . . . 90
3.5.1 Concepts and types of information loss and its measures . . . 90
3.5.2 Information loss measures for categorical data . . . 91
3.5.3 Information loss measures for continuous data . . . 94
3.5.4 Complex measures of information loss . . . 97
3.5.5 Practical realization of trade-off between safety and utility of microdata . . . 99
3.5.6 Example . . . 100
3.5.7 References . . . 103
3.6 Software . . . 105
3.6.1 $\mu$-ARGUS . . . 105
3.6.2 sdcMicro . . . 106
3.7 Introductory example: rules at Statistics Netherlands . . . 108
3.8 Further examples . . . 110
3.8.1 Labour Force Survey . . . 111
3.8.2 Community Innovation Survey . . . 112
3.8.3 References . . . 114

**4 Magnitude tabular data 115**
4.1 Introduction . . . 115
4.2 Disclosure Control Concepts for Magnitude Tabular Data . . . 118
4.2.1 Sensitive Cells in Magnitude Tables . . . 118
4.2.2 Secondary tabular data protection methods . . . 129
4.3 The $\tau$-ARGUS Implementation of Cell Suppression . . . 134
4.3.1 Setting up a Tabular Data Protection Problem in Practice . . . 135
4.3.2 Evaluation of Secondary Cell Suppression algorithms offered by $\tau$-ARGUS . . . 140
4.3.3 Processing table protection efficiently . . . 146
4.3.4 Introductive Example . . . 152
4.4 Methodological concepts of secondary cell suppression algorithms in $\tau$-ARGUS . . . 153
4.4.1 Optimal . . . 154
4.4.2 Modular {#sec-modular} . . . 155
4.4.3 Network . . . 157
4.4.4 Hypercube . . . 157
4.5 Controlled Tabular Adjustment . . . 159
4.6 Cell Key Method for Magnitude Tables . . . 161

4.7 Measurement of disclosure risk and information loss . . . 163
4.7.1 Disclosure risk . . . 163
4.7.2 Information loss . . . 164
4.7.3 References . . . 166
4.8 References . . . 167

**5 Frequency tables** **171**
5.1 Introduction . . . 171
5.2 Disclosure risks . . . 172
5.3 Methods . . . 179
5.4 Cell Perturbation - the Cell Key Method . . . 182
5.4.1 Software implementing the Cell Key Method . . . 184
5.5 Rounding . . . 185
5.5.1 Software - How to use Controlled Rounding in $\tau$-ARGUS . . . 190
5.6 Targeted Record Swapping . . . 192
5.6.1 The TRS noise mechanism . . . 193
5.6.2 Pros and cons of targeted record swapping . . . 194
5.7 Publication of mean values . . . 195
5.8 Information loss . . . 197
5.9 References . . . 199

**6 Remote access issues** **200**
6.1 Introduction . . . 200
6.2 Research Data Centres (RDCs) . . . 200
6.3 Remote execution . . . 201
6.4 Remote access . . . 202
6.5 Licensing . . . 203
6.6 Confidentiality protection of the analysis results . . . 203
6.6.1 Output checking . . . 203
6.6.2 Rules for designing programs for controlled teleprocessing using microdata of official statistics in Germany . . . 204
6.7 References . . . 211

**Glossary** **213**

**Index** **229**

# Preface to the second edition

In 2006 Eurostat took the initiative of setting up Centres of Excellence. The idea behind this scheme is to combine the strengths of the leading National Statistical Institutes (NSIs) in Europe on a certain topic. Often in several NSIs small isolated groups are working on specific topics. Other NSIs even lack the resources to pay enough attention to certain methodological issues. This situation led to the Eurostat initiative on Centres of Excellence. A Centre of Excellence could bring together the knowledge on a certain topic at a higher level by supporting the research in the leading countries and to spread this work to the other NSIs. Statistical Disclosure Control (SDC) was selected as a pilot topic. One of the reasons for selecting SDC was a longer tradition in the field of SDC with respect to European cooperation, like the EU framework projects SDC (https://cordis.europa.eu/project/id/20462) and CASC (https://research.cbs.nl/casc/CASCIndex.htm).

One of the tasks of a Centre of Excellence on SDC launched in 2005 (grant agreement No 25200.2005.001-2005.619) was to write a handbook on Statistical Disclosure Control. The writing of the handbook was coordinated by Anco Hundepool, who also wrote parts of the chapter on the software tools (Sections 3.6, 4.3 and 4.4). Jane Longhurst and Luisa Franconi wrote the introduction chapter (Chapter 1), while Eric Schulte Nordholt wrote the regulation chapter (Chapter 2). Josep Domingo-Ferrer was responsible for the microdata chapter (Chapter 3) with contributions by Anco Hundepool, Luisa Franconi, Jane Longhurst and Peter-Paul de Wolf. Sarah Giessing was responsible for the magnitude chapter (Chapter 4) and Jane Longhurst for the frequency chapter (Chapter 5). The remote access chapter (Chapter 6) was written by Anco Hundepool and Jane Longhurst. But various other Centre of Excellence partners contributed smaller parts. Glòria Pujol Crespo and Caroline Tudor did valuable proofreading. Also the work of other partners, who did the cumbersome task of proofreading and helping with valuable comments, is gratefully acknowledged.

One of the tasks of a Centre of Excellence on SDC launched in 2020 (for 4 years), as part of the STACE project (Statistical methods and tools Centers of Excellence, grant agreement 899218–2019-BG-Methodology, 2020–2024), was to update this handbook. With permission of all the original authors, the current second edition was produced by the partners in the STACE project. The format of the handbook was revised to facilitate online as well as offline publication and some new content was added.

The original authors were asked to revise their chapters. In addition to the original authors three additional authors contributed to this second edition of the handbook. Reinhard

Tent wrote the new parts on the Cell Key Method for Magnitude Tables and Frequency Tables (Sections 4.6 and 5.4). Andrzej Młodak contributed to the parts on Risk assessment and Information loss in microdata protection, and Measurement of disclosure risk and information loss in the chapters on Magnitude Tabular Data, and Frequency tables. Johannes Gussenbauer wrote the new part on Targeted Record Swapping (Section Section 5.6). Julien Jamme led the transfer of the handbook into a new technical set-up.

Some parts of the handbook are considered to contain a bit more advanced information than necessary for a first read. Those sections are marked in this way:

> ⚠ Expert level
>
> In the online quarto-book version of the handbook these parts can be made hidden or visible by clicking the "∨" sign or the ">" sign respectively.
> In the pdf version of the handbook these parts will always be visible but marked like this.

Extended examples are marked in this way:

> ℹ **Example.** Assume a non-negative continuous variable $X$...

In spite of all the improvements and additions, the authors still see this version of the handbook as a "live document" that could be updated from time to time. Therefore, we will appreciate your comments and suggestions for further improvements. Please submit your comments or suggestions to the Issues page of the User Support repository on GitHub (https://github.com/sdctools/UserSupport).

## License

Handbook on Statistical Disclosure Control © 2024 by Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Naylor, Eric Schulte Nordholt, Giovanni Seri, Peter-Paul De Wolf, Reinhard Tent, Andrzej Młodak, Johannes Gussenbauer, Kamil Wilak, is licensed under CC BY-SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/

# 1 Introduction

National Statistical Institutes (NSIs) publish a wide range of trusted, high quality statistical outputs. To achieve their objective of supplying society with rich statistical information these outputs are as detailed as possible. However, this objective conflicts with the obligation NSIs have to protect the confidentiality of the information provided by the respondents. Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL) seeks to protect statistical data in such a way that they can be released without giving away confidential information that can be linked to specific individuals or entities.

In addition to official statistics there are several other areas of application of SDC techniques, including:

- *Health information.* This is one of the most sensitive areas regarding privacy.
- *E-commerce.* Electronic commerce results in the automated collection of large amounts of consumer data. This wealth of information is very useful to companies, which are often interested in sharing it with their subsidiaries or partners. Such consumer information transfer is subject to strict regulations.

This handbook aims to provide technical guidance on statistical disclosure control for NSIs on how to approach this problem of balancing the need to provide users with statistical outputs and the need to protect the confidentiality of respondents. Statistical disclosure control should be combined with other tools (administrative, legal, IT) in order to define a proper data dissemination strategy based on a risk management approach.

A data dissemination strategy offers many different statistical outputs covering a range of different topics for many types of users. Different outputs require different approaches to SDC and different mixture of tools.

*Tabular data protection*
Tabular data protection is the oldest and best established part of SDC, because tabular data have been the traditional output of NSIs. The goal here is to publish static aggregate information, i.e. tables, in such a way that no confidential information on specific individuals among those to which the table refers can be inferred. In the majority of cases confidentiality protection is achieved only by statistical tools due to the absence of legal and IT restrictions.

*Dynamic databases*
The scenario here is a database to which the user can submit statistical queries (sums, averages etc.). The aggregate information obtained as a result of successive queries should

not allow him to infer information on specific individuals. The mixture of tools here may vary according to the setting and the data provided.

*Microdata protection*
In recent years, with the widespread use of personal computers and the public demand for data, microdata (that is data sets containing for each respondent the scores on a number of variables) are being disseminated to users in universities, research institutes and interest groups (Unece, 2006). Microdata protection is the youngest subdiscipline and has experienced continuous evolution in the last years. If microdata are freely disseminated then statistical disclosure limitation methods will be very severe to protect confidentiality of respondents; if, on the other hand, legal restrictions are in place (such as Commission Regulation 831/2002, see section 2.3) a different amount of information may be released.

*Protection of output of statistical analyses*
The need to allow access to microdata has encouraged the creation of Microdata Laboratories (Safe Centres) in many NSI. Due to an IT protected environment, legal and administrative restrictions users may analyse detailed microdata. Checking the output of these analyses to avoid confidentiality breaches is another field which is developing in SDC research.

This handbook provides guidance on how to protect confidentiality for all of these types of output using statistical methods.

This first chapter provides a brief introduction to some of the key concepts and definitions involved with this field of work as well as a high level overview of how to approach problems associated with confidentiality.

## 1.1 Concepts and Definitions

*Disclosure*
A disclosure occurs when a person or organisation recognises or learns something that they did not know already about another person or organisation, via released data. There are two types of disclosure risk; identity disclosure and attribute disclosure.

Identity disclosure occurs with the association of a respondent's identity with a disseminated data record containing confidential information. (Duncan, et al (2001)).

Attribute disclosure occurs with the association of either an attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent. (Duncan, et al (2001)).

Some NSIs may also be concerned with the perception of disclosure risk. For example if small values appear in tabular output users may perceive that no (or insufficient) protection has been applied. More emphasis has been placed on this type of disclosure risk in recent years because of declining response rates and decreasing data quality.

*Statistical disclosure control*
Statistical disclosure control (SDC) techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. SDC methods minimise the risk of disclosure to an acceptable level while releasing as much information as possible. There are two types of SDC methods; perturbative and non-perturbative methods.

Perturbative methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. Non-perturbative methods reduce the amount of information released by suppression or aggregation of data.

A wide range of different SDC methods are available for different types of outputs. There are two types of tabular output; magnitude and frequency tables.

*Magnitude table*
In a magnitude table each cell value represents the sum of a particular response, across all respondents that belong to that cell. Magnitude tables are commonly used for business or economic data providing, for example, turnover of all businesses of a particular industry within a region.

*Frequency table*
In a frequency table each cell value represents the number of respondents that fall into that cell. Frequency tables are commonly used for Census or social data providing, for example the number of individuals within a region who are unemployed.

*Microdata*
Microdata, or unit record data, is the form from which all other data outputs are derived and is the primary form that data is stored in. While in the past NSIs simply derived outputs of other forms, more and more, microdata is becoming a key output by itself.

*Risk and Utility*
NSIs should aim to determine optimal SDC methods and solutions that minimize disclosure risk while maximizing the utility of the data. Figure 1.1 contains an R-U confidentiality map developed by Duncan, et. al. (2001) where R is a quantitative measure of disclosure risk and U is a quantitative measure of data utility.

In the lower left hand quadrant of the graph low disclosure risk is achieved but also low utility, where no data is released at all. In the upper right hand quadrant of the graph high disclosure risk is realised but also high utility, represented by the point where the original data is released. The NSI must set the maximum tolerable disclosure risk based on standards, policies and guidelines. The goal in this disclosure risk–data utility decision problem is to then find the balance in maintaining the utility of the data but reducing the risk below the maximum tolerable threshold.

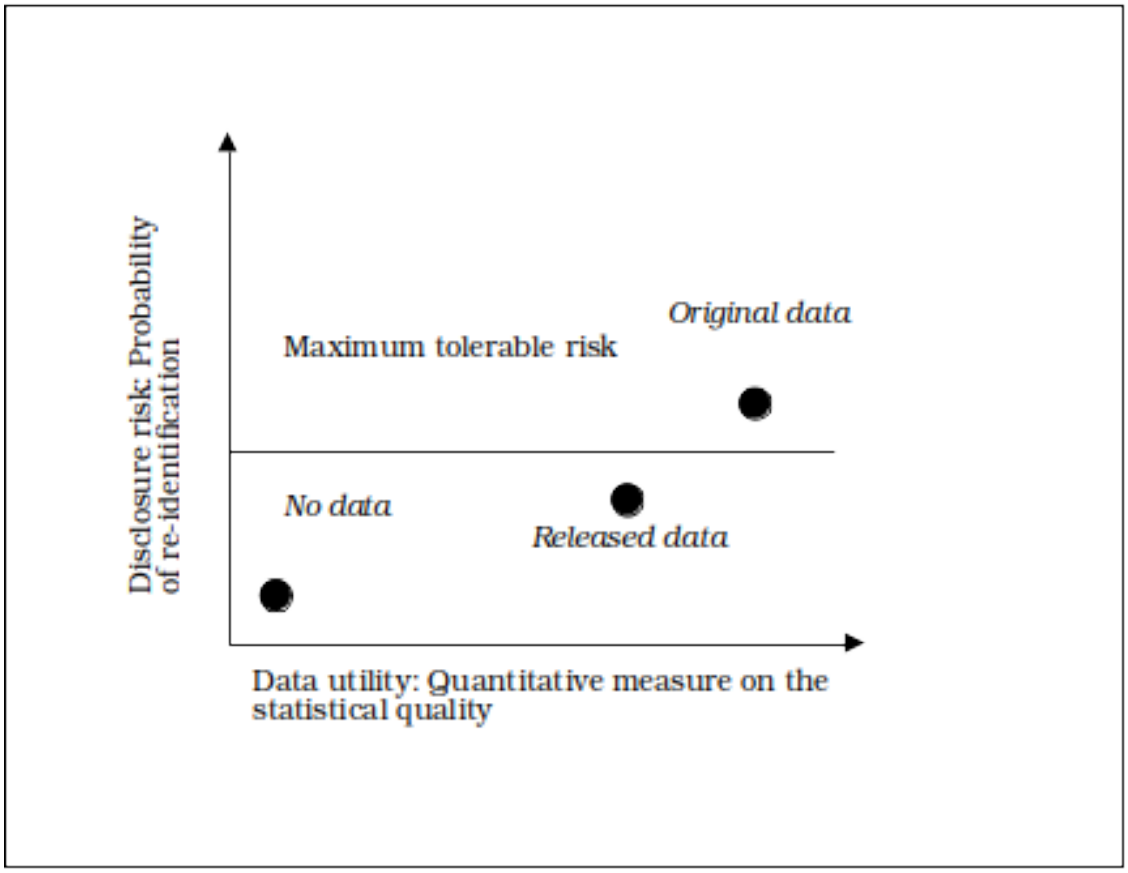


Figure 1.1: R-U Confidentiality map

## 1.2 An approach to Statistical Disclosure Control

This section describes the approach that a data provider within an NSI should take in order to meet data users' needs while managing confidentiality risks. A general framework for addressing the question of confidentiality protection for different statistical outputs is proposed based on the following five key stages and we outline how the handbook provides guidance on the different aspects of this process.

- Why is confidentiality protection needed?
- What are the key characteristics and uses of the data?
- What disclosure risks need to be protected against?
- Disclosure control methods
- Implementation

*Why is confidentiality protection needed?*
There are three main reasons why confidentiality protection is needed for statistical outputs.

- It is a fundamental principle for Official Statistics that the statistical records of individual persons, businesses or events used to produce Official Statistics are strictly confidential, and are to be used only for statistical purposes. Principle 6 of the UN Economic Commission report 'Fundamental Principles for Official Statistics', April 1992 states:
  'Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes'. The disclosure control methods applied for the outputs from an NSI should meet the requirements of this principle.
- There may be legislation that places a legal obligation on an NSI to protect individual business and personal data. In addition where public statements are made about the protection of confidentiality or pledges are made to respondents of business or social surveys these place a duty of confidence on the NSI that the NSI must legally comply with.
- One of the reasons why the data collected by NSIs is of such high quality is that data suppliers or respondents have confidence and trust in the NSI to preserve the confidentiality of individual information. It is essential that this confidence and trust is maintained and that identifiable information is held securely, only used for statistical purposes and not revealed in published outputs.

More information on regulations and legislation is provided in Chapter 2.

*What are the key characteristics and uses of the data?*
When considering confidentiality protection of a statistical output it is important to understand the key characteristics of the data since all of these factors influence both disclosure risks and appropriate disclosure control methods. This includes knowing the type of data, e.g. full population or sample survey; sample design, an assessment of quality e.g. the level

of non-response and coverage of the data; variables and whether they are categorical or continuous; type of outputs, e.g. microdata, magnitude or frequency tables. Producers of statistics should design publications according to the needs of users, as a first priority. It is therefore vital to identify the main users of the statistics, and understand why they need the figures and how they will use them. This is necessary to ensure that the design of the output is relevant and the amount of disclosure protection used has the least possible adverse impact on the usefulness of the statistics. Section 3.2 addresses some examples on how to carry out this initial analysis.

*What disclosure risks need to be protected against?*
Disclosure risk assessment then combines the understanding gained above with a method to identify situations where there is a likelihood of disclosure. Risk is a function of likelihood (related to the design of the output), and impact of disclosure (related to the nature of the underlying data). In order to be explicit about the disclosure risks to be managed one should consider a range of potentially disclosive situations or scenarios and take action to prevent them. A disclosure scenario describes (i) which information is potentially available to an 'intruder' and (ii) how the intruder would use the information to identify an individual. A range of intruder scenarios should be determined for different outputs to provide an explicit statement of what the disclosure risks are, and what elements of the output pose an unacceptable risk of disclosure. Issues in developing disclosure scenarios are provided in Section 3.3.2. Risk assessment methods for microdata are covered in Section 3.3 and different rules applied to assess the risk of magnitude and frequency tables are described in Chapter 4 and 5 respectively.

*Disclosure control methods*
Once an assessment of risk has been undertaken an NSI must then take steps to manage any identified risks. The risk within the data is not entirely eliminated but is reduced to an acceptable level, this can be achieved either through the application of statistical disclosure control methods or through the controlled use of outputs, or through a combination of both. Several factors must be balanced through the choice of approach. Some measure of information loss and impact on main uses of the data can be used to compare alternatives. Any method must be implemented within a given production system so available software and efficiency within demanding production timetables must be considered. Statistical disclosure control methods used to reduce the risk of microdata, magnitude tables and frequency tables are covered in Chapter 3, 4 and 5 respectively. Chapter 6 provides information on how disclosure risk can be managed by restricting access.

*Implementation*
The final stage in this approach to a disclosure control problem is implementation of the methods and dissemination of the statistics. This will include identification of the software to be used along with any options and parameters. The proposed guidance will allow data providers to set disclosure control rules and select appropriate disclosure control methods to protect different types of outputs. The most important consideration is maintaining confidentiality but these decisions will also accommodate the need for clear, consistent and practical solutions that can be implemented within a reasonable time and using available

resources. The methods used will balance the loss of information against the likelihood of individuals' information being disclosed. Data providers should be open and transparent in this process and document their decisions and the whole risk assessment process so that these can be reviewed. Users should be aware that a dataset has been assessed for disclosure risk, and whether methods of protection have been applied. For quality purposes, users of a dataset should be provided with an indication of the nature and extent of any modification due to the application of disclosure control methods. Any technique(s) used may be specified, but the level of detail made available should not be sufficient to allow the user to recover disclosive data. Each chapter of the handbook provides details of software that can be used to assess and manage disclosure risk of the different statistical outputs.

## 1.3 The chapters of the handbook

This book starts with an overview of regulations describing the legal underpinning of Statistical Disclosure Control in Chapter 2. Microdata are covered in Chapter 3, magnitude tables are addressed in Chapter 4 and Chapter 5 provides guidance for frequency tables. Chapter 6 describes the confidentiality problems associated with microdata access issues. Within each section different approaches to assessing and managing disclosure risks are described and the advantages and disadvantages of different SDC methods are discussed. Where appropriate recommendations are made for best practice.

In Chapter 7 a glossary of statistical terms used in Statistical Disclosure Control has been included.

## 1.4 References

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001) *Disclosure Risk vs. Data Utility: the R-U Confidentiality Map,* Technical Report LA-UR-01-6428, Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory

Trewin, D et al. (2007) *Principles and Guidelines of Good Practice for Managing Statistical Confidentiality and Microdata Access*, UNECE United Nations Economic commission for Europe: https://unece.org/…/Managing.statistical.confidentiality.and.microdata.access.pdf

# 2 Regulations, a general Background

## 2.1 Introduction

Up to the late 1980's microdata were rarely sent to Eurostat, the European statistical office. There was a general reliance on submission by National Statistical Institutes (NSIs) of agreed tabular data. National confidentiality rules in some of the European countries made it impossible to harmonise European statistics. This was an unwanted situation for all NSIs, and especially for Eurostat. Therefore, a regulation on the transmission of confidential data to Eurostat has been prepared and was finally adopted by the Council in June 1990 as Regulation 1588/90.

*Committee on Statistical Confidentiality*
In January 1994, these measures have been defined and formally adopted by the Member States through the Committee on Statistical Confidentiality (CSC). This Committee met at least once a year at the Eurostat office in Luxembourg. This Committee discussed the implementation and evaluation of European Regulations on the dissemination of microdata and tabular data. Also revisions to the basic statistical legal framework were considered. The last meeting of the CSC was held in 2008.

Another relevant Council Regulation is No 322/97 of February 1997. This Regulation defined the general principles governing Community statistics, the processes for the production of these statistics and established detailed rules on confidentiality. This Regulation could be considered as the general statistical law of the European Union.

*European Statistical System Committee*
A new statistical legal framework was introduced in 2009. One of the new aspects concern statistical confidentiality: the need to enhance the role of the NSIs and Eurostat for organisational, co-ordination and representation purposes was noted. In this context the former Statistical Programme Committee was replaced by a new Committee, the European Statistical System Committee (ESSC). This new Committee is also entrusted with the functions of the CSC, which thus ceased to exist.

The European Statistical System (ESS) is defined by Regulation 223/2009 on European statistics on 1 April 2009 as the partnership between the Community statistical authority (the Commission (Eurostat)) and all national authorities responsible for the development, production and dissemination of European Statistics (ES). Regulation 223/2009 was amended by Regulation 2015/759 of the European Parliament and the Council in order

to further strengthen the governance of the ESS, in particular its professional independence.

Currently, Regulation 223/2009 is being replaced by a more modern version that gives e.g. better possibilities for the NSIs to use privately held data. However, the new regulation will not include principle changes regarding the situation of Statistical Disclosure Control in the European Union.

The availability of confidential data for the needs of the ESS is of particular importance in order to maximise the benefits of the data with the aim of increasing the quality of European statistics and to ensure a flexible response to the newly emerging Community statistical needs.

The transmission of confidential data between ESS partners is allowed if necessary for the production, development and dissemination of ES and also for increasing the quality of these statistics. The conditions for their further transmission, in particular for scientific purposes, are also strictly defined.

The ESSC is consulted on all draft comitology measures submitted by the Commission in the domain of statistical confidentiality.

This section gives some background by discussing a few ethical codes and laws. It does not contain any national specialities.

## 2.2 Ethical codes

Many Member States have an ethical code that forms the basis of the production of official statistics.

*ISI Declaration on Professional Ethics*
After an intense preparation process taking place from 1979 to 1985, the International Statistical Institute (ISI) adopted the ISI Declaration on Professional Ethics in 1985. A newer version was adopted in 2010 by the ISI Council. Finally, an updated version was endorsed by the ISI Executive Committee. Whilst the 2010 Declaration content remains largely valid, the increasing use of a diversity of data sources, linked data sets and computationally heavy statistical methods has required updates introduced in 2023.

*European Statistics Code of Practice*
On 24 February 2005 the Statistical Programme Committee adopted the European Statistics Code of Practice. On 17 November 2017 a renewed version of this Code was adopted by the European Statistical System Committee (ESSC). This Code of Practice has the dual purpose of:

- Improving trust and confidence in the independence, integrity and accountability of both National Statistical Authorities and Eurostat, and in the credibility and quality of the statistics they produce and disseminate (i.e. an external focus);

- Promoting the application of best international statistical principles, methods and practices by all producers of European Statistics to enhance their quality (i.e. an internal focus).

The Code of Practice is based on 15 Principles. Governance authorities and statistical authorities in the European Union commit themselves to adhering to the principles fixed in this code and to reviewing its implementation periodically by the use of Indicators of Good Practice for each of the 15 Principles, which are to be used as references. Principle 5 concerns statistical confidentiality and is cited below.

> Principle 5: Statistical Confidentiality
>
> *The privacy of data providers, the confidentiality of the information they provide, its use only for statistical purposes and the security of the data are absolutely guaranteed.*

Indicators

- 5.1 Statistical confidentiality is guaranteed in law.
- 5.2 Staff sign legal confidentiality commitments on appointment.
- 5.3 Penalties are prescribed for any wilful breaches of statistical confidentiality.
- 5.4 Guidelines and instructions are provided to staff on the protection of statistical confidentiality throughout the statistical processes. The confidentiality policy is made known to the public.
- 5.5 The necessary regulatory, administrative, technical and organisational measures are in place to protect the security and integrity of statistical data and their transmission, in accordance with best practices, international standards, as well as European and national legislation.
- 5.6 Strict protocols apply to external users accessing statistical microdata for research purposes.

*UNECE principles and Guidelines of Good Practice for Managing Statistical Confidentiality and Microdata Access*
The 2003 Conference of European Statisticians (CES) of the United Nations Statistical Commission for Europe installed a Task Force, chaired by Dennis Trewin (at that time the Australian Statistician), to draft Principles and Guidelines of Good Practice for Managing Statistical Confidentiality and Microdata Access. In their final report of 2007 the following two key objectives in these guidelines are mentioned:

- To foster greater uniformity of approach by countries whilst facilitating better access to microdata by the research community for worthwhile papers;
- Through these guidelines and supporting case studies, to enable countries to improve their arrangements for providing access to microdata.

The sixth United Nations Fundamental Principle of Official Statistics, which was mentioned in Section 1.2 of this handbook, is very clear on statistical confidentiality: "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes". Any principles for microdata access must be consistent with this Fundamental Principle.

According to the report by Trewin c.s. the following principles should be used for managing the confidentiality of microdata. Each is discussed in the following paragraphs.

> Principle 1: It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected.
>
> Principle 2: Microdata should only be made available for statistical purposes.
>
> Principle 3: Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected.
>
> Principle 4: The procedures for researcher access to microdata, as well as the uses and users of microdata, should be transparent and publicly available.

Making microdata available for research is not in contradiction with the sixth UN Fundamental Principle as long as it is not possible to identify data referring to an individual. Principle 1 does not constitute an obligation to provide microdata. The National Statistical Office should be the one to decide whether to provide microdata or not. There may be other concerns (for example, quality) that make it inappropriate to provide access to microdata. Or there may be specific persons or institutions to which it would be inappropriate to provide microdata.

For Principle 2, a distinction has to be made between statistical or analytical uses and administrative uses. In the case of statistical or analytical use, the aim is to derive statistics that refer to a group (be it of persons or legal entities). In the case of administrative use, the aim is to derive information about a particular person or legal entity to make a decision that may bring benefit or harm to the individual. For example, some requests for data may be legal (a court order) but inconsistent with this principle. It is in the interest of public confidence in the official statistical system that these requests are refused. If the use of the microdata is incompatible with statistical or analytical purposes, then microdata access should not be provided. Ethics committees or a similar arrangement may assist in situations where there is uncertainty whether to provide access or not.

Researchers are accessing microdata for research purposes but to support this research they may need to compile statistical aggregations of various forms, compile statistical distributions, fit statistical models, or analyse statistical differences between sub-populations. These uses would be consistent with statistical purposes. To the extent that this is how the microdata are being used, it could also be said to support research purposes.

With respect to Principle 3, legal arrangements to protect confidentiality should be in place before any microdata are released. However, the legal arrangements have to be complemented with administrative and technical measures to regulate the access to microdata and to ensure that individual data cannot be disclosed. The existence and visibility of such arrangements (whether in law or supplementary regulations, ordinances, etc.) are necessary to increase public confidence that microdata will be used appropriately. Legal arrangements are clearly preferable but in some countries this may not be possible and some other form of administrative arrangement should be put in place. The legal (or other arrangements) should also be cleared with the privacy authorities of countries where they exist before they are established by law. If such authorities do not exist, there may be NGOs who have a "watchdog" role on privacy matters. It would be sensible to get their support for any legal or other arrangements, or at least to address any serious concerns they might have.

In some countries, authorising legislation does not exist. At a minimum, release of microdata should be supported by some form of authority. However, an authorising legislation is a preferable approach.

Principle 4 is important to increase public confidence that microdata are being used appropriately and to show that decisions about microdata release are taken on an objective basis. It is up to the NSO to decide whether, how and to whom microdata can be released. But their decisions should be transparent. The NSO web site is an effective way of ensuring compliance and also for providing information on how to access research reports based on released microdata.

The guidelines of the report were endorsed by the CES plenary session in 2006. They addressed the need to unify the approaches internationally and to agree on core principles for dissemination of microdata. They also suggested moving towards a risk management rather than a risk avoidance approach in the provision of microdata.

The report originally contained an annex with 22 case studies describing good practices in different countries. It is a dynamic document that is updated from time to time.

*UNECE Principles and Guidelines on Confidentiality Aspects of Data Integration*
In 2007 and 2008 a CES Task Force chaired by Brian Pink, at that time the Australian Statistician, drafted Principles and Guidelines on Confidentiality Aspects of Data Integration.

Data integration is concerned with integrating unit record data from different administrative and/or survey sources to compile new official statistics which can then be released in their own right. In addition these integrated data sets may be used to support a range of economic and social research not possible using traditional sources.

The drafted principles and associated guidelines expand on the sixth UN Fundamental Principle by providing a common framework for assessing and mitigating legislative and other confidentiality aspects of the creation and use of integrated datasets for statistical

and associated research purposes. In particular they recognise that the fundamental principles of official statistics apply equally to integrated data sets as to any other source of official statistics.

In developing these principles, it is recognised that integration of statistical data sets has become a normal part of the operations of a number of statistical offices and is generally most advanced in those countries where a heavy reliance is placed on obtaining statistical information from administrative registers. Countries that regularly undertake statistical integration usually already have a strong legislative basis and clear rules about protection of the confidentiality of personal and individual business data irrespective of whether the data has been integrated from different sources or not.

However, for many other countries the notion of integrating data from different sources for statistical and related research purposes is relatively new. The drafted principles and associated guidelines are designed to provide some clarity and consistency of application.

These Principles and Guidelines were endorsed by the CES at their June 2009 meeting.

## 2.3 Laws

For statistical disclosure control the following two laws are of importance.

- **Commission Regulation (EC) No 223/2009 of the European Parliament and Council of 11 March 2009 on European statistics.** This regulation has entered into force on the 1 April 2009 and replaced the Regulation 322/97 on Community Statistics and Regulation 1101/2008 (previously 1588/90) on the transmission of data subject to statistical confidentiality.

This Regulation establishes the legal framework for the development, production and dissemination of European statistics, including the rules on confidentiality.

Article 2, clause 1(e) defines statistical confidentiality as ”the protection of confidential data related to single statistical units which are obtained directly for statistical purposes or indirectly from administrative or other sources and implying the prohibition of use for non-statistical purposes of the data obtained and of their unlawful disclosure”.

Confidential data are defined as ”data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all relevant means that might reasonably be used by a third party to identify the statistical unit”.

Chapter V 'Statistical Confidentiality' describes in detail rules and measures that shall apply to ensure that confidential data are exclusively used for statistical purposes and how their unlawful disclosure shall be prevented (Articles 20 -26). Article 23, in particular, makes provision for the access to confidential data for scientific purposes.

- **Commission Regulation (EC) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002.** This Regulation establishes the conditions under which access to confidential data transmitted to the Commission (Eurostat) may be granted for enabling statistical analyses for scientific purposes, and the rules of cooperation between the Commission (Eurostat) and national statistical authorities in order to facilitate such access.

In Article 3 general principles are described. The Commission (Eurostat) may grant access to confidential data for scientific purposes held by it for the development, production or dissemination of European statistics as referred to in Article 1 of Regulation (EC) No 223/2009, provided that the following conditions are satisfied:

- access is requested by a recognized research entity;
- an appropriate research proposal has been submitted;
- the requested type of confidential data for scientific purposes has been submitted;
- access is provided either by the Commission (Eurostat) or by another access facility accredited by the Commission (Eurostat);
- the relevant national statistical authority which provided the data has given its approval.

The original Regulation 831/2002 covered four surveys: Labour Force Survey (LFS), Continuing Vocational Training Survey (CVTS), European Community Household Panel (ECHP) and Community Innovation Survey (CIS). Later other surveys were added and the European Union Statistics on Income and Living Conditions (EU-SILC) replaced the ECHP.

Discussions about the level of detail for the microdata (to which the researchers get access) take place in the relevant Working Groups. Currently, the task to approve this level of detail is delegated to the Directors' Group on Methodology (DIME). In addition to the DIME two Expert Groups in the ESS exist that deal with Statistical Disclosure Control (SDC): the Expert Group on SDC and the Microdata Access Network Group (MANG).

## 2.4 References

Commission Regulation (EC) No 831/2002 concerning access to confidential data for scientific purposes
https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32002R0831

Commission Regulation (EC) No 223/2009 on European statistics
http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32009R0223

Commission Regulation (EC) No 557/2013 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013R0557

Council Regulation 1588/90
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31990R1588

Council Regulation 322/97
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31997R0322

European Statistics Code of Practice https://ec.europa.eu/eurostat/web/products-catalogues/-/european-statistics-code-of-practice-revised-edition-2017

ISI (1985) *Declaration on Professional Ethics:* http://isi-web.org/declaration-professional-ethics

Pink, B et al. (2009) *Principles and Guidelines on Confidentiality Aspects of Data Integration* UNECE United Nations Economic commission for Europe
https://digitallibrary.un.org/record/651518?v=pdf

Trewin, D et al. (2007) *Principles and Guidelines of Good Practice for Managing Statistical Confidentiality and Microdata Access* UNECE United Nations Economic commission for Europe
https://unece.org/…/Managing.statistical.confidentiality.and.microdata.access.pdf

# 3 Microdata

## 3.1 Introduction

There is a strong, widespread and increasing demand for NSIs to release Microdata Files (MF), that is, data sets containing for each respondent the score on a number of variables. Microdata files are samples generated from business or social surveys or from the Census or originate from administrative sources. It is in the interest of users to make the microdata as detailed as possible but this interest conflicts with the obligation that NSIs have to protect the confidentiality of the information provided by the respondent.

In Section 1.1 two definitions of disclosure were provided: re-identification disclosure and attribute disclosure. In the microdata setting the re-identification disclosure concept is used as we are releasing information at individual level. When releasing microdata, an NSI must assess the risk of re-identifying statistical units and disclosing confidential information. There are different options available to NSIs for managing these disclosure risks, namely applying statistical disclosure control techniques, restricting access or a combination of the two.

Applying SDC methods leads to a loss of information and statistical content and affects the inferences that users are able to make on the data. The goal for an effective statistical disclosure control strategy is to choose optimum SDC techniques which maximize the utility of the data while minimizing the disclosure risk. On the other hand, the user of protected microdata should obtain most important information on the expected information loss due to SDC process. It enables the assessment of the impact of changes in the original data resulting from the need to protect statistical confidentiality on the quality of the final results of the estimates and analyzes carried out by him/her.

In general, two types of microdata files are released by NSIs, namely public use files (PUF) and research use files (MFR). The disclosure risk in public use files is entirely managed by the design of the file and the application of SDC methods. For research use microdata files SDC methods will be applied in addition to some restrictions on access and use, e.g. under a licence or access agreement, such as those provided by Commission Regulation 831/2202, see Section 2.3. Necessarily the research release files contain more detail than the public use files. The estimated expected information loss should be computed both as total and for each variable separately, if possible.

Some NSIs will also provide access to microdata in datalaboratories/research centres or via remote access/execution. Datalabs allow approved users on-site access to more identifiable

microdata. Typically datalab users are legally prohibited from disclosing information and are subject to various stringent controls, e.g. close supervision on-site to protect the security of the data and output checking, to assist with disclosure control. For remote execution researchers are provided with a full description of the microdata. They then send prepared scripts to the NSI who run the analysis, check and return the results. Remote access is a secure on-line facility where the researchers connect to the NSI's server (via passwords and other security devices) where the data and programs are located. The researchers can submit code for analysis of microdata or in some instances see the files and programs 'virtually' on their desktops. Confidentiality protection is by a combination of microdata modification, automatic checks on output requested, manual auditing of output and a contractual agreement. The researcher does not have complete access to the whole data itself; however they may have access to a small amount of unit record information for the purpose of seeing the data structure before carrying out their analysis.

Section 3.2 goes through the whole process of creating a microdata file for external users from the original microdata. The aim of such section is to briefly analyse the different stages of the disclosure process providing references to the relevant sections where each step will be described in more details. Section 3.6 is dedicated to the software. Sections 3.7 and 3.8 provide some examples. Further and more detailed examples can be found in Case studies available on the (CASC-website). Chapter 6 of this handbook provides more details on different microdata access issues such as research centres, remote access/execution and licensing.

## 3.2 A roadmap to the release of a microdata file

This section aims at introducing the reader to the process that, starting from the original microdata file as it is produced by survey specialists, ends with the creation of a file for external users. This roadmap will drive you through the six stage process for disclosure, mostly outlined in Section 1.2 i.e.

1. why is confidentiality protection needed;
2. what are the key characteristics and use of the data;
3. disclosure risk (ex ante);
4. disclosure control methods;
5. implementation;
6. assessment of disclosure risk and utility (ex post)

Specifying, for each stage, the peculiarities of microdata release. In Table 3.1 we present an overview of the process.

| **Stage of disclosure process** | **Analyses to be carried out / problem to be addressed**<br>⇓<br>***Results expected*** |
|---|---|
| 1. Why is confidentiality protection needed | Does the data refer to individuals or legal entity?<br>⇓<br>*We need to protect the statistical unit* |
| 2. What are the key characteristics and use of the data | Analysis of the type/structure of the data<br>⇓<br>*Clear vision of which units need protections* |
| | Analysis of survey methodology<br>⇓<br>*Type of sampling frame, sample/complete enumeration of strata, further analysis of survey methodology, calibration* |
| | Analysis of NSI objectives<br>⇓<br>*Type of release (PUF, MFR),dissemination policies, peculiarities of the phenomenon, coherence between multiple releases (PUF and MFR), coherence with released tables and on-line databases, etc.* |
| | Analysis of user needs<br>⇓<br>*Priorities for variables, type of analysis, etc.* |
| | Analysis of the questionnaire<br>⇓<br>*List of variables to be removed, variables to be included, some ideas of level of details of structural variables* |
| 3. Disclosure risk (ex ante) | Disclosure scenario<br>⇓<br>*List of identifying variables* |
| | Definition of risk<br>⇓<br>*Risk measure* |
| | Risk assessment<br>⇓<br>*If the risk is deemed too high need of disclosure limitation method* |
| 4. Disclosure limitation methods | Analysis of type of data involved, NSI policies and users need<br>⇓<br>*Identification of a disclosure limitation method* |
| 5. Implementation methods | Choice of software, parameters and thresholds for different methods |
| 6. assessment of disclosure risk and utility (ex post) | Ex post analysis of disclosure risk and information loss<br>⇓<br>*In case disclosure risk and/or utility loss is too high, return to step 5.* |

Table 3.1: Roadmap to releasing a microdata file

The idea is to identify for each stage of the process choices that have to be made, analyses that need to be done, problems that need to be addressed and methods to be selected. References to the relevant sections where technical topics are discussed in detail will help the beginners in following the process without getting lost in too technical aspects.

We now analyse in turn each of the six stages.

### 3.2.1 Need of confidentiality protection

The starting point deals with the need of confidentiality protection which is at the base of any release of microdata. If the microdata do not refer to legal entity or individual persons it can be released without confidentiality protection: an example is the amount of rain fall in a region. If microdata pertain only of public variables, in most cases they might be released: the legislation usually treats such data as excluded from statistical confidentiality. However, in general, data refer to individual or enterprises and contains confidential variables (health related data, income, turnover, expenses, etc.) and therefore need to be protected.

### 3.2.2 Characteristics and uses of microdata

Of course different levels of protection are needed for different type of users. This theme leads us to the second stage of the process i.e. the study of the key uses and characteristics of the data. Here the initial question is whether the microdata file we are going to release is intended for a general public (public use file) or whether it is created for research purpose (research use files). In the latter case the microdata will be released according to predefined procedures and legal binding (see also Section 6.5). The difference in user's type implies different user's needs, different disclosure scenarios, different types of analyses we expect to be performed with the released data, different statistics we may intend to preserve and different amount of protection we intend to apply. We now analyse all these issues in terms.

*Type and structure of data*
Analysis of user needs involves first a study of the survey information content. This should be done together with a survey expert that has a deeper knowledge of the data, phenomenon and possible types of analysis that can be performed on the data.

Typical questions that need to be addressed are:

*Which statistical units are involved in the survey?* Individuals, enterprises, households, etc. The type of units has a big influence on the risk assessment stage.

*Do data present a particular structure*? Hierarchical data: students inside schools, graduates inside universities, employees inside an enterprise, individual inside household etc. If this is the case care needs to be taken in checking both levels/types of

units involved. *E.g.*, do schools/universities/enterprises need to be protected besides students/graduates/employees?

*What type of sampling design has been used?* Are there strata (or units of earlier stages in a two- or multistage sampling design) which have been censured? Of course a complete enumeration of a strata (typical in business surveys) implies different and higher risks than a sample. Is two- or multistage sampling used with different types of units in the different stages?

An analysis of the questionnaire is useful to analyse the type of information present in the file: possible identifying variables (of which identifiers and quasi-identifiers), confidential variables and sensitive variables.

*Preliminary work on variables*
In this stage the setting of objectives from the viewpoint of the NSI and the user are defined. From the NSI side dissemination policies are clarified (e.g. level of dissemination of NACE, geography, etc. or coherence with published tables). From the user point of view a list of priorities in the structural variables of the survey, requests for minimum level of details for such variables and type of analysis to be performed (ratios, weighted totals, regressions, etc).

The characteristics of the phenomenon under study should also be considered as well as the dissemination policy of the Statistical Institute. This is particularly true for example in business data where some NACE classifications may never be released by their own, but always aggregated with others. Such a-priori aggregations generally depend on the economic structure of the country. It is not a sampling or dissemination problem, but rather a feature of the surveyed phenomenon. This will bring to aggregation of categories of some identifying variables deemed too detailed.

The output of this questionnaire analysis should be a preliminary list of variables to be removed and those to be released (because relevant to users need) together with some ideas of their level of details (depending on whether we are releasing a public use file or a research use file). Some examples to clarify these ideas. Variables that shouldn't be released comprise variables used as internal checks (*e.g.* some paradata), flags for imputation, variables that were not validated, variables deemed as not useful because containing too many missing values, information on the design stratum from which the unit comes from etc. Obviously also direct identifiers should not be released. The case studies A1 and A2 on microdata release provide examples of such stage.

Categories of identifying variables with too significant identifying power are commonly aggregated into a single category.

This is particularly true when releasing public use files as certain variables when too detailed could retain a level of "sensitivity". This may not be felt useful and/or appropriate for the general public. For example, in an household expenditure survey we might avoid releasing for the public use file very detailed information on the expenditure for housing (mortgage, rent) or detailed information on the age of the house or its number of rooms

(when this is very high) as these might be considered as giving too much information for particular outlying cases.

*Geography*
Another example is related to the level of geographical details that maybe different for a public use file or a research use file (especially if a data limitation technique is used). This happens because geographical information is a strongly identifying variable. Moreover, the geographical information collected from the respondent may be available in different variables for different purposes (place of birth, place of residence, place of work, place of study, commuting, etc.). All such geographical details need to be coherent/consistent throughout the file. To this end it may be convenient releasing relative information instead of absolute one: for example place of residence can be given at a certain detail (*e.g.* region) and then the other geographical information (place of work, study etc.) can be released with respect to this one. Examples of possible relative recodings (*e.g.* with respect to region of residence) are: region of work same as region of residence, different region but same macroregion, different macroregion.

*Coherence with published tables*
At this initial stage of the analysis information should be collected on what has already been published/what it is going to be released from the microdata set: dissemination plan, which type of tables and what classification/aggregation was used for the variables. This is to avoid different classifications in different release: the geographical breakdown, as well as classification of other variables in the survey (e.g. age, type of work etc.), should be coherent with the published marginals. For example, if a certain classification of the variable age is published in a table the microdata file should use a classification which has compatible break points so that to avoid gaining information by differencing. Release of date of birth is highly discouraged. Also, as far as possible, published totals should be preserved for transparency.

### 3.2.3 Disclosure risk (ex ante)

Moreover, in case of multiple release of the same survey (e.g. PUF and microdata for research) coherence should be maintained also between different released files in the sense that releasing different files at the same time shouldn't allow the gaining of more information than for one file alone (see, Trottini et al., 2006). The principles apply also to the release of longitudinal or panel microdata, where the differences between records pertaining to the same case in different waves will reflect 'events' that have occurred to that case, as well as the attributes of the individuals.

Once the characteristics and uses of the survey data are clear, it is time to start the real analysis of the disclosure risk in relation to files with originally collected data – *ex ante* assessment (in one of next subsections we will indicated also on a necessity of making *ex post* assessment of disclosure risk to verify effciency of used SDC methods). This implies

first a definition of possible situations at risk (disclosure scenarios) and second a proper definition of the 'risk' in order to quantify the phenomenon (risk assessment).

*Disclosure scenario*
A disclosure scenario is the definition of realistic assumptions about what an intruder might know about respondents and what information would be available to him to match against the microdata to be released and potentially make an identification and disclosure.

Again different types of releases may require different disclosure scenarios and different definitions of risk. For example the nosy neighbourhood scenario described in Section 3.3.2, possibly with knowledge of the presence of the respondent in the sample (implying that sample uniques are a relevant quantity of interest for risk definition), may be deemed adequate for a public use file. A different trust might be put in a researcher that needs to perform an analysis for research purposes. This implies, as a minimum step, a higher level of acceptable risk and a different scenario the spontaneous identification scenario.

*Spontaneous recognition*
Spontaneous recognition is possible when the researchers unintentionally recognize some units. For example, when releasing enterprise microdata, it is publicly known that the largest enterprises are generally included in the microdata file because of their significant impact on the studied phenomenon. Moreover, the largest enterprises are also the most identifiable ones as recognisable by all (the largest car producer factory, the national mail delivery enterprise, etc.). Consequently, a spontaneous identification or recognition might occur. A description of different scenarios is presented in Section 3.3.2; examples of spontaneous identification scenarios for MFR are reported in case studies A1 and A2.

*Definition of risk*
From the adopted scenario we can extract the list of identifying variables i.e. the variables that may allow the identification of a unit. These will be the basis for defining the risk of disclosure. Intuitively, a unit is at risk of identification when it cannot be confused with several other units. The difficulty is to express this simple concept using sound statistical methodology.

Different approaches are used if the identifying variables are categorical of continuous. In the former case at the basis of the definition is the concept of a 'key' (i.e. the combination of categories of the identifying variables): see Section 3.3.1 for a classification of different definitions. Whereas if continuous identifying variables are present in the file a possibility is to use the concept of density: see Ichim (2009) for a detailed analysis of definitions of risk in the case of continuous variables. Of course, the problem is even more complicated when we deal with a mixture of categorical and numerical key variables; for an example of this situation (quite common in enterprise microdata) see case study A1 (Community Innovation Survey). Another solution in this context can be the assessment of disclosure risk based on the (expected) number of units for which a value of the given continuous variable falls into the respectively defined (using established threshold of deviation) neighborhood of a given observation. This approach can be perceived as some variation

of $k$-anonymity in this case.

*Risk assessment*
Once a formal definition of risk has been chosen we need to measure/estimate it. There are several possibilities for categorical identifying variables (these are reported in various subsections of Section 3.3) and for a mixture of categorical and continuous identifying variable we have already mentioned Case study A1. The final step of the risk assessment is the definition of a threshold to define when a unit or a file presents an acceptable risk and when, on the contrary, it has to be considered at risk. This threshold depends of course on the type of measure adopted and details on how to choose a threshold are reported in the relevant subsequent sections.

Choice of scenarios and level of acceptable risk are extremely dependent on different cultural situations in different member states, different policies applied by different institutes, different approaches to statistical analysis, different perceived risk. To this end it must be stressed that different countries may have extremely different situation/phenomenon therefore different scenarios and risk methods are indeed necessary.

Currently there is no general agreement on which risk methodology is best although different methods give in general similar answers for the extreme cases. However, as already stated in Section 3.3.1, there is a strong need to further compare and understand differences between available methods. Pros and cons of each method are described in the relevant sections may be used as a guidelines for the most appropriate choice of the risk estimation in different situations. Further advice can be gained by studying of the examples and case studies.

### 3.2.4 SDC-methods

If the risk assessment stage shows that the disclosure risk is high then the application of statistical disclosure limitation methods is necessary to produce a microdata file for external users.

*Masking methods*
Microdata protection methods can generate a protected microdata set either by *masking original data*, *i.e.* generating a modified version of the original microdata set or by *generating synthetic data* that preserve some statistical properties of the original data. Synthetic data are still difficult to implement; a description can be found in Section 3.4.7. Masking methods are divided into two categories depending on their effect on the original data (Willenborg and De Waal, 2001): *perturbative* and *non perturbative masking* methods.

Perturbative methods either modify the identifying variables or modify the confidential variables before publication. In the former way, unique combinations of scores of identifying variables in the original dataset may disappear and new unique combinations may

appear in the perturbed dataset. In this way a user cannot be certain of an identification. Alternatively confidential variables can be modified; in this case even if an identification occurs, the wrong value is associated and disclosure of the original value is avoided (for an example of this case see case study A2). For a description of a variety of perturbative methods see sections 3.4.2, 3.4.4, 3.4.5 and 3.4.6.

Non-perturbative methods do not alter the values of the variables (either identifying or confidential); rather, they produce a reduction of detail in the original dataset. Examples of non-perturbative masking are presented in Section 3.4.3.

The choice between a data reduction and a data perturbation method strongly depends on the policy of an institute and on the type of data/survey to be released. While the policy of an institute is outside of this debate, technical reasons may suggest the use of perturbative methods for the protection of continuous variables (mainly business data). Analysis of information loss should always be part of the selection process. The usual difference between types of release remains valid and it is linked to the difference between users needs. Again the examples and the case studies A1 and A2 may help in clarifying different situations.

*User needs and types of protection*
From the needs of the users and the types of analyses that could be performed on the data one could gain information for the choice of the type of protection that could be applied to the microdata. Also users could express priorities in the need of maintaining some variables intact (e.g., for business usually NACE is the most important variable, then employees, and so on).

*Information loss*
For research purposes maybe we could be interested in maintaining the possibility of being able to reproduce the published tables. For a public use file maybe we could avoid, as much as possible, the use of local suppression as this may render data analysis difficult for non sophisticated users. In general, the implementation of perturbative methods should take into account what variables and relationships among them need to be kept from the user point of view. An assessment of information loss caused by the protection methods adopted is highly recommended. A brief description of information loss measures is reported in Section 3.5; examples of how to check in practice the amount of distortion or modification in the protected microdata is presented in case studies A1 and A2.

Finally, every time a data perturbation method is applied attention should be placed at relationships between different types of release (PUF, MFR, tables) so as to avoid as much as possible, different marginal totals from different sources.

An example of the application of this reasoning for the definition of a dissemination strategy can be found, for example, in Trottini et al. (2006).

### 3.2.5 Implementation

The next stage of the SDC process is the implementation of the whole procedure, choice of software, parameters and levels of acceptable risks.

Documentation is an essential part of any dissemination strategy both for auditing from external authorities and transparency towards users. The former may include description of legal and administrative steps for a risk management policy together with the technical solution applied. The latter is essential for a user to understand what has been changed or limited in the data because of confidentiality constraints. If a data perturbation method has been applied then, for transparency reasons, this should be clearly stated. Information on which statistics have been preserved and which have been modified and some order of magnitude of possible changes should be provided as far as possible. If a data reduction method has been applied with some local suppression then the distribution of such suppressions should be given for a series of different dimensions of interest (distribution by variables, by household size, household type, etc.) and any other statistics that are deemed relevant for the user. The released microdata should be obviously accompanied by all necessary metadata and information on methodologies used at various stage of the survey process (sampling, imputation, validation, etc.) together with information on magnitude of sampling errors, estimation domains etc.

### 3.2.6 Ex post assessment of disclosure risk and information loss

The last - but, of course, not least - stage of the procedure is the ex post assessment of disclosure risk and computation of the expected information loss due to SDC. The ex post risk assessment (usually made using the same measures as in the case of ex ante assessement, for comparability) allows for confirmation whether the used procedure eliminates or sufficiently reduces the threat of unit identification or not. If not, a modification of used methods (*e.g.* by changing some tools, modification of parameters, etc.) should be made. This means back to either the step "Disclosure limitation methods" or the step "Implementation".

An assessment of information loss caused by the applied protection methods is highly recommended. The knowledge of possible loss of information is key for data utility for possible users. If the information loss is too great then the used methods or their parameterization should be changed (coming back to step "Disclosure limitation methods"). One should remain that simultaneously the disclosure risk should be also as small as possible. Thus, these quantities should be harmonized. A detailed description of information loss measures is reported in Section 3.5; examples of how to check in practice the amount of distortion or modification in the protected microdata is presented in case studies A1 and A2.

Of course, the results of computation of discloure risk (both final and indirect, if applicable) and information loss should be saved in the documentation of the whole process. However,

whereas the values of measures of disclosure risk are confidential and known only by entitled staff of the data holder, the level of expected information loss should be made available to the user. It is a very important factor influencing the quality of the final analysis results obtained by him/her.

## 3.3 Risk assessment

### 3.3.1 Overview

Microdata has many analytical advantages over aggregated data, but also poses more serious disclosure issues because of the many variables that are disseminated in one file. For microdata, disclosure occurs when there is a possibility that an individual can be re-identified by an intruder using information contained in the file, and when on the basis of that, confidential information is obtained. Microdata are released only after taking out directly identifying variables, such as names, addresses, and identity numbers. However, other variables in the microdata can be used as indirect identifying variables. For individual microdata this are variables such as gender, age, occupation, place of residence, country of birth, family structure, etc. and for business microdata variables such as economic activity, number of employees, etc. These (indirect) identifying variables are mainly publicly available variables or variables that are present in public databases such as registers.

If the identifying variables are categorical then the compounding (cross-classification) of these variables defines a key. The disclosure risk is a function of such identifying variables/keys either in the sample alone or in both the sample and the population.

To assess the disclosure risk, we first need to make realistic assumptions about what an intruder might know about respondents and what information will be available to him to match against the microdata and potentially make an identification and disclosure. These assumptions are known as disclosure risk scenarios and more details and examples are provided in the next section of this handbook. Based on the disclosure risk scenario, the identifying variables are determined. The other variables in the file are confidential or sensitive variables and represent the data not to be disclosed. NSIs usually view all non-publicly available variables as confidential/sensitive variables regardless of their specific content, though there can be some variables, *e.g.* sexual identity, health conditions, income, that can be more sensitive.

In order to undertake a risk assessment of microdata, NSIs might rely on *ad-hoc* methods, experience and checklists based on assessing the detail and availability of identifying variables. There is a clear need for obtaining quantitative and objective disclosure risk measures for the risk of re-identification in the microdata. For microdata containing censuses or registers, the disclosure risk is known as we have all identifying variables available for the whole population. However, for microdata containing samples the population base

is unknown or partially known through marginal distributions. Therefore, probabilistic modelling or heuristics are used to estimate disclosure risk measures at population level, based on the information available in the sample. This section provides an overview of methods and tools that are available in order to estimate quantitative disclosure risk measures.

Intuitively, a unit is at risk if we are able to single it out from the rest. The idea at the base of the definition of risk is a way to measure rareness of a unit either in the sample or in the population.

When the identifying variables are categorical (as it is usually the case in social surveys) the risk is cast in terms of the cells of the contingency table built by cross-tabulating the identifying variables: the keys. Consequently all the records in the same cell have the same value of the risk.

*A classification of risk measures*
Several definitions of risk have been proposed in the literature; here we focus mainly on those for which tools are available to compute/estimate them easily. We can broadly classify disclosure risk measures into three types: risk measures based on keys in the sample, those based on keys in the population and that make use of statistical models or heuristics to estimate the quantities of interest and those based on the theory of record linkage. Whereas the first two classes are devoted to risk assessment for categorical identifying variables the third one may be used for categorical and continuous variables.

*Risk based on keys in the sample*
For the first class of risk measures a unit is at risk if its combination of scores on the identifying variables is below a given threshold. The threshold rule used within the software package $\mu$-ARGUS is an example of this class of risk measures.

*Risk based on keys in the population*
For the second type of approach we are concerned with the risk of a unit as determined by its combination of scores on the identifying variables within the population or its probability of re-identification. The idea then is that a unit is at risk if such quantity is above a given threshold. Because the frequency in the population is generally unknown, it may be estimated through a modelling process. Examples of this reasoning are the individual risk of disclosure based on the Negative Binomial distribution developed by Benedetti and Franconi (1998) and Franconi and Polettini (2004), which is outlined in Section 3.3.5, and the one based on the Poisson distribution and log-linear models developed by Skinner and Holmes (1998) and Elamir and Skinner (2004) which is described in Section 3.3.6 along with current research on other probabilistic methods. Another approach based on keys in the population is the Special Uniques Detection (SUDA) Algorithm developed by Elliot et al. (2002) that uses a heuristic method to estimate the risk; this is outlined in Section 3.3.7.

*Risk based on record linkage*
When identifying variables are continuous we cannot exploit the concept of rareness of the

keys and we transform such concept into rareness in the neighbourhood of the record. A way to measure rareness in the neighbourhood is through record linkage techniques. This third class of disclosure risk is covered in Section 3.3.8.

Section 3.3.2 provides an introduction to disclosure risk scenarios and Section 3.3.3 introduces concepts and notation used throughout this chapter. Sections to 3.3.8 describe different approaches to microdata risk assessment as specified above. However, as microdata risk assessment is a novelty in statistical research there isn't yet agreement on what method is the best, or at least best under given circumstances. In the following sections we comment on various approaches to risk measures and try to give advice on situations where they could or could not be applied. In any case, it has been recognised that research should be undertaken to evaluate these different approaches to microdata risk assessment, see for example Shlomo and Barton (2006).

The focus of these methods and this section of the handbook is for microdata samples from social surveys. For microdata samples from censuses or registers the disclosure risk is known. Business survey microdata are not typically released due to their disclosive nature (skewed distributions and very high sampling fractions).

In Section 3.7 we make some suggestions on practical implementation and in Section 3.8 we give examples of real data sets and ways in which risk assessment could be carried out.

### 3.3.2 Disclosure risk scenarios

The definition of a disclosure scenario is a first step towards the development of a strategy for producing a "safe" microdata file (MF). A scenario synthetically describes (i) which is the information potentially available to the intruder, and (ii) how the intruder would use such information to identify an individual i.e. the intruder's attack means and strategy. Often, defining more than one scenario might be convenient, because different sources of information might be alternatively or simultaneously available to the intruder. Moreover, re-identification risk can be assessed keeping into account different scenarios at the same time.

We refer to the information available to the intruder as an External Archive (EA), where information is provided at individual level, jointly with directly identifying data, such as name, surname, etc. The disclosure scenario is based on the assumption that the EA available to the intruder is an individual microdata archive. That is, for each individual directly identifying variables, and some other variables are available. Some of these further variables are assumed to be available also in the MF that we want to protect. The intruder's strategy of attack would be to use this overlapping information to match direct identifier to a record in the MF. The matching variables are then the *identifying variables*.

We consider two different types of re-identification, *spontaneous recognition* and *re-identification via record matching* (or *linkage*) according to the information we assume to be available to the intruder. In the first case we consider that the intruder might rely on personal knowledge about one or a few target individuals, and spontaneously recognize a surveyed individual (*Nosy Neighbour scenario*). In such a case the External Archive contains one (or a few) records relative to *detailed* personal information. In the second case, we assume that the intruder (who might be an MF user) has access to a *public register* and that he or she tries to match the information provided by this EA, with that provided by the MF, in order to identify surveyed units. In such a case, the intruder's chance of identifying a unit depends on the EA main characteristics, such as completeness, accuracy and data classification. Broadly speaking, we assume that the intruder has a lower chance of correctly identifying an individual when the information provided by the EA is not update, complete, accurate, or is classified according to standards different by those used in the statistical survey.

Moreover, as far as statistical disclosure control is concerned, experts are used to distinguish between social and economic microdata (without loss of generality we can consider respectively individuals and enterprises). In fact, the concept of disclosure risk is mainly based on the idea of rareness with respect to a set of identifying variables. For social survey microdata, because of the characteristics of the population under investigation and the nature of the data collected, identifying variables are mainly (or exclusively) categorical. For much of the information collected on enterprises however the identifying variables often take the form of quantitative variables with asymmetric distributions (Willenborg and de Waal, 2001). Disclosure scenarios are then described according to this statement.

The case study part of the Handbook contains examples of the Nosy Neighbour scenario and the EA scenario for social survey data. The issues involved with hierarchical and longitudinal data are also addressed. Finally, scenarios for business survey data are discussed.

In any case the definition of the scenario is essential as it defines the hypothesis underneath the risk estimation and the subsequent protection of the data.

### 3.3.3 Concepts and notation

For microdata, disclosure risk measures quantify the risk of re-identification. Individual per record disclosure risk measures are useful for identifying high-risk records and targeting the SDC methods. These individual risk measures can be aggregated to obtain global file level disclosure risk measures. These global risk measures are particularly useful to NSIs for their decision making process on whether the microdata is safe to be released and allows comparisons across different files.

*Microdata disclosure*
Disclosure in a microdata context means a correct record re-identification operation that is achieved by an intruder when comparing a target individual in a sample with an available

list of units (external file) that contains individual identifiers such as name and address plus a set of identifying variables. Re-identification occurs when the unit in the released file and a unit in the external file belong to the same individual in the population. The underlying hypothesis is that the intruder will always try to match a unit in the sample $s$ to be released and a unit in the external file using the identifying variables only. In addition, it is likely that the intruder will be interested in identifying those sample units that are unique on the identifying variables. A re-identification occurs when, based on a comparison of scores on the identifying variables, a unit $i^*$ in the external file is selected as matching to a unit $i$ in the sample and this link is correct and therefore confidential information about the individual is disclosed using the direct identifiers.

To define the disclosure scenario, the following assumptions are made. Most of them are conservative and contribute to the definition of a worst case scenario:

1. a sample $s$ from a population $\mathcal{P}$ is to be released, and sampling design weights are available;
2. the external file available to the intruder covers the whole population $\mathcal{P}$; consequently for each $i \in s$ the matching unit $i^*$ does always exist in $\mathcal{P}$;
3. the external file available to the intruder contains the individual direct identifiers and a set of categorical identifying variables that are also present in the sample;
4. the intruder tries to match a unit $i$ in the sample with a unit $i^*$ in the population register by comparing the values of the identifying variables in the two files;
5. the intruder has no extra information other than that contained in the external file;
6. a re-identification occurs when a link between a sample unit $i$ and a population unit $i^*$ is established and $i^*$ is actually the individual of the population from which the sampled unit $i$ was derived; e.g. the match has to be a correct match before an identification takes place.

Moreover we add the following assumptions:

7. the intruder tries to match all the records in the sample with a record in the external file;
8. the identifying variables agree on correct matches, that is no errors, missing values or time-changes occur in recording the identifying variables in the two microdata file.

*Notation*
The following notation is introduced here and used throughout the chapter when describing different methods for estimating the disclosure risk of microdata.

Suppose the key has $K$ cells and each cell $k = 1, \ldots, K$ is the cross-product of the categories of the identifying variables. In general, we will be looking at a contingency table spanned by the identifying variables in the microdata and not a single vector. The contingency table contains the sample counts and is typically very large and very sparse. Let the population size in cell $k$ of the key be $F_k$ and the sample size $f_k$. Also:

$$\sum_{k=1}^{K} F_k = N, \quad \sum_{k=1}^{K} f_k = n.$$

Formally the sample and population sizes in the models introduced in Section 3.3.5 and 3.3.6 are random and their expectations are denoted by $n$ and $N$ respectively. In practice, the sample and population size are usually replaced by their natural estimators; the actual sample and population sizes, assumed to be known.

Observing the values of the key on individual $i \in s$ will classify such individual into one cell. We denote by $k(i)$ the index of the cell into which individual $i \in s$ is classified based on the values of the key.

According to the concept of re-identification disclosure given above, we define the (base) individual risk of disclosure of unit $i$ in the sample as its probability of re-identification under the worst case scenario. Therefore the risk $r_i$ that we get is certainly not smaller than the actual risk, the individual risk is a conservative estimate of the actual risk:

$$r_i = \mathbb{P}\left(i \text{ correctly linked with } i^* \mid s, \mathcal{P}, \text{ worst case scenario }\right) \tag{3.1}$$

All of the methods based on keys in the population described in this chapter aim to estimate this individual per-record disclosure risk measure that can be formulated as $1/F_k$. The population frequencies $F_k$ are unknown parameters and therefore need to be estimated from the sample. A global file-level disclosure risk measure can be calculated by aggregating the individual disclosure risk measures over the sample:

$$\tau_1 = \sum_k \frac{1}{F_k}$$

An alternative global risk measure can be calculated by aggregating the individual disclosure risk measures over the sample uniques of the cross-classified identifying variables. Since the uniques in the population $F_k = 1$, are the dominant factor in the disclosure risk measure, we focus our attention on sample uniques $f_k = 1$:

$$\tau_2 = \sum_k I(f_k = 1)\frac{1}{F_k}$$

where $I$ represents an indicator function obtaining the value 1 if $f_k = 1$ or 0 if not.

Both of these global risk measures can also be presented as rates by dividing by $n$, the sample size or the number of uniques.

We assume that the $f_k$ are observed but the $F_k$ are not observed.

### 3.3.4 ARGUS threshold rule

The ARGUS threshold rule is based on easily applicable rules and views of safety/unsafety of microdata that is used at Statistics Netherlands. The implementation of these rules was the main reason to start the development of the software package $\mu$-ARGUS.

In a disclosure scenario, keys a combination of identifying variables, are supposed to be used by an intruder to re-identify a respondent. Re-identification of a respondent can occur when this respondent is rare in the population with respect to a certain key value, i.e. a combination of values of identifying variables. Hence, rarity of respondents in the population with respect to certain key values should be avoided. When a respondent appears to be rare in the population with respect to a key value, then disclosure control measures should be taken to protect this respondent against re-identification.

Following the Nosy Neighbour scenario, the aim of the $\mu$-ARGUS threshold rule is to avoid the occurrence of combinations of scores that are rare in the population and not only avoiding population-uniques. To define what is meant by rare the data protector has to choose a threshold value for each key. If a key occurs more often than this threshold the key is considered safe, otherwise the key must be protected because of the risk of re-identification.

The level of the threshold and the number and size of the keys to be inspected depend of course on the level of protection you want to achieve. Public use files require much more protection than microdata files under contract that are only available to researchers under a contract. How this rule is used in practice is given in the example of Section 3.7.

If a key is considered unsafe according to this rule, protection is required. Therefore often global recoding and local suppression are applied. These techniques are described in the sections 3.4.3.2 and 3.4.3.4.

### 3.3.5 ARGUS individual risk methodology

If a distinction between units rare in the sample from a unit rare in the population wants to be made then an inferential step may be followed. In the initial proposal by Benedetti and Franconi (1998), further developed in Franconi and Polettini (2004) and implemented in $\mu$-ARGUS, the uncertainty on $F_k$ is accounted for in a Bayesian fashion by introducing the distribution of the population frequencies given the sample frequencies. The individual risk of disclosure is then measured as the (posterior) mean of $\frac{1}{F_k}$ with respect to the distribution of $F_k|f_k$:

$$r_i = \mathbb{E}\left(\frac{1}{F_k} \mid f_k\right) = \sum_{h \geq f_k} \frac{1}{h} \mathbb{P}\left(F_k = h \mid f_k\right). \tag{3.2}$$

where the posterior distribution of $F_k|f_k$ is negative binomial with success probability $p_k$ and number of successes $f_k$. As the risk is a function of $f_k$ and $p_k$ its estimate can be obtained by estimating $p_k$. Benedetti and Franconi (1998) propose to use

$$\hat{p}_k = \frac{f_k}{\sum_{i:k(i)=k} w_i} \tag{3.3}$$

where $\sum_{i:k(i)=k} w_i$ is an estimate of $F_k$ based on the sampling design weights $w_i$, possibly calibrated (Deville and Särndal, 1992).

*When is it possible to apply the individual risk estimation*
The procedure relies on the assumption that the available data are a sample from a larger population. *If the sampling weights are not available, or if data represent the whole population, the strategy used to estimate the individual risk is not meaningful.*

In the $\mu$-ARGUS manual (Hundepool *et al.*, 2004) a fully detailed description of the approach is reported. This brief note is based on Polettini (2004).

*Assessing the risk for the whole file*
The individual risk provides a measure of risk *at the individual level.* A *global* measure of disclosure risk for the whole file can be expressed in terms of the expected number of re-identifications in the file. The *expected number of re-identifications* is a measure of disclosure that depends on the number of records. For this reason, $\mu$-ARGUS evaluates also the *re-identification rate* that is independent of $n$:

$$\xi = \frac{1}{n}\sum_{k=1}^{K} f_k r_k \quad .$$

$\xi$ provides a measure of *global risk, i.e.* a measure of disclosure risk for the whole file, which does not depend on the sample size and can be used to assess the risk of the file or to compare different types of release; for the mathematical details see Polettini (2004).

The *percentage of expected re-identifications*, *i.e.* the value $\psi = 100 \cdot \xi\%$ provides an equivalent measure of global risk.

*Application of local suppression within the individual risk methodology*
After the risk has been estimated, protection takes place. One option in protection is the application of *local suppression* (see Section 3.4.3.4).

In $\mu$-ARGUS the technique of local suppression, when combined with the individual risk, is applied only to unsafe cells or combinations. Therefore, the user must input a *threshold* in terms of risk, *e.g.* probability of re-identification, to classify these as either safe or unsafe. Local suppression is applied to the unsafe individuals, so as to lower their probability of being re-identified under the given threshold.

In order to select the risk threshold, that represents a level of *acceptable risk*, *i.e.* a risk value under which an individual can be considered safe, the *re-identification rate* can be used. A *release* will be considered *safe* when the expected rate of correct re-identifications is below a level the NSI considers acceptable. As the re-identification rate is cast in terms of the individual risk, a threshold on the re-identification rate can be transformed into a threshold on the individual risk (see below). Under this approach, individuals are at risk because their probability of re-identification contributes a large proportion of expected re-identifications in the file.

In order to reduce the number of local suppressions, the procedure of releasing a safe file considers preliminary steps of protection using techniques such as *global recoding* (see Section 3.4.3.2). Recoding of selected variables will indeed lower the individual risk and therefore the re-identification rate of the file.

⚠ Expert level

*Threshold setting using the re-identification rate*
Consider the re-identification rate $\xi$: a key $k$ contributes to $\xi$ an amount $r_k f_k$ of expected re-identifications. Since units belonging to the same key $k$ have the same individual risk, keys can be arranged in increasing order of risk $r_k$. Let the subscript $(k)$ denotes the $k$-th element in this ordering. A threshold $r^*$ on the individual risk can be set. Consequently, unsafe cells are those for which $r_k \geq r^*$ that can be indexed by $(k) = k^* + 1, \ldots, K$. The key $k^*$ is in a one-to-one correspondence to $r^*$. This allows setting an upper bound $\xi^*$ on the re-identification rate of the released file (after data protection) substituting $r_k f_k$ with $r^* f_{(k)}$ for each $(k)$. For the mathematical details see Polettini (2004) and the Argus manual (Hundepool *et al.*, 2004).
The approach pursued so far can be reversed. Therefore, selecting a *threshold $\tau$ on the re-identification rate $\xi$* determines a key index $k^*$ which corresponds to a value for $r^*$. Using $r^*$ as a threshold for the individual risk keeps the re-identification rate $\xi$ of the released file below $\tau$. The search of such a $k^*$ is performed by a simple iterative algorithm.
*Releasing hierarchical files*
A relevant characteristic of social microdata is its inherent hierarchical structure, which allows us to recognise groups of individuals in the file, the most typical case being the *household*. When defining the re-identification risk, it is important to take into account this dependence among units: indeed re-identification of an individual in the group may affect the probability of disclosure of all its members. So far, implementation of a hierarchical risk has been performed only with reference to households, *i.e.* a *household risk*.
Allowing for dependence in estimating the risk enables us to attain a higher level of safety than when merely considering the case of independence.
*The household risk*
The household risk makes use of the same framework defined for the individual risk.

In particular, the concept of re-identification holds with the additional assumption that *the intruder attempts a confidentiality breach by re-identification of individuals in households.*
The *household risk* is defined as the probability that *at least* one individual in the household is re-identified. For a given household $g$ of size $|g|$, whose members are labelled $i_1, \dots, i_{|g|}$, the household risk is:

$$r^h(g) = \mathbb{P}\left(i_1 \cup i_2 \cup \dots \cup i_{|g|} \text{ re-identified }\right)$$

and is the same for all the individuals in household $g$ and equals $r_g^h$.
*Threshold setting for the household risk*
Since all the individuals in a given household have the same household risk, the expected number of re-identified records in household $g$ equals $|g|r_g^h$. The re-identification rate in a hierarchical file can be then defined as $\xi^h = \frac{1}{n}\sum_{g=1}^{G}|g|r_g^h$, where $G$ is the total number of households in the file. The re-identification rate can be used to define a threshold $r^{h^*}$ on the household risk $r^h$, much in the same way as for the individual risk. For the mathematical details see Polettini (2004) and the Argus manual (Hundepool *et al.*, 2004).
Note that the household risk $r_g^h$ of household $g$ is computed by the individual risks of its household members. For a given household, it might happen that a household is unsafe ($r_g^h$ exceeds the threshold) because just one of its members, $i$, say, has a high value $r_i$ of the individual risk. To protect the households, the followed approach is therefore to protect individuals in households, first protecting those individuals who contribute most to the household risk. For this reason, inside *unsafe households*, detection of *unsafe individuals* is needed. In other words, the threshold on the household risk $r^h$ has to be transformed into a threshold on the individual risk $r_i$. To this aim, it can be noticed that the household risk is bounded by the sum of the individual risks of the members of the household: $r_g^h \leq \sum_{j=1}^{|g|} r_{i_j}$.
Consider to apply a threshold $r^{h^*}$ on the household risk. In order for household $g$ to be classified safe (*i.e.* $r_g^h < r^{h^*}$) it is *sufficient* that all of its components have individual risk less than $\delta_g = r^{h^*}/|g|$.
This is clearly an approach possibly leading to overprotection, as we check whether a *bound* on the household risk is below a given threshold.
It is important to remark that the threshold $\delta_g$ just defined depends on the size of the household to which individual $i$ belongs. This implies that for two individuals that are classified in the same key $k$ (and therefore have the same individual risk $r_k$), but belong to different households with different sizes, it might happen that one is classified safe, while the other unsafe (unless the household size is included in the set of identifying variables).
In practice, denoting by $g(i)$ the household to which record $i$ belongs, the approach

pursued so far consists in turning a threshold $r^{h^*}$ on the household risk into a *vector of thresholds* on the *individual risks* $r_i = 1, \ldots, n$:

$$\delta_g = \delta_{g(i)} = \frac{r^{h^*}}{|g(i)|} \quad .$$

*Individuals* are finally set to unsafe whenever $r_i \geq \delta_{g(i)}$; local suppression is then applied to those records, if requested. Suppression of these records ensures that after protection the household risk is below the threshold $\delta_g$.

*Choice of identifying variables in hierarchical files*

For household data it is important to include in the identifying variables that are used to estimate the household risks also the available information on the household, such as the number of components or the household type.

Suppose one computes the risk using the household size as the only identifying variable in a household data file, and that such file contains households whose risk is above a fixed threshold. Since information on the number of components in the household cannot be removed from a file with household structure, these records cannot be safely released, and no suppression can make them safe. This permits to check for presence of very peculiar households (usually, the very large ones) that can be easily recognised in the population just by their size and whose main characteristic, namely their size, can be immediately computed from the file. For a discussion on this issue see Polettini (2004).

### 3.3.6 The Poisson model with log-linear modelling

As defined in Skinner and Elamir (2004), assuming that the $F_k$ are independently Poisson distributed with means $\{\lambda_k\}$ and assuming a Bernoulli sampling scheme with equal selection probably $\pi$, then $f_k$ and $F_k - f_k$ are independently Poisson distributed as: $f_k \mid \lambda_k \sim \text{Pois}(\pi\lambda_k)$ and $F_k - f_k \mid \lambda_k \sim \text{Pois}((1-\pi)\lambda_k)$ . The individual risk measure for a sample unique is defined as $r_k = \mathbb{E}_{\lambda_k}\left(\frac{1}{F_k} \mid f_k = 1\right)$ which is equal to:

$$r_k = \frac{1}{\lambda_k(1-\pi)}\left[1 - e^{-\lambda_k(1-\pi)}\right]$$

In this approach the parameters $\{\lambda_k\}$ are estimated by taking into account the structure and dependencies in the data through log-linear modelling. Assuming that the sample frequencies $f_k$ are independently Poisson distributed with a mean of $u_k = \pi\lambda_k$, a log-linear model for the $u_k$ can be expressed as: $\log(u_k) = x_k^{'}\beta$ where $x_k$ is a design vector denoting the main effects and interactions of the model for the key variables. Using standard procedures, such as iterative proportional fitting, we obtain the Poisson maximum-likelihood estimates for the vector $\beta$ and calculate the fitted values: $\widehat{u}_k = \exp(x_k^{'}\widehat{\beta})$. The estimate for

$\hat{\lambda}_k$ is equal to $\frac{\hat{u}_k}{\pi}$ which is substituted for $\lambda_k$ in the above formula for $r_k$. The individual disclosure risk measures can be aggregated to obtain a global (file-level) measure:

$$\hat{\tau}_2 = \sum_{k \in \mathrm{SU}} \hat{r_k} = \sum_{k \in \mathrm{SU}} \frac{1}{\hat{\lambda}_k(1-\pi)}[1 - e^{-\hat{\lambda}_k(1-\pi)}]$$

where SU is the set of all sample uniques.

More details on this method are available from Skinner and Shlomo (2005, 2006) and Shlomo and Barton (2006).

⚠ Expert level

Skinner and Shlomo (2005, 2006) have developed goodness-of-fit criteria for selecting the most robust log-linear model that will provide accurate estimates for the global disclosure risk measure detailed above. The method begins with a log-linear model where a high test statistic indicates under-fitting (*i.e.*, the disclosure risk measures will be over-estimated). Then a forward search algorithm is employed by gradually adding in higher order interaction terms into the model until the test statistic approaches the level (based on a Normal distribution approximation) where the fit of the log-linear model is accepted.
This method is still under development. At present there is a need to develop clear and user-friendly software to implement the method. However, the Office for National Statistics in the UK has used it to inform microdata release decisions. The method is based on theoretical well-defined disclosure risk measures and goodness of fit criteria which ensure the fit of the log-linear model and the accuracy of the disclosure risk measures. It requires a model search algorithm which takes some computer time and requires intervention.
New methods for probabilistic risk assessment are under development based on a generalized Negative Binomial smoothing model for sample disclosure risk estimation which subsumes both the model used in $\mu$-ARGUS and the Poisson log-linear model above. The method is useful for key variables that are ordinal where local neighbourhoods can be defined for inference on cell $k$. The Bayesian assumption of $\lambda_k \sim \mathrm{Gamma}(\alpha_k, \beta_k)$ is added independently to the Poisson model above which then transforms the marginal distribution to the generalized Negative Binomial Distribution:

$$f_k \sim \mathrm{NB}(\alpha_k, p_k = \frac{1}{1 + \mathrm{N}\pi_k\beta_k})$$

and

$$F_k | f_k \sim \mathrm{NB}(\alpha_k + f_k, \rho_k = \frac{1 + \mathrm{N}\pi_k\beta_k}{1 + \mathrm{N}\beta_k})$$

where $\pi_k$ is the sampling fraction. In each local neighbourhood of cell $k$ a smooth-

ing polynomial regression model is carried out to estimate $\alpha_k$ and $\beta_k$, and disclosure risk measures are estimated based on the Negative Binomial Distribution, $\hat{\tau}_2 = \sum_{k \in SU} \hat{r_k} = \sum_{k \in SU} \frac{\hat{\rho}_k (1-\hat{\rho}_k)^{\hat{\alpha}_k}}{\hat{\alpha}_k (1-\hat{\rho}_k)}$ , see: Rinott and Shlomo (2005, 2006).

### 3.3.7 SUDA

The Special Uniques Detection Algorithm (SUDA) (Elliot et.al., 2005) is a software system (windows application available as freeware under restricted licence) that provides disclosure risk broken down by record, variable, variable value and by interactions of those. It is based on the concept of a "special unique". A special unique is a record that is a sample unique on a set of variables and that is also unique on a subset of those variables. Empirical work has shown that special uniques are more likely to be population unique than random uniques. Special uniques can be classified according to the size and number of the smallest subset of key variables that defines the record as unique, known as minimal sample uniques (MSU). In the algorithm, all MSUs are found for each record on all possible subsets of the key variables where the maximum size of the subsets $m$ is specified by the user.

⚠ Expert level

SUDA grades and orders records within a microdata file according to the level of risk. The method assigns a per record matching probability to a sample unique based on the number and size of minimal uniques. The DIS Measure (Skinner and Elliot, 2000) is the conditional probability of a correct match given a unique match:

$$p(cm \mid um) = \frac{\sum_{k=1}^{K} I\left(f_k = 1\right)}{\sum_{k=1}^{K} F_k I\left(f_k = 1\right)}$$

and is estimated by a simple sample-based measure which is approximately unbiased without modelling assumptions. Elliot (2005) describes a heuristic which combines the DIS measure with scores resulting from the algorithm (*i.e.*, SUDA scores). This method known as DIS-SUDA produces estimates of intruder confidence in a match against a given record being correct. This is closely related to the probability that the match is correct and is heuristically linked to the estimate of

$$\tau_2 = \sum_k I(f_k = 1) \frac{1}{F_k}$$

The advantage of this method is that it relates to a practical model of data intrusion, and it is possible to compare different values directly. The disadvantages are that it is sensitive to level of the max MSU parameter and is calculated in a heuristic manner.

In addition it is difficult to compare disclosure risk across different files. However, the method has been extensively tested and was used successfully for the detection of high-risk records in the UK Sample of Anonymized Records (SAR) drawn from the 2001 Census (Merrett, 2004). The assessment showed that the DIS-SUDA measure calculated from the algorithm provided a good estimate for the individual disclosure risk measure, especially for the case when the number of key variables, $m = 6$. The algorithm also identifies the variables and value of variables that are contributing most to the disclosure risk of the record.
A new algorithm, SUDA2 has been developed, Elliot et al (2005), that improves SUDA using several methods. The development provides a much faster tool that can handle larger datasets.

### 3.3.8 Record Linkage

Roughly speaking, record linkage consists of linking each record $a$ in file $A$ (protected file) to a record $b$ in file $B$ (original file). The pair $(a, b)$ is a match if $b$ turns out to be the original record corresponding to $a$.

To apply this method to measure the risk of identity disclosure, it is assumed that an intruder has got an external dataset sharing some (key or outcome) variables with the released protected dataset and containing additionally some identifier variables (*e.g.* passport number, full name, etc.). The intruder is assumed to try to link the protected dataset with the external dataset using the shared variables. The number of matches gives an estimation of the number of protected records whose respondent can be re-identified by the intruder. Accordingly, disclosure risk is defined as the proportion of matches among the total number of records in $A$.

The main types of record linkage used to measure identity disclosure in SDC are discussed below. An illustrative example can be found on the CASC-website as one of the case-studies linked to this handbook (see https://research.cbs.nl/casc/handbook.htm).

#### 3.3.8.1 Distance-based record linkage

⚠ Expert level

Distance-based record linkage consists of linking each record $a$ in file $A$ to its nearest record $b$ in file $B$. Therefore, this method requires a definition of a distance function for expressing *nearness* between records. This record-level distance can be constructed from distance functions defined at the level of variables. Construction of record-level distances requires standardizing variables to avoid scaling problems and assigning each variable a weight on the record-level distance.

Distance-based record linkage was first proposed in Pagliuca and Seri (1999) to assess the disclosure risk after microaggregation, see Section 3.4.2.3. Those authors used the Euclidean distance and equal weights for all variables. (Domingo-Ferrer and Torra, 2001) later used distance-based record linkage for evaluating other masking methods as well; in their empirical work, distance-based record linkage outperforms probabilistic record linkage (described below). Recently, (Torra and Miyamoto, 2004) have shown that method-specific distance functions might be defined to increase the proportion of matches for particular SDC methods.
The record linkage algorithm introduced in (Bacher, Brand and Bender, 2002) is similar in spirit to distance-based record linkage. This is so because it is based on cluster analysis and, therefore, links records that are near to each other.
The main advantages of using distances for record linkage are simplicity for the implementer and intuitiveness for the user. Another strong point is that subjective information (about individuals or variables) can be included in the re-identification process by properly modifying distances. In fact, the next version of the $\mu$-ARGUS microdata protection package (Hundepool *et al.*, 2005) will incorporate distance-based record linkage as a disclosure risk assessment method.
The main difficulty of distance-based record linkage consists of coming up with appropriate distances for the variables under consideration. For one thing, the weight of each variable must be decided and this decision is often not obvious. Choosing a suitable distance is also especially thorny in the cases of categorical variables and of masking methods such as local recoding where the masked file contains new labels with respect to the original dataset.

#### 3.3.8.2 Probabilistic record linkage

⚠ Expert level

Like distance-based record linkage, probabilistic record linkage aims at linking pairs of records $(a, b)$ in datasets $A$ and $B$, respectively. For each pair, an index is computed. Then, two thresholds $LT$ and $NLT$ in the index range are used to label the pair as linked, clerical or non-linked pair: if the index is above $LT$, the pair is linked; if it is below $NLT$, the pair is non-linked; a clerical pair is one that cannot be automatically classified as linked or non-linked and requires human inspection. When independence between variables is assumed, the index can be computed from the following conditional probabilities for each variable: the probability $\mathbb{P}(1 \mid M)$ of coincidence between the values of the variable in two records $a$ and $b$ given that these records are a real match, and the probability $\mathbb{P}(0 \mid U)$of non-coincidence between the values of the variable given that $a$ and $b$ are a real unmatch.
Like in the previous section, disclosure risk is defined as the number of matches (linked pairs that are correctly linked) over the number of records in file $A$.

To use probabilistic record linkage in an effective way, we need to set the thresholds $LT$ and $NLT$ and estimate the conditional probabilities $\mathbb{P}(1 \mid M)$ and $\mathbb{P}(0 \mid U)$ used in the computation of the indices. In plain words, thresholds are computed from: (i) the probability $\mathbb{P}(\text{LP} \mid U)$ of linking a pair that is an unmatched pair (a *false positive* or *false linkage*) and (ii) the probability $\mathbb{P}(\text{NP} \mid M)$of not linking a pair that is a match (a *false negative* or *false unlinkage*). Conditional probabilities $\mathbb{P}(1 \mid M)$ and $\mathbb{P}(0 \mid U)$ are usually estimated using the EM algorithm (Dempster, Laird and Rubin 1977).
Original descriptions of this kind of record linkage can be found in Fellegi and Sunter (1969) and Jaro (1989). Torra and Domingo-Ferrer (2003) describe the method in detail (with examples) and Winkler (1993) presents a review of the state of the art on probabilistic record linkage. In particular, this latter paper includes a discussion concerning non-independent variables. A (hierarchical) graphical model has recently been proposed (Ravikumar and Cohen, 2004) that compares favourably with previous approaches.
Probabilistic record linkage methods are less simple than distance-based ones. However, they do not require rescaling or weighting of variables. The user only needs to provide two probabilities as input: the maximum acceptable probability $\mathbb{P}(\text{LP} \mid U)$ of false positive and the maximum acceptable probability $\mathbb{P}(\text{NP} \mid M)$ of false negative.

#### 3.3.8.3 Other record linkage methods

⚠ Expert level

Recently, the use of other record linkage methods has also been considered for disclosure risk assessment. While in the previous record linkage methods it is assumed that the two files to be linked share a set of variables, other methods have been developed where this constraint is relaxed. Under appropriate conditions, (Torra, 2004) shows that re-identification is still possible when files do not share any variables. Domingo-Ferrer and Torra (2003) propose the use of such methods for disclosure risk assessment.

### 3.3.9 References

Bacher J., Brand R., and Bender S. (2002), *Re-identifying register data by survey data using cluster analysis: an empirical study.* International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10(5):589–607, 2002.

Benedetti, R. and Franconi, L. (1998). *Statistical and technological solutions for controlled data dissemination*, Pre-proceedings of New Techniques and Technologies for Statistics, 1, 225-232.

Coppola, L. and Seri, G. (2005). *Confidentiality aspects of household panel survey: the case study of Italian sample from EU-SILC.* Monographs of official statistics – Proceedings of the Work session on statistical data confidentiality – Geneve 9-11 November 2005, 175-180.

Cox, L.H. (1995). *Protecting confidentiality in business surveys.* Business Survey Methods, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. e Kott, P.S. (Eds.), New-York: Wiley, 443-476.

Dempster A. P., Laird N. M., and Rubin D. B. (1977), *Maximum likelihood from incomplete data via the em algorithm.* Journal of the Royal Statistical Society, 39:1–38, 1977.

Deville, J.C. and Särndal, C.E. (1992). *Calibration estimators in survey sampling*, Journal of the American Statistical Association 87, 367–382.

Domingo-Ferrer J., and Torra, V. (2001), *A quantitative comparison of disclosure control methods for microdata.* In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pages 111–134, Amsterdam, 2001. North-Holland. http://vneumann.etse.urv.es/publications/bcpi.

Domingo-Ferrer, J., and Torra, V. (2003), *Disclosure risk assessment in statistical microdata protection via advanced record linkage.* Statistics and Computing, 13:343–354.

Elamir, E., Skinner, C. (2004) *Record-level Measures of Disclosure Risk for Survey Microdata*, Journal of Official Statistics (forthcoming). See also: Southampton Statistical Sciences Research Institute, University of Southampton, methodology working paper: http://eprints.soton.ac.uk/8175/01/s3ri-workingpaper-m04-02.pdf

Elliot, M. J., (2000). *DIS: A new approach to the Measurement of Statistical Disclosure Risk.* International Journal of Risk Management 2(4), pp 39-48.

Elliot, M. J., Manning, A. M.& Ford, R. W. (2002**).** '*A Computational Algorithm for Handling the Special Uniques Problem*'. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 5(10), pp 493-509.

Elliot, M. J., Manning, A., Mayes, K., Gurd, J. & Bane, M. (2005). '*SUDA: A Program for Detecting Special Uniques'.* Proceedings of the UNECE/Eurostat work session on statistical data confidentiality, Geneva, November 2005

Elliot, M. J., Skinner, C. J., and Dale, A. (1998). *'Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk'.* Research in Official Statistics 1(2), pp 53-67.

Fellegi, I. P., and Sunter, A.B. (1969), *A theory for record linkage.* Journal of the American Statistical Association, 64(328):1183–1210.

Franconi, L. and Polettini, S. (2004). *Individual risk estimation in μ-ARGUS: a review.* In: Domingo-Ferrer, J. (Ed.), Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer, 262-272

Franconi, L. and Seri, G. (2000). *Microdata Protection at the Italian National Statistical Insititute (Istat): A User Perspective. Of Significance* Journal of the Association of Public Data Users – Volume 2 Number 1 2000, page. 57-64.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., and Giessing, S. (2005), *μ-ARGUS version 4.0 Software and User's Manual*. Statistics Netherlands, Voorburg NL, may 2005. https://research.cbs.nl/casc.

Jackson, P., Longhurst, J. (2005), *Providing access to data and making microdata safe, experiences of the ONS*, proceedings of the UNECE/Eurostat work session on statistical data confidentiality, Geneva, November 2005

Jaro, M.A. (1989), *Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida.* Journal of the American Statistical Association, 84(406):414–420.

Pagliuca, D. and Seri, G. (1999), *Some results of individual ranking method on the system of enterprise accounts annual survey*, Esprit SDC Project, Deliverable MI-3/D2.

Polettini, S. and Seri, G (2004). *Revision of "Guidelines for the protection of social microdata using the individual risk methodology".* Deliverable 1.2-D3, available at CASC web site.

Ravikumar, P., and Cohen, W.W. (2004),. *A hierarchical graphical model for record linkage.* In UAI 2004, USA, 2004. Association for Uncertainty in Artificial Intelligence.

Rinott, Y. ,Shlomo, N (2006) *A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation* ,. PSD'2006 Privacy in Statistical Databases, Springer LNCS proceedings, to appear.

Rinott, Y., Shlomo, N. (forthcoming) *A Smoothing Model for Sample Disclosure Risk Estimation*, Volume in memory of Yehuda Vardi in the IMS Lecture Notes Monograph Series.

Shlomo, N. (2006), *Review of statistical disclosure control methods for census frequency tables*, ONS Survey Methodology Bulletin.

Shlomo, N., Barton, J. (2006) *Comparison of Methods for Estimating Disclosure Risk Measures for Microdata at the UK Office for National Statistics*, PSD'2006 Privacy in Statistical Databases Conference, CD Proceedings, to appear

Skinner, C., Shlomo, N. (2005), *Assessing disclosure risk in microdata using record-level measures*, proceedings of the UNECE/Eurostat work session on statistical data confidentiality, Geneva, November 2005

Skinner, C.J., Shlomo, N. (2006) *Assessing Identification Risk in Survey Microdata Using Log-linear Models*, see: http://eprints.soton.ac.uk/41842/01/s3ri-workingpaper-m06-14.pdf

Skinner, C., Holmes, D. (1998), *Estimating the re-identification risk per record in microdata*, JOS, Vol.14.

Torra, V. (2004), *Owa operators in data modeling and re-identification.* IEEE Trans. on Fuzzy Systems, vol. 12, no. 5, pp. 652-660.

Torra V., and Domingo-Ferrer J. (2003). *Record linkage methods for multidatabase data mining.* In V. Torra, editor, Information Fusion in Data Mining, pages 101–132, Germany, Springer.

Torra, V., and Miyamoto, S. (2004),. *Evaluating fuzzy clustering algorithms for microdata protection.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of LNCS, pages 175–186, Berlin Heidelberg. Springer.

Willenborg, L. and De Waal, T. (1996). *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics, 111, New-York: Springer Verlag.

Willenborg, L. and De Waal, T. (2001). *Elements of statistical disclosure control.* Lecture Notes in Statistics, 115, New York: Springer-Verlag.

Winkler, W. E. (1993),. *Matching and record linkage.* Technical Report RR93/08, Statistical Research Division, U. S. Bureau of the Census (USA), 1993.

## 3.4 Microdata protection methods

### 3.4.1 Overview of concepts and methods

In this section we explain the basic concepts and methods related to microdata protection. Sections 3.4.2, 3.4.3, 3.4.4 and 3.4.5 give in-depth descriptions of some particularly complex methods: microaggregation, rank swapping, additive noise and synthetic data (the first two implemented in $\mu$-ARGUS).

A microdata set $\mathbf{X}$ can be viewed as a file with $n$ records, where each record contains $m$ variables on an individual respondent. The variables can be classified in four categories which are not necessarily disjoint:

- *Identifiers.* These are variables that *unambiguously* identify the respondent. Examples are the passport number, social security number, etc.
- *Quasi-identifiers or key variables.* These are variables which identify the respondent with some degree of ambiguity. (Nonetheless, a combination of quasi-identifiers may provide unambiguous identification.) Examples are name, address, gender, age, telephone number, etc.

- *Confidential outcome variables.* These are variables which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- *Non-confidential outcome variables.* Those variables which do not fall in any of the categories above.

The purpose of SDC is to prevent confidential information from being linked to specific respondents. Therefore, we will assume in what follows that original microdata sets to be protected have been pre-processed so as to remove identifiers and quasi-identifiers with low ambiguity (such as name).

The purpose of microdata SDC mentioned in the previous section can be stated more formally by saying that, given an original microdata set $\mathbf{X}$, the goal is to release a protected microdata set $\mathbf{X}'$ in such a way that:

1. Disclosure risk (*i.e.* the risk that a user or an intruder can use $\mathbf{X}'$ to determine confidential variables on a specific individual among those in $\mathbf{X}$) is low.
2. User analyses (regressions, means, etc.) on $\mathbf{X}'$ and on $\mathbf{X}$ yield the same or at least similar results.

Microdata protection methods can generate the protected microdata set $\mathbf{X}'$

- either by *masking original data*, *i.e.* generating $\mathbf{X}'$ a modified version of the original microdata set $\mathbf{X}$;
- or by *generating synthetic data* $\mathbf{X}'$ that preserve some statistical properties of the original data $\mathbf{X}$.

Masking methods can in turn be divided in two categories depending on their effect on the original data (Willenborg and DeWaal, 2001):

- *Perturbative masking.* The microdata set is distorted before publication. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset.
- *Non-perturbative masking.* Non-perturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail in the original dataset. Global recoding, local suppression and sampling are examples of non-perturbative masking.

At a first glance, synthetic data seem to have the philosophical advantage of circumventing the re-identification problem: since published records are invented and do not derive from any original record, some authors claim that no individual having supplied original data can complain from having been re-identified. At a closer look, some authors (*e.g.*, Winkler, 2004 and Reiter, 2005) claim that even synthetic data might contain some records that allow for re-identification of confidential information. In short, synthetic data overfitted

to original data might lead to disclosure just as original data would. On the other hand, a clear problem of synthetic data is data utility: only the statistical properties explicitly selected by the data protector are preserved, which leads to the question whether the data protector should not directly publish the statistics he wants preserved rather than a synthetic microdata set.

So far in this section, we have classified microdata protection methods by their operating principle. If we consider the type of data on which they can be used, a different dichotomic classification applies:

- *Continuous data.* A variable is considered continuous if it is numerical and arithmetic operations can be performed with it. Examples are income and age. Note that a numerical variable does not necessarily have an infinite range, as is the case for age. When designing methods to protect continuous data, one has the advantage that arithmetic operations are possible, and the drawback that every combination of numerical values in the original dataset is likely to be unique, which leads to disclosure if no action is taken.
- *Categorical data.* A variable is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Ordinal scales and nominal scales can be distinguished among categorical variables. In ordinal scales the order between values is relevant, whereas in nominal scales it is not. In the former case, max and min operations are meaningful while in the latter case only pairwise comparison is possible. The instruction level is an example of ordinal variable, whereas eye colour is an example of nominal variable. In fact, all quasi-identifiers in a microdata set are normally categorical nominal. When designing methods to protect categorical data, the inability to perform arithmetic operations is certainly inconvenient, but the finiteness of the value range is one property that can be successfully exploited.

### 3.4.2 Perturbative masking

Perturbative statistical disclosure control (SDC) methods allow for the release of the entire microdata set, although perturbed values rather than exact values are released. Not all perturbative methods are designed for continuous data; this distinction is addressed further below for each method.

Most perturbative methods reviewed below (including noise addition, rank swapping, microaggregation and post-randomization) are special cases of matrix masking. If the original microdata set is $\mathbf{X}$, then the masked microdata set $\mathbf{X}'$ is computed as

$$\mathbf{X}' = \mathbf{AXB} + \mathbf{C}$$

where **A** is a record-transforming mask, **B** is a variable-transforming mask and **C** is a displacing mask or noise (Duncan and Pearson, 1991).

Table 3.2 lists the perturbative methods described below. For each method, the table indicates whether it is suitable for continuous and/or categorical data.

Table 3.2: Perturbative methods vs. data types. "X" denotes applicable and "(X)" denotes applicable with some adaptation.

| *Method* | *Continuous data* | *Categorical data* |
|---|---|---|
| Noise addition | X | |
| Microaggregation | X | (X) |
| Rank swapping | X | X |
| Rounding | X | |
| Resampling | X | |
| PRAM | | X |
| MASSC | | X |

#### 3.4.2.1 Noise addition

The main noise addition algorithms in the literature are:

- Masking by uncorrelated noise addition
- Masking by correlated noise addition
- Masking by noise addition and linear transformation
- Masking by noise addition and nonlinear transformation (Sullivan, 1989).

For more details on specific algorithms, the reader can check Brand (2002).

In practice, only a limited set of noise addition methods is more commonly used: the first three listed methods. When using linear transformations, a decision has to be made whether to reveal to the data user the parameter $c$ determining the transformations to allow for bias adjustment in the case of sub-populations.

With the exception of the not very practical method of Sullivan(1989), noise addition is not suitable to protect categorical data. On the other hand, it is well suited for continuous data for the following reasons:

- It makes no assumptions on the range of possible values for $\mathbf{X}_i$ (which may be infinite).
- The noise being added is typically continuous and with mean zero, which suits well with continuous original data.
- No exact matching is possible with external files. Depending on the amount of noise added, approximate (interval) matching might be possible. More details can be found in Section 3.4.2.

#### 3.4.2.2 Multiplicative Noise

One main challenge regarding additive noise with constant variance is that on one hand small values are strongly perturbed and on the other large values are weakly perturbed. For instance, in a business microdata set the large enterprises -- which are much easier to re-identify than the smaller ones -- remain still high at risk after noise addition. A possible way out is given by the multiplicative noise approach explained below.

⚠ Expert level

Let $\mathbf{X}$ be the matrix of the original data and $\mathbf{Z}$ the matrix of continuous perturbation variables with expectation 1 and variance $\sigma^2_{\mathbf{Z}} > 0$. The corresponding anonymised data $\mathbf{X}^a$ is then obtained as

$$(\mathbf{X}^a)_{ij} := \mathbf{Z}_{ij} \cdot \mathbf{X}_{ij}$$

for each pair $(i, j)$.
The following approach has been suggested by Höhne (2004). In a first step, for each record it is randomly decided whether its values are increased or decreased, each with 0.5-probability. This is done using the main factors $1 - f$ and $1 + f$. In order to avoid that all values of some record are perturbed with the same noise, these main factors are themselves perturbed with some additive noise $s$ (where $s < f/2$). The following transformation is needed to preserve the first and second moments of the distribution:

$$\mathbf{X}_i^{a^R} := \frac{\sigma_{\mathbf{X}}}{\sigma_{\mathbf{X}^a}} (\mathbf{X}_i^a - \mu_{\mathbf{X}^a}) + \mu_{\mathbf{X}},$$

where $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{X}^a}$ define the average of the original and anonymised variables, $\sigma_{\mathbf{X}}$ and $\sigma_{\mathbf{X}^a}$ the corresponding standard deviations, respectively.
Particularly if the original data follow a strongly skewed distribution, the deviations using this method may strongly depend on the configuration of the noise factors for some few, but large values. That is, despite consistency, means and sums might be unsatisfactorily reproduced. For this reason, (Höhne, 2004) suggests a slight modification of the method. At first, we generate normal distributed random variables $\mathbf{W}_i$ with expectation greater than zero and 'small' variance, s.t. the realisation of $\mathbf{W}_i$ yields a positive value. Afterwards, the data is sorted in descending order by the variable under consideration. Then, the record with the largest entry in this variable is diminished by

$$\mathbf{X}_1^a = (1 - \mathbf{W}_1)\, \mathbf{X}_1 \quad .$$

The records $\mathbf{X}_2, \ldots, \mathbf{X}_{n-1}$ are now perturbed as follows:

$$\mathbf{X}_i^a = \begin{cases} (1 - \mathbf{W}_i)\,\mathbf{X}_i, & \text{if} \quad \sum_{k=1}^{i-1} \mathbf{X}_k^a > \sum_{k=1}^{i-1} \mathbf{X}_k \\ (1 + \mathbf{W}_i)\,\mathbf{X}_i, & \text{if} \quad \sum_{k=1}^{i-1} \mathbf{X}_k^a \leq \sum_{k=1}^{i-1} \mathbf{X}_k \end{cases} .$$

Hence, means and sums are preserved and the diminishing and enlarging effects of single values cancel out each other. For the remaining record $\mathbf{X}_n$ we set

$$\mathbf{X}_n^a = \mathbf{X}_n - \left( \sum_{k=1}^{n-1} \mathbf{X}_k^a - \sum_{k=1}^{n-1} \mathbf{X}_k \right)$$

in order to preserve the overall sum.

#### 3.4.2.3 Microaggregation

Microaggregation is a family of SDC techniques for continuous microdata. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of $k$ or more individuals, where no individual dominates (*i.e.* contributes too much to) the group and $k$ is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation.

To obtain microaggregates in a microdata set with $n$ records, these are combined to form $g$ groups of size at least $k$. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) records can be published.

Microaggregation exists in several variants:

- Fixed vs. variable group (Defays and Nanopoulos, 1993), (Mateo-Sanz and Domingo-Ferrer, 1999), (Domingo-Ferrer and Mateo-Sanz, 2002), (Sande, 2002).
- Exact optimal vs. heuristic microaggregation (Hansen and Mukherjee, 2003), (Oganian and Domingo-Ferrer, 2001).
- Categorical microaggregation (V. Torra, 2004).

More details on the microaggregation implemented in $\mu$-ARGUS are given in Section 3.4.5.

#### 3.4.2.4 Data swapping and rank swapping

Data swapping was originally presented as an SDC method for databases containing only categorical variables (Dalenius and Reiss, 1978). The basic idea behind the method is to transform a database by exchanging values of confidential variables among individual records. Records are exchanged in such a way that low-order frequency counts or marginals are maintained.

Even though the original procedure was not very used in practice (see Fienberg and McIntyre, 2004), its basic idea had a clear influence in subsequent methods. In Reiss, Post and Dalenius (1982) and Reiss (1984) data swapping was introduced to protect continuous and categorical microdata, respectively. Another variant of data swapping for microdata is *rank swapping*. Although originally described only for ordinal variables (Greenberg, 1987), rank swapping can also be used for any numerical variable (Moore, 1996). First, values of a variable $\mathbf{X}_i$ are ranked in ascending order, then each ranked value of $\mathbf{X}_i$ is swapped with another ranked value randomly chosen within a restricted range (*e.g.* the rank of two swapped values cannot differ by more than $p\%$ of the total number of records, where $p$ is an input parameter). This algorithm is independently used on each original variable in the original data set.

It is reasonable to expect that multivariate statistics computed from data swapped with this algorithm will be less distorted than those computed after an unconstrained swap. In earlier empirical work by these authors on continuous microdata protection (Domingo-Ferrer and Torra, 2001), rank swapping has been identified as a particularly well-performing method in terms of the trade-off between disclosure risk and information loss. Consequently, it is one of the techniques that have been implemented in the $\mu$-ARGUS package (Hundepool *et al.*, 2005).

**Example.** In Table 3.5, we can see an original microdata set on the left and its rankswapped version on the right. There are four variables and ten records in the original dataset; the second variable is alphanumeric, and the standard alphabetic order has been used to rank it. A value of $p = 10$ has been used for all variables.

Table 3.3: Original file

| | | | |
|---|---|---|---|
| 1 | K | 3.7 | 4.4 |
| 2 | L | 3.8 | 3.4 |
| 3 | N | 3.0 | 4.8 |
| 4 | M | 4.5 | 5.0 |
| 5 | L | 5.0 | 6.0 |
| 6 | H | 6.0 | 7.5 |
| 7 | H | 4.5 | 10.0 |
| 8 | F | 6.7 | 11.0 |
| 9 | D | 8.0 | 9.5 |
| 10 | C | 10.0 | 3.2 |

Table 3.4: Rankswapped file

| | | | |
|---|---|---|---|
| 1 | H | 3.0 | 4.8 |
| 2 | L | 4.5 | 3.2 |
| 3 | M | 3.7 | 4.4 |
| 4 | N | 5.0 | 6.0 |
| 5 | L | 4.5 | 5.0 |
| 6 | F | 6.7 | 9.5 |
| 7 | K | 3.8 | 11.0 |
| 8 | H | 6.0 | 10.0 |
| 9 | C | 10.0 | 7.5 |
| 10 | D | 8.0 | 3.4 |

Table 3.5: Example of rank swapping.

#### 3.4.2.5 Rounding

Rounding methods replace original values of variables with rounded values. For a given variable $X_i$, rounded values are chosen among a set of rounding points defining a *rounding set*. In a multivariate original dataset, rounding is usually performed one variable at a time (*univariate* rounding); however, multivariate rounding is also possible (Willenborg and DeWaal, 2001). The operating principle of rounding makes it suitable for continuous data.

**Example** Assume a non-negative continuous variable $X$. Then we have to determine a set of rounding points$\{p_0, \cdots, p_r\}$. One possibility is to take rounding points as multiples of a base value $b$, that is, $p_i = bi$ for $i = 1, \cdots, r$. The set of attraction for each rounding point $p_i$ is defined as the interval $[p_i - b/2, p_i + b/2)$, for $i = 1$ to $r - 1$. For $p_0$ and $p_r$, respectively, the sets of attraction are $[0, b/2)$ and $[p_r - b/2, X_{\max}]$, where $X_{\max}$ is the largest possible value for variable $X$. Now an original value $x$ of $X$ is replaced with the rounding point corresponding to the set of attraction where $x$ lies.

#### 3.4.2.6 Resampling

Originally proposed for protecting tabular data (Heer, 1993), (Domingo-Ferrer and Mateo-Sanz, 1999), resampling can also be used for microdata. Take $t$ independent samples $S_1, \cdots, S_t$ of the values of an original variable $X_i$. Sort all samples using the same ranking criterion. Build the masked variable $Z_i$ as $\overline{x}_1, \cdots, \overline{x}_n$, where $n$ is the number of records and $\overline{x}_j$ is the average of the $j$-th ranked values in $S_1, \cdots, S_t$.

#### 3.4.2.7 PRAM

The Post-RAndomization Method or PRAM (Gouweleeuw *et al.*, 1997) is a probabilistic, perturbative method for disclosure protection of categorical variables in microdata files. In the masked file, the scores on some categorical variables for certain records in the original file are changed to a different score according to a prescribed probability mechanism, namely a Markov matrix. The Markov approach makes PRAM very general, because it encompasses noise addition, data suppression and data recoding.

PRAM information loss and disclosure risk largely depend on the choice of the Markov matrix and are still (open) research topics (De Wolf et al., 1999).

The PRAM matrix contains a row for each possible value of each variable to be protected. This rules out using the method for continuous data. More details on PRAM can be found in Section 3.4.6.

#### 3.4.2.8 MASSC

MASSC (Singh, Yu and Dunteman, 2003) is a masking method whose acronym summarizes its four steps: Micro Agglomeration, Substitution, Subsampling and Calibration. We briefly recall the purpose of those four steps:

1. Micro agglomeration is applied to partition the original dataset into risk strata (groups of records which are at a similar risk of disclosure). These strata are formed using the key variables, *i.e.* the quasi-identifiers in the records. The idea is that those records with rarer combinations of key variables are at a higher risk.
2. Optimal probabilistic substitution is then used to perturb the original data. (i.e. substitution is governed by a Markov matrix like in PRAM, see [Singh, Yu and Dunteman, 2003] for details)
3. Optimal probabilistic subsampling is used to suppress some variables or even entire records (i.e. variables and/or records are suppressed with a certain probability set as parameters).
4. Optimal sampling weight calibration is used to preserve estimates for outcome variables in the treated database whose accuracy is critical for the intended data use.

MASSC, to the best of our knowledge, is the first attempt at designing a perturbative masking method in such a way that disclosure risk can be analytically quantified. Its main shortcoming is that its disclosure model simplifies reality by considering only disclosure resulting from linkage of key variables with external sources. Since key variables are typically categorical, the uniqueness approach can be used to analyze the risk of disclosure; however, doing so ignores the fact that continuous outcome variables can also be used for respondent re-identification. As an example, if respondents are companies and turnover is one outcome variable, everyone in a certain industrial sector knows which is the company with largest turnover. Thus, in practice, MASSC is a method only suited when continuous variables are not present.

### 3.4.3 Non-perturbative masking

Non-perturbative masking does not rely on distortion of the original data but on partial suppressions or reductions of detail. Some of the methods are usable on both categorical and continuous data, but others are not suitable for continuous data. Table 3.6 lists the non-perturbative methods described below. For each method, the Table 3.6 indicates whether it is suitable for continuous and/or categorical data.

Table 3.6: Non-perturbative methods vs. data types.

| *Method* | *Continuous data* | *Categorical data* |
|---|---|---|
| Sampling | | X |
| Global recoding | X | X |
| Top and bottom coding | X | X |
| Local suppression | | X |

#### 3.4.3.1 Sampling

Instead of publishing the original microdata file, what is published is a sample $S$ of the original set of records.

Sampling methods are suitable for categorical microdata, but their adequacy for continuous microdata is less clear in a general disclosure scenario. The reason is that such methods leave a continuous variable $V_i$ unperturbed for all records in $S$. Thus, if variable $V_i$ is present in an external administrative public file, unique matches with the published sample are very likely: indeed, given a continuous variable $V_i$ and two respondents $o_1$ and $o_2$, it is highly unlikely that $V_i$ will take the same value for both $o_1$ and $o_2$ unless $o_1 = o_2$ (this is true even if $V_i$ has been truncated to represent it digitally).

If, for a continuous identifying variable, the score of a respondent is only approximately known by an attacker (as assumed in Willenborg and De Waal, 1996) it might still make

sense to use sampling methods to protect that variable. However, assumptions on restricted attacker resources are perilous and may prove definitely too optimistic if good quality external administrative files are at hand. For the purpose of illustration, the example below gives the technical specifications of a real-world application of sampling.

**Example** Statistics Catalonia released in 1995 a sample of the 1991 population census of Catalonia. The information released corresponds to 36 categorical variables (including the recoded versions of initially continuous variables); some of the variables are related to the individual person and some to the household. The technical specifications of the sample were as follows:

- Sampling algorithm: Simple random sampling.
- Sampling unit: Individuals in the population whose residence was in Catalonia as of March 1, 1991.
- Population size: 6,059,494 inhabitants
- Sample size: 245,944 individual records
- Sampling fraction: *0.0406*

With the above sampling fraction, the maximum absolute error for estimating a maximum-variance proportion is 0.2 percent.

#### 3.4.3.2 Global recoding

For a categorical variable $V_i$, several categories are combined to form new (less specific) categories, thus resulting in a new $V_i'$ with $|D(V_i')| < |D(V_i)|$ where $|\cdot|$ is the cardinality operator and $D(V_i)$ denotes the domain of variable $V_i$, *i.e.*, the possible values $V_i$ can have. For a continuous variable, global recoding means replacing $V_i$ by another variable $V_i'$ which is a discretized version of $V_i$. In other words, a potentially infinite range $D(V_i)$ is mapped onto a finite range $D(V_i')$. This is the technique used in $\mu$-ARGUS (Hundepool *et al.* 2005).

This technique is more appropriate for categorical microdata, where it helps disguise records with strange combinations of categorical variables. Global recoding is used heavily by statistical offices.

**Example.** If there is a record with "Marital status = Widow/er" and "Age = 17", global recoding could be applied to "Marital status" to create a broader category "Widow/er or divorced", so that the probability of the above record being unique would diminish. Global recoding can also be used on a continuous variable, but the inherent discretization leads very often to an unaffordable loss of information. Also, arithmetical operations that were straightforward on the original $V_i$ are no longer easy or intuitive on the discretized $V_i'$.

**Example.** We can recode the variable 'Occupation', by combining the categories 'Statistician' and 'Mathematician' into a single category 'Statistician or Mathematician'. When the number of female statisticians in Urk (a small town) plus the number of female mathematicians in Urk is sufficiently high, then the combination 'Place of residence = Urk', 'Gender = Female' and 'Occupation = Statistician or Mathematician' is considered safe for release. Note that instead of recoding 'Occupation' one could also recode 'Place of residence' for instance.

It is important to realise that global recoding is applied to the whole data set, not only to the unsafe part of the set. This is done to obtain a uniform categorisation of each variable. Suppose, for instance, that we recode the 'Occupation' in the above way. Suppose furthermore that both the combinations 'Place of residence = Amsterdam', 'Gender = Female' and 'Occupation = Statistician', and 'Place of residence = Amsterdam', 'Gender = Female' and 'Occupation = Mathematician' are considered safe. To obtain a uniform categorisation of 'Occupation' we would, however, not publish these combinations, but only the combination 'Place of residence = Amsterdam', 'Gender = Female' and 'Occupation = Statistician or Mathematician'.

#### 3.4.3.3 Top and bottom coding

Top and bottom coding is a special case of global recoding which can be used on variables that can be ranked, that is, continuous or categorical ordinal. The idea is that top values (those above a certain threshold) are lumped together to form a new category. The same is done for bottom values (those below a certain threshold). See the $\mu$-ARGUS manual (Hundepool *et al.* 2005).

#### 3.4.3.4 Local suppression

Certain values of individual variables are suppressed with the aim of increasing the set of records agreeing on a combination of key values. Ways to combine local suppression and global recoding are discussed in (De Waal and Willenborg, 1995) and implemented in $\mu$-ARGUS (Hundepool *et al.* 2005).

If a continuous variable $V_i$ is part of a set of key variables, then each combination of key values is probably unique. Since it does not make sense to systematically suppress the values of $V_i$, we conclude that local suppression is rather oriented to categorical variables.

When local suppression is applied, one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. For instance, in the above example we can protect the unsafe combination 'Place of residence = Urk', 'Gender = Female' and 'Occupation = Statistician' by suppressing the value of 'Occupation', assuming that the number of females in Urk is sufficiently high. The resulting combination is then given by 'Place of

residence = Urk', 'Gender = Female' and 'Occupation = missing'. Note that instead of suppressing the value of 'Occupation' one could also suppress the value of another variable of the unsafe combination. For instance, when the number of female statisticians in the Netherlands is sufficiently high then one could suppress the value of 'Place of residence' instead of the value of 'Occupation' in the above example to protect the unsafe combination. A local suppression is only applied to a particular value. When, for instance, the value of 'Occupation' is suppressed in a particular record, then this does not imply that the value of 'Occupation' has to be suppressed in another record. The freedom that one has in selecting the values that are to be suppressed allows one to minimise the number of local suppressions.

#### 3.4.3.5 References

Brand, R. (2002*). Microdata protection through noise addition.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 97–116, Berlin Heidelberg, 2002. Springer.

Dalenius T., and Reiss, S. P. (1978). *Data-swapping: a technique for disclosure control* (extended abstract). In Proc. of the ASA Section on Survey Research Methods, pages 191–194, Washington DC, 1978. American Statistical Association.

*Defays, D., and Nanopoulos, P. (1993). Panels of enterprises and confidentiality: the small aggregates method.* In Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, pages 195–204, Ottawa, 1993. Statistics Canada.

De Waal, A. G., and Willenborg, L.C.R.J. (1995). *Global recodings and local suppressions in microdata sets.* In Proceedings of Statistics Canada Symposium'95, pages 121–132, Ottawa, 1995. Statistics Canada.

De Waal, A.G. and Willenborg, L. C. R. J. (1999). *Information loss through global recoding and local suppression.* Netherlands Official Statistics, 14:17–20, 1999. special issue on SDC.

De Wolf, P.-P., Gouweleeuw, J. M., Kooiman, P., and Willenborg, L.C.R.J. (1999). *Reflections on PRAM.* In J. Domingo-Ferrer, editor, Statistical Data Protection, pages 337–349, Luxemburg, 1999. Office for Official Publications of the European Communities.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (1999). *On resampling for statistical confidentiality in contingency tables.* Computers & Mathematics with Applications, 38:13–32, 1999.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002). *Practical data-oriented microaggregation for statistical disclosure control.* IEEE Transactions on Knowledge and Data Engineering, 14(1):189–201, 2002.

Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001). *Comparing sdc methods for microdata on the basis of information loss and disclosure risk.* In Pre-proceedings of ETK-NTTS'2001 (vol. 2), pages 807–826, Luxemburg, 2001. Eurostat.

Domingo-Ferrer, J., and Torra, V., (2001*). Disclosure protection methods and information loss for microdata.* In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pages 91–110, Amsterdam, 2001. North-Holland. http://vneumann.etse.urv.es/publications/bcpi.

Duncan, G. T., and Pearson, R. W. (1991). *Enhancing access to microdata while protecting confidentiality: prospects for the future.* Statistical Science, 6:219–239, 1991.

Fienberg, S. E., and McIntyre, J. (2004). *Data swapping: variations on a theme by dalenius and reiss.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of LNCS, pages 14–29, Berlin Heidelberg, 2004. Springer.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and De Wolf, P.-P. (1997*). Post randomisation for statistical disclosure control: Theory and implementation*, Research paper no. 9731 (Voorburg: Statistics Netherlands).

Greenberg, B. (1987). *Rank swapping for ordinal data*, Washington, DC: U. S. Bureau of the Census (unpublished manuscript).

Hansen, S. L., and Mukherjee, S. (2003). *A polynomial algorithm for optimal univariate microaggregation.* IEEE Transactions on Knowledge and Data Engineering, 15(4):1043–1044, 2003.

Heer, G. R. (1993). *A bootstrap procedure to preserve statistical confidentiality in contingency tables.* In D. Lievesley, editor, Proc. of the International Seminar on Statistical Confidentiality, pages 261–271, Luxemburg, 1993. Office for Official Publications of the European Communities.

Höhne (2004), Varianten von Zufallsüberlagerung (German), working paper of the project group 'De facto anonymisation of business microdata', Wiesbaden.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P.-P., Domingo-Ferrer, J., Torra, V. and Giessing, S. (2005). *$\mu$-ARGUS version 4.0 Software and User's Manual.* Statistics Netherlands, Voorburg NL, may 2005. https://research.cbs.nl/casc.

Kooiman, P. L, Willenborg, L, and Gouweleeuw, J. M. (1998). *PRAM: A method for disclosure limitation of microdata.* Technical report, Statistics Netherlands (Voorburg, NL), 1998.

Mateo-Sanz, J. M., and Domingo-Ferrer, J. (1999) . *A method for data-oriented multivariate microaggregation.* In J. Domingo-Ferrer, editor, Statistical Data Protection, pages 89–99, Luxemburg, 1999. Office for Official Publications of the European Communities.

Moore, R. (1996). *Controlled data swapping techniques for masking public use microdata sets, 1996.* U. S. Bureau of the Census, Washington, DC, (unpublished manuscript).

Oganian, A., and Domingo-Ferrer, J. (2001). *On the complexity of optimal microaggregation for statistical disclosure control.* Statistical Journal of the United Nations Economic Commissions for Europe, 18(4):345–354, 2001.

Reiss, S. P, (1984). *Practical data-swapping: the first steps.* ACM Transactions on Database Systems, 9:20–37, 1984.

Reiss, S. P., Post, M. J., and Dalenius, T. (1982). *Non-reversible privacy transformations.* In Proceedings of the ACM Symposium on Principles of Database Systems, pages 139–146, Los Angeles, CA, 1982. ACM.

Reiter, J. P. (2005). *Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study.* Journal of the Royal Statistical Society, Series A, 168:185–205, 2005.

Sande, G. (2002). *Exact and approximate methods for data directed microaggregation in one or more dimensions.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):459–476, 2002.

Singh, A. C, Yu, F., and Dunteman, G. H. (2003) . *Massc: A new data mask for limiting statistical information loss and disclosure.* In H. Linden, J. Riecan, and L. Belsby, editors, Work Session on Statistical Data Confidentiality 2003, Monographs in Official Statistics, pages 373–394, Luxemburg, 2004. Eurostat.

Sullivan, G. R. (1989). *The Use of Added Error to Avoid Disclosure in Microdata Releases.* PhD thesis, Iowa State University, 1989.

Torra, V. (2004). *Microaggregation for categorical variables: a median based approach.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of LNCS, pages 162–174, Berlin Heidelberg, 2004. Springer.

Willenborg, L. and De Waal, T. (1996) . *Statistical Disclosure Control in Practice.* Springer-Verlag, New York, 1996.

Willenborg, L., and De Waal, T. (2001). *Elements of Statistical Disclosure Control.* Springer-Verlag, New York, 2001.

Winkler, W. E. (2004). *Re-identification methods for masked microdata.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of LNCS, pages 216–230, Berlin Heidelberg, 2004. Springer.

### 3.4.4 Noise addition

We sketch in this Section the operation of the main noise addition algorithms in the literature for microdata protection. For more details on specific algorithms, the reader can check (Brand, 2002).

#### 3.4.4.1 Masking by uncorrelated noise addition

Expert level

Masking by additive noise assumes that the vector of observations $x_j$ for the $j$-th variable of the original dataset $X_j$ is replaced by a vector where $\varepsilon_j$ is a vector of normally distributed errors drawn from a random variable $\varepsilon_j \sim N\left(0, \sigma^2_{\varepsilon_j}\right)$, such that $\text{Cov}\left(\varepsilon_t, \varepsilon_l\right) = 0$ for all $t \neq l$ (white noise).
The general assumption in the literature is that the variances of the $\varepsilon_j$ are proportional to those of the original variables. Thus, if $\sigma^2_j$ is the variance of $X_j$, then $\sigma^2_{\varepsilon_j} := \alpha\sigma^2_j$.
In the case of a $p$-dimensional dataset, simple additive noise masking can be written in matrix notation as $Z = X + \epsilon$, where $X \sim (\mu, \Sigma)$, $\varepsilon \sim (0, \Sigma_\varepsilon)$ and
$\Sigma_\varepsilon = \alpha \cdot \text{diag}\left(\sigma^2_1, \sigma^2_2, \cdots, \sigma^2_p\right)$, for $\alpha > 0$
This method preserves means and covariances, *i.e.*

$$\mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(\epsilon) = \mathbb{E}(X) = \mu \tag{3.4}$$

$$\text{Cov}(Z_j, Z_l) = \text{Cov}(X_j, X_l) \quad \forall j \neq l \tag{3.5}$$

Unfortunately, neither variances nor correlation coefficients are preserved:

$$\text{Var}\left(Z_j\right) = \text{Var}\left(X_j\right) + \alpha\,\text{Var}\left(X_j\right) = (1+\alpha)\,\text{Var}\left(X_j\right)$$

$$\rho(Z_j, Z_l) = \frac{\text{Cov}(Z_j, Z_l)}{\sqrt{\text{Var}(X_j)\,\text{Var}(X_l)}} = \frac{1}{1+\alpha}\rho(X_j, X_l), \forall j \neq l$$

#### 3.4.4.2 Masking by correlated noise addition

Expert level

Correlated noise addition also preserves means and additionally allows preservation of correlation coefficients. The difference with the previous method is that the covariance matrix of the errors is now proportional to the covariance matrix of the original data, *i.e.* $\varepsilon \sim (0, \Sigma)$, where $\Sigma_\varepsilon = \alpha\Sigma$.
With this method, we have that the covariance matrix of the masked data is

$$\Sigma_z = \Sigma + \alpha\Sigma = (1+\alpha)\Sigma. \tag{3.6}$$

Preservation of correlation coefficients follows, since

$$\rho(Z_j, Z_l) = \frac{1+\alpha}{1+\alpha} \frac{\text{Cov}\left(X_j, X_l\right)}{\sqrt{\text{Var}\left(X_j\right)\text{Var}\left(X_l\right)}} = \rho(X_j, X_l)$$

Regarding variances and covariances, we can see from Equation 3.6 that masked data only provide biased estimates for them. However, it is shown in Kim (1990) that the covariance matrix of the original data can be consistently estimated from the masked data as long as $\alpha$ is known.
As a summary, masking by correlated noise addition outputs masked data with higher analytical validity than masking by uncorrelated noise addition. Consistent estimators for several important statistics can be obtained as long as $\alpha$ is revealed to the data user. However, simple noise addition as discussed in this section and in the previous one is seldom used because of the very low level of protection it provides (Tendick, 1991), (Tendick and Matloff, 1994).

#### 3.4.4.3 Masking by noise addition and linear transformations

Expert level

In Kim (1986), a method is proposed that ensures by additional transformations that the sample covariance matrix of the masked variables is an unbiased estimator for the covariance matrix of the original variables. The idea is to use simple additive noise on the $p$ original variables to obtain overlayed variables

$$Z_j = X_j + \varepsilon_j, \quad \text{for } j = 1, \ldots, p$$

As in the previous section on correlated masking, the covariances of the errors $\varepsilon_j$ are taken proportional to those of the original variables. Usually, the distribution of errors is chosen to be normal or the distribution of the original variables, although

in Roque (2000) mixtures of multivariate normal noise are proposed.
In a second step, every overlayed variable $Z_j$ is transformed into a masked variable $G_j$ as

$$G_j = cZ_j + d_j$$

In matrix notation, this yields

$$Z = X + \varepsilon$$

$$G = cZ_j + D = c(X + \varepsilon) + D$$

where $X \sim N(\mu, \Sigma), \varepsilon \sim (0, \alpha\Sigma), G \sim (\mu, \Sigma)$ and $D$ is a matrix whose $j$-th column contains the scalar $d_j$ in all rows. Parameters $c$ and $d_j$ are determined under the restrictions that $\mathbb{E}(G_j) = \mathbb{E}(X_j)$ and $\text{Var}(G_j) = \text{Var}(X_j)$ for $j = 1, \cdots, p$. In fact, the first restriction implies that $d_j = (1-c)\mathbb{E}(X_j)$, so that the linear transformations depend on a single parameter $c$.
Due to the restrictions used to determine $c$, this methods preserves expected values and covariances of the original variables and is quite good in terms of analytical validity. Regarding analysis of regression estimates in subpopulations, it is shown in Kim (1990) that (masked) sample means and covariances are asymptotically biased estimates of the corresponding statistics on the original subpopulations. The magnitude of the bias depends on the parameter $c$, so that estimates can be adjusted by the data user as long as $c$ is revealed to her —revealing $c$ to the user has a fundamental disadvantage, though: the user can undo the linear transformation, so that this method becomes equivalent to plain uncorrelated noise addition (Domingo-Ferrer, Sebé, and Castellà, 2004)
The most prominent shortcomings of this method are that it does not preserve the univariate distributions of the original data and that it cannot be applied to discrete variables due to the structure of the transformations.

#### 3.4.4.4 Masking by noise addition and nonlinear transformations

⚠ Expert level

An algorithm combining simple additive noise and nonlinear transformation is proposed in Sullivan (1989). The advantages of this proposal are that it can be applied to discrete variables and that univariate distributions are preserved.
The method consists of several steps:

1. Calculate the empirical distribution function for every original variable.
2. Smooth the empirical distribution function.

3. Convert the smoothed empirical distribution function into a uniform random variable and this into a standard normal random variable.
4. Add noise to the standard normal variable.
5. Back-transform to values of the distribution function.
6. Back-transform to the original scale.

In the European project CASC (IST-2000-25069), the practicality and usability of this algorithm was assessed. Unfortunately, the internal CASC report by Brand (2002b) concluded that:
*"All in all, the results indicate that an algorithm as complex as the one proposed by Sullivan can only be applied by experts. Every application is very time-consuming and requires expert knowledge on the data and the algorithm."*

#### 3.4.4.5 Summary on noise addition

In practice, only simple noise addition or noise addition with linear transformation are used. When using linear transformations, a decision has to be made whether to reveal to the data user the parameter $c$ determining the transformations to allow for bias adjustment in the case of subpopulations.

With the exception of the not very practical method of Sullivan (1989), additive noise is not suitable to protect categorical data. On the other hand, it is well suited for continuous data for the following reasons:

- It makes no assumptions on the range of possible values for $\mathbf{X}_i$ (which may be infinite).
- The noise being added is typically continuous and with mean zero, which suits well continuous original data.
- No exact matching is possible with external files. Depending on the amount of noise added, approximate (interval) matching might be possible.

#### 3.4.4.6 References

Brand, R. (2002). *Microdata protection through noise addition.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 97–116, Berlin Heidelberg, 2002. Springer.

Brand, R. (2002b). *Tests of the applicability of sullivan's algorithm to synthetic data and real business data in official statistics*, European Project IST-2000-25069 CASC, Deliverable 1.1-D1, https://research.cbs.nl/casc.

Domingo-Ferrer, J., Sebé, F., and Castellà, J. (2004). *On the security of noise addition for privacy in statistical databases.* In J. Domingo-Ferrer and V. Torra, editors, Privacy

in Statistical Databases, volume 3050 of LNCS, pages 149–161, Berlin Heidelberg, 2004. Springer.

Kim, J. J. (1986). *A method for limiting disclosure in microdata based on random noise and transformation.* In Proceedings of the Section on Survey Research Methods, pages 303–308, Alexandria VA, American Statistical Association.

Kim, J. J. (1990). *Subpopulation estimation for the masked data.* In Proceedings of the ASA Section on Survey Research Methods, pages 456–461, Alexandria VA, 1990. American Statistical Association.

Roque, G. M. (2000).. *Masking Microdata Files with Mixtures of Multivariate Normal Distributions.* PhD thesis, University of California at Riverside, 2000.

Sullivan, G. R. (1989*). The Use of Added Error to Avoid Disclosure in Microdata Releases.* PhD thesis, Iowa State University.

Tendick, P. (1991). *Optimal noise addition for preserving confidentiality in multivariate data.* Journal of Statistical Planning and Inference, 27:341–353, 1991.

Tendick, P., and Matloff, N. (1994). *A modified random perturbation method for database security.* ACM Transactions on Database Systems, 19:47–63.

### 3.4.5 Microaggregation: further details

Consider a microdata set with $p$ continuous variables and $n$ records (*i.e.*, the result of recording $p$ variables on $n$ individuals). A particular record can be viewed as an instance of $\mathbf{X}' = \left(\mathbf{X}_1, \cdots, \mathbf{X}_p\right)$, where the $\mathbf{X}_i$ are the variables. With these individuals, $g$ groups are formed with $n_i$ individuals in the $i$-th group ($n_i \geq k$ and $n = \Sigma n_i$). Denote by $x_{\mathrm{ij}}$ the $j$-th record in the $i$-th group; denote by $\overline{x}_i$ the average record over the $i$-th group, and by $\overline{x}$ the average record over the whole set of $n$ individuals.

The optimal $k$-partition (from the information loss point of view) is defined to be the one that maximizes within-group homogeneity; the higher the within-group homogeneity, the lower the information loss, since microaggregation replaces values in a group by the group centroid. The sum of squares criterion is common to measure homogeneity in clustering. The within-groups sum of squares SSE is defined as

$$\mathrm{SSE} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(x_{\mathrm{ij}} - \overline{x}_i\right)^T \left(x_{\mathrm{ij}} - \overline{x}_i\right)$$

The lower SSE, the higher the within group homogeneity. The total sum of squares is

$$\mathrm{SST} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(x_{\mathrm{ij}} - \overline{x}\right)^T \left(x_{\mathrm{ij}} - \overline{x}\right)$$

In terms of sums of squares, the optimal $k$-partition is the one that minimizes SSE.

For a microdata set consisting of $p$ variables, these can be microaggregated together or partitioned into several groups of variables. Also the way to form groups may vary. We next review the main proposals in the literature.

**Example.** This example illustrates the use of microaggregation for SDC and, more specifically, for $k$-anonymization (Samarati and L. Sweeney, 1998), (Samarati, 2001), (Sweeney, 2002), (Domingo-Ferrer and Torra, 2005). A $k$-anonymous dataset allows no re-identification of a respondent within a group of at least $k$respondents. We show in Table 3.7 a dataset giving, for 11 small or medium enterprises (SMEs) in a certain town, the company name, the surface in square meters of the company's premises, its number of employees, its turnover and its net profit. Clearly, the company name is an identifier. We will consider that turnover and net profit are confidential outcome variables. A first SDC measure is to suppress the identifier "Company name" when releasing the dataset for public use. However, note that the surface of the company's premises and its number of employees can be used by a snooper as key variables. Indeed, it is easy for anybody to gauge to a sufficient accuracy the surface and number of employees of a target SME. Therefore, if the only privacy measure taken when releasing the dataset in Table 3.7 is to suppress the company name, a snooper knowing that company K&K Sarl has about a dozen employees crammed in a small flat of about 50 m will still be able to use the released data to link company K&K Sarl with turnover 645,223 Euros and net profit 333,010 Euros. Table 3.8 is a 3-anonymous version of the dataset in Table 3.7. The identifier "Company name" was suppressed and optimal bivariate microaggregation with $k = 3$ was used on the key variables "Surface" and "No. employees" (in general, if there are $p$ key variables, multivariate microaggregation with dimension $p$ should be used to mask all of

#### 3.4.5.1 Fixed vs. variable group size

Classical microaggregation algorithms (Defays and Nanopoulos, 1993) required that all groups except perhaps one be of size $k$; allowing groups to be of size $k$ depending on the structure of data was termed *data-oriented microaggregation* (Mateo-Sanz and Domingo-Ferrer, 1999), (Domingo-Ferrer and Mateo-Sanz, 2002). Figure 3.1 illustrates the advantages of variable-sized groups. If classical fixed-size microaggregation with $k = 3$ is used, we obtain a partition of the data into three groups, which looks rather unnatural for the data distribution given. On the other hand, if variable-sized groups are allowed then the five data on the left can be kept in a single group and the four data on the right in another group; such a variable-size grouping yields more homogeneous groups, which implies lower information loss.

However, except for specific cases such as the one depicted in Figure 3.1, the small gain in within-group homogeneity obtained with variable-sized groups hardly justifies the higher computational overhead of this option with respect to fixed-sized groups. This is particularly evident for multivariate data, as noted by Sande (2002).

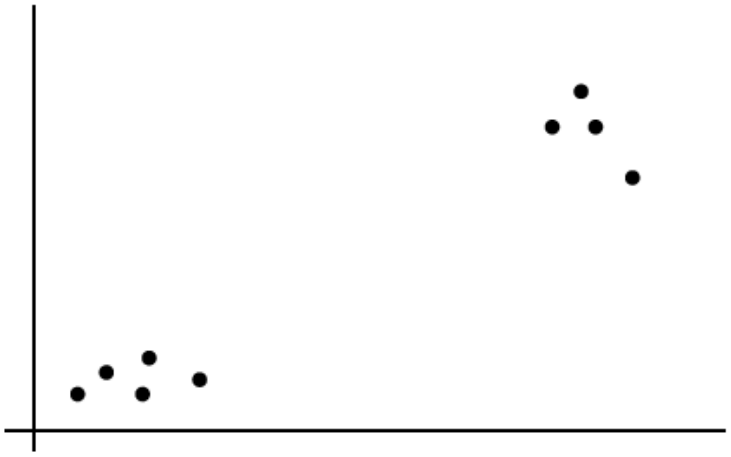

Figure 3.1: Variable-sized groups versus fixed-sized groups

#### 3.4.5.2 Exact optimal vs. heuristic microaggregation

For $p = 1$, *i.e.* a univariate dataset or a multivariate dataset where variables are microaggregated one at a time, an exact polynomial shortest-path algorithm exists to find the $k$-partition that optimally solves the microaggregation problem (Hansen and Mukherjee, 2003). See its description in Section 3.4.5.3.

For $p > 1$, finding an exact optimal solution to the microaggregation problem, *i.e.* finding a grouping where groups have maximal homogeneity and size at least $k$, has been shown to be NP-hard (Oganian and Domingo-Ferrer, 2001).

Unfortunately, the univariate optimal algorithm by Hansen and Mukherjee (2003) is not very useful in practice and this for two reasons: i) microdata sets are normally multivariate and using univariate microaggregation to microaggregate them one variable at a time is not good in terms of disclosure risk (see Domingo-Ferrer et al., 2002); ii) although polynomial-time, the optimal algorithm is quite slow when the number of records is large.

Thus, practical methods in the literature are heuristic:

- Univariate methods deal with multivariate datasets by microaggregating one variable at a time, *i.e.* variables are sequentially and independently microaggregated. These heuristics are known as individual ranking (Defays and Nanopoulos, 1993). While they are fast and cause little information loss, these univariate heuristics have the same problem of high disclosure risk as univariate optimal microaggregation.
- Multivariate methods either rank multivariate data by projecting them onto a single axis (*e.g.* using the first principal component or the sum of $z$-scores (Defays and Nanopoulos, 1993) or directly deal with unprojected data (Mateo-Sanz and Domingo-Ferrer, 1999), (Domingo-Ferrer and Mateo-Sanz, 2002). When working on unprojected data, we can microaggregate all variables of the dataset at a time, or independently microaggregate groups of two variables at a time, three variables at a time, etc. In any case, it is preferable that variables within a group which is microaggregated at a time be correlated (W.E. Winkler, 2004) in order to keep as much as possible the analytic properties of the file.

We next describe the two microaggregation algorithms implemented in $\mu$-ARGUS.

#### 3.4.5.3 Hansen-Mukherjee's optimal univariate microaggregation

⚠ Expert level

In Hansen and Mukherjee (2003) a polynomial-time algorithm was proposed for *univariate* optimal microaggregation. Authors formulate the microaggregation problem as a shortest-path problem on a graph. They first construct the graph and then show that the optimal microaggregation corresponds to the shortest path in this graph. Each arc of the graph corresponds to a possible group that may be part of an optimal partition. The arc label is the SSE that would result if that group were to be included in the partition. We next detail the graph construction.
Let $\mathbf{V} = \{v_1, \cdots, v_n\}$ be a vector consisting of $n$ real numbers sorted into ascending order, so that $v_1$ is the smallest value and $v_n$ the largest value. Let $k$ be an integer group size such that $1 \leq k < n$. Now, a graph $G_{n,k}$ is constructed as follows:

1. For each value $\mathbf{X}_i$ in $\mathbf{X}$, create a node with label $i$. Create also an additional node with label 0.
2. For each pair of graph nodes $(i, j)$ such that $1+k \leq j < i+2k$, create a directed

arc $(i, j)$from node $i$ to node $j$.

3. Map each arc $(i, j)$ to the group of values $C(i, j) = \{\mathbf{X}_h : i < h \leq j\}$. Let the length $L(i, j)$ of the arc be the within group sum of squares for $C(i, j)$, that is,

$$L(i,j) = \sum_{h=i+1}^{j} \left(\mathbf{X}_h - \overline{\mathbf{X}}_{(i,j)}\right)^2$$

where $\overline{\mathbf{X}}_{(i,j)} = \frac{1}{j-i} \sum_{h=i+1}^{j} \mathbf{X}_h$

It is proven in Hansen and Mukherjee (2003) that the optimal $k$-partition for $V$ is found by taking as groups the $C(i, j)$ corresponding to the arcs in the shortest path between nodes 0 and $n$. For minimal group size $k$ and a dataset of $n$ real numbers sorted in ascending order, the complexity of this optimal univariate microaggregation is $O\left(k^2 n\right)$, that is, linear in the size of the dataset.

#### 3.4.5.4 The MDAV heuristic for multivariate microaggregation

Expert level

The multivariate microaggregation heuristic implemented in $\mu$-ARGUS is called MDAV (Maximum Distance to Average Vector). MDAV performs multivariate fixed group size microaggregation on unprojected data. MDAV is also described in Domingo-Ferrer and Torra (2005).
MDAV Algorithm

1. Compute the average record $\overline{x}$ of all records in the dataset. Consider the most distant record $x_r$ to the average record $\overline{x}$ (using the squared Euclidean distance).
2. Find the most distant record $x_s$ from the record $x_r$ considered in the previous step.
3. Form two groups around $x_r$ and $x_s$, respectively. One group contains $x_r$ and the $k-1$ records closest to $x_r$. The other group contains $x_s$ and the $k-1$ records closest to $x_s$.
4. If there are at least 3k records which do not belong to any of the two groups formed in Step 3, go to Step 1 taking as new dataset the previous dataset minus the groups formed in the last instance of Step 3.
5. If there are between $3k-1$ and $2k$ records which do not belong to any of the two groups formed in Step 3:

   a) compute the average record $\overline{x}$ of the remaining records;
   b) find the most distant record $x_r$ from $\overline{x}$;

c) form a group containing $x_r$ and the $k-1$ records closest to $x_r$;
d) form another group containing the rest of records. Exit the Algorithm.

6. If there are less than $2k$ records which do not belong to the groups formed in Step 3, form a new group with those records and exit the Algorithm.

The above algorithm can be applied independently to each group of variables resulting from partitioning the set of variables in the dataset.

#### 3.4.5.5 Categorical microaggregation

Recently (Torra, 2004), microaggregation has been extended to categorical data. Such an extension is based on existing definitions for aggregation and clustering, the two basic operations required in microaggregation. Specifically, the median is used for aggregating ordinal data and the plurality rule (voting) for aggregating nominal data. Clustering of categorical data is based on the $k$-modes algorithm, which is a partitive clustering method similar to $c$-means.

#### 3.4.5.6 References

Defays, D., and Nanopoulos, P. (1993*). Panels of enterprises and confidentiality: the small aggregates method.* In Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, pages 195–204, Ottawa, 1993. Statistics Canada.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002). *Practical data-oriented microaggregation for statistical disclosure control.* IEEE Transactions on Knowledge and Data Engineering, 14(1):189–201, 2002.

Domingo-Ferrer, J., Mateo-Sanz, J. M., Oganian, A., and Torres, À. (2002). *On the security of microaggregation with individual ranking: analytical attacks.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):477–492, 2002.

Domingo-Ferrer, J., and Torra, V. (2005). *Ordinal, continuous and heterogenerous k-anonymity through microaggregation.* Data Mining and Knowledge Discovery, 11(2):195–212, 2005.

Hansen, S. L. and Mukherjee, S. (2003*). A polynomial algorithm for optimal univariate microaggregation.* IEEE Transactions on Knowledge and Data Engineering, 15(4):1043–1044, 2003.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P.-P., Domingo-Ferrer, J., Torra, V., and Giessing, S. (2005*). μ-ARGUS version 4.0 Software and User's Manual.* Statistics Netherlands, Voorburg NL, May 2005. https://research.cbs.nl/casc.

Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1999). *A method for data-oriented multivariate microaggregation.* In J. Domingo-Ferrer, editor, Statistical Data Protection, pages 89–99, Luxemburg, 1999. Office for Official Publications of the European Communities.

Oganian, A:, and Domingo-Ferrer, J. (2001). *On the complexity of optimal microaggregation for statistical disclosure control.* Statistical Journal of the United Nations Economic Comission for Europe, 18(4):345–354, 2001.

Samarati, P. (2001). *Protecting respondents' identities in microdata release.* IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, 2001.

Samarati, P., and Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical report, SRI International, 1998.

Sande, G. (2002). *Exact and approximate methods for data directed microaggregation in one or more dimensions.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):459–476, 2002.

Sweeney, L. (2002*). k-anonimity: A model for protecting privacy.* International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10(5):557–570, 2002.

Torra, V. (2004). *Microaggregation for categorical variables: a median based approach.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of Lecture Notes in Computer Science, pages 162–174, Berlin Heidelberg, 2004. Springer.

Winkler, W. E. (2004). *Masking and re-identification methods for public-use microdata: overview and research problems.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of Lecture Notes in Computer Science, pages 231–246, Berlin Heidelberg, 2004. Springer.

### 3.4.6 PRAM

PRAM is a disclosure control technique that can be applied to categorical data. Basically, it is a form of intended misclassification, using a known and predetermined probability mechanism. Applying PRAM means that for each record in a microdata file, the score on one or more categorical variables is changed with a certain probability. This is done independently for each of the records. PRAM is thus a perturbative method. Since PRAM uses a probability mechanism, the disclosure risk is directly influenced by this method. An intruder can never be certain that a record she thinks she has identified is indeed the identified person: with a certain probability this has been a perturbed record.

Since the probability mechanism that is used when applying PRAM is known, characteristics of the (latent) true data can still be estimated from the perturbed data file. To that end, one can make use of correction methods similar to those used in case of misclassification and randomised response situations.

PRAM was used in 2001 UK Census to produce an end-user licence version of the Samples of Anonymised Records (SARs). See Gross(2004) www.ccsr.ac.uk/sars/events/2004-09-30/gross.pdf for a full description.

#### 3.4.6.1 PRAM, the method

Expert level

In this section a short theoretical description of PRAM is given. For a detailed description of the method, see e.g., Gouweleeuw *et al.* (1998a and 1998b). For a discussion of several issues concerning the method and its consequences, see *e.g.*, De Wolf *et al.* (1998).
Let $\xi$ denote a categorical variable in the original file to which PRAM will be applied and let $X$ denote the same variable in the perturbed file. Moreover, assume that $\xi$, and hence $X$ as well, has $K$ categories, labelled $1, \ldots, K$. The probabilities that define PRAM are denoted as

$$p_{\text{kl}} = \mathbb{P}(X = l \mid \xi = k)$$

*i.e.*, the probability that an original score $\xi = k$ is changed into the score $X = l$. These so called transition probabilities are defined for all $k, l = 1, ..., K$.
Using these transition probabilities as entries of a $K \times K$ matrix, we obtain a Markov matrix that we will call the PRAM-matrix, denoted by $\mathbf{P}$.
Applying PRAM now means that, given the score $\xi = k$ for record $r$, the score $X$ for that record is drawn from the probability distribution $p_{k1}, \ldots, p_{kK}$. For each record in the original file, this procedure is performed independently of the other records.
To illustrate the ideas, suppose that the variable $\xi$ is gender, with scores $\xi = 1$ if male and $\xi = 2$ if female. Applying PRAM with $p_{11} = p_{22} = 0.9$ on a microdata file with 110 males and 90 females, would yield a perturbed microdata file with *in expectation*, 108 males and 92 females. However, in expectation, 9 of these males were originally female, and similarly, 11 of the females were originally male.
*Correcting analyses*
More generally, the effect of PRAM on one-dimensional frequency tables is that

$$\mathbb{E}(T_X \mid \xi) = \mathbf{P}^t T_\xi$$

where $T_\xi = (T_\xi(1), \ldots, T_\xi(K))^T$ denotes the frequency table according to the original microdata file and $T_X$ the frequency table according to the perturbed microdata file. A conditionally unbiased estimator of the frequency table in the original file is then given by

$$\hat{T}_\xi = \left(\mathbf{P}^{-1}\right)^t T_X$$

This can be extended to two-dimensional frequency tables, by vectorizing such tables. The corresponding PRAM-matrix is then given by the Kronecker product of the PRAM-matrices of the individual dimensions.
Alternatively, one could use the two-dimensional frequency tables[1] $T_{\xi\eta}$ for the original data and $T_{XY}$ for the perturbed data directly in matrix notation:

$$\hat{T}_{\xi\eta} = \left(\mathbf{P}_X^{-1}\right)^t T_{XY} \mathbf{P}_Y^{-1}$$

where $\mathbf{P}_X$ denotes the PRAM-matrix corresponding to the categorical variable $X$ and $\mathbf{P}_Y$ denotes the PRAM-matrix corresponding to the categorical variable $Y$.
For more information about correction methods for statistical analyses applied to data that have been protected with PRAM, we refer to e.g., Gouweleeuw *et al.* (1998a) and Van den Hout (1999 and 2004).
*Choice of PRAM-matrix*
The exact choice of transition probabilities influences both the amount of information loss as well as the amount of disclosure limitation. Moreover, in certain situations 'illogical' changes could occur, e.g., changing the gender of a female respondent with ovarian cancer to male. These kind of changes would attract the attention of a possible intruder which should be avoided.
It is thus important to choose the transition probabilities in an appropriate way. Illogical changes could be avoided by appointing a probability of 0 to the illogical scores. In the example given above, PRAM should not be applied to the variable gender individually, but to the crossing of the variables gender and diseases. In that case, each transition probability of changing a score into the score (male, ovarian cancer) should be set equal to 0.
The choice of the transition probabilities in relation to the disclosure limitation and the information loss is more delicate. An empirical study on these effects is given in De Wolf and Van Gelder (2004). A theoretical discussion on the possibility to choose the transition probabilities in an optimal way (in some sense) is given in Cator *et al.* (2005).

#### 3.4.6.2 When to use PRAM

In certain situations methods like global recoding, local suppression and top-coding would yield too much loss of detail in order to produce a safe microdata file. In these circumstances, PRAM is an alternative. Using PRAM, the amount of detail is preserved whereas the level of disclosure control is achieved by introducing uncertainty in the scores on identifying variables.

However, in order to make adequate inferences on a microdata file to which PRAM has been applied, the statistician needs to include sophisticated changes to the standard meth-

[1] http://www.ihsn.org/software/disclosure-control-toolbox

ods. This demands a good knowledge of both PRAM and the statistical analysis that is to be applied.

In case a researcher is willing to make use of a remote execution facility, PRAM might be used to produce a microdata file with the same structure as the original microdata file, but with some kind of synthetic data. Such microdata files might be used as a 'test' microdata file on which a researcher can try her scripts before sending these scripts to the remote execution facility. Since the results of the script are not used directly, the amount of distortion of the original microdata file can be chosen to be quite large. That way a safe microdata file is produced that still exhibits the same structure (and amount of detail) as the original microdata file.

In other situations, PRAM might produce a microdata file that is safe and leaves certain statistical characteristics of that file (more or less) unchanged. In that case, a researcher might perform his research on that microdata file in order to get an idea on the eventually needed research strategy. Once that strategy has been determined, the researcher might come to an on-site facility in order to perform the analyses once more on the original microdata hence reducing the amount of time that she has to be at the on-site facility.

#### 3.4.6.3 References on PRAM

Gross, B., Guiblin,Ph, and K. Merrett (2004), *Implementing the Post Randomisation method To the Individual Sample of Anonymised Records (SAR) from the 2001 Census* , Office for National Statistics. www.ccsr.ac.uk/sars/events/2004-09-30/gross.pdf

Cator, E., Hensbergen A. and Y. Rozenholc (2005), *Statistical Disclosure Control using PRAM*, Proceedings of the 48th European Study Group Mathematics with Industry, Delft, The Netherlands, 15-19 March 2004. Delft University Press, 2005, p. 23 – 30.

Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg and P.P. de Wolf (1998a), *Post Randomisation for Statistical Disclosure Control: Theory and Implementation*, Journal of Official Statistics, Vol. 14, 4, pp. 463 – 478.

Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg and P.P. de Wolf (1998b), *The post randomisation method for protecting microdata*, Qüestiió, Quaderns d'Estadística i Investigació Operativa, Vol. 22, 1, pp. 145 – 156.

Van den Hout, A. (2000), *The analysis of data perturbed by PRAM*, Delft University Press, ISBN 90-407-2014-2.

Van den Hout, A. (2004), *Analyzing misclassified data: randomized response and post randomization*, Ph.D. thesis, Utrecht University.

De Wolf, P.P. and I. Van Gelder (2004), *An empirical evaluation of PRAM*, Discussion paper 04012, Statistics Netherlands. This paper can also be found on the CASC-Website (https://research.cbs.nl/casc Related Papers)

De Wolf, P.P., J.M. Gouweleeuw, P. Kooiman and L.C.R.J. Willenborg (1998), *Reflections on PRAM*, Proceedings of the conference "Statistical Data Protection", March 25-27 1998, Lisbon, Portugal. This paper can also be found on the CASC-Website (https://research.cbs.nl/casc/related/Sdp_98_2.pdf)

### 3.4.7 Synthetic microdata

Publication of synthetic —*i.e.* simulated— data was proposed long ago as a way to guard against statistical disclosure. The idea is to randomly generate data with the constraint that certain statistics or internal relationships of the original dataset should be preserved.

We next review some approaches in the literature to synthetic data generation and then proceed to discuss the global pros and cons of using synthetic data.

#### 3.4.7.1 A forerunner: data distortion by probability distribution

Data distortion by probability distribution was proposed in 1985 (Liew, Choi and Liew, 1985) and is not usually included in the category of synthetic data generation methods. However, its operating principle is to obtain a protected dataset by randomly drawing from the underlying distribution of the original dataset. Thus, it can be regarded as a forerunner of synthetic methods.

This method is suitable for both categorical and continuous variables and consists of three steps:

1. Identify the density function underlying to each of the confidential variables in the dataset and estimate the parameters associated with that density function.
2. For each confidential variable, generate a protected series by randomly drawing from the estimated density function.
3. Map the confidential series to the protected series and publish the protected series instead of the confidential ones.

In the identification and estimation stage, the original series of the confidential variable (*e.g.* salary) is screened to determine which of a set of predetermined density functions fits the data best. Goodness of fit can be tested by the Kolmogorov-Smirnov test. If several density functions are acceptable at a given significance level, selecting the one yielding the smallest value for the Kolmogorov-Smirnov statistics is recommended. If no density in the predetermined set fits the data, the frequency imposed distortion method can be used. With the latter method, the original series is divided into several intervals (somewhere between 8 and 20). The frequencies within the interval are counted for the original series, and become a guideline to generate the distorted series. By using a uniform random number generating subroutine, a distorted series is generated until its frequencies become

the same as the frequencies of the original series. If the frequencies in some intervals overflow, they are simply discarded.

Once the best-fit density function has been selected, the generation stage feeds its estimated parameters to a random value generating routine to produce the distorted series.

Finally, the mapping and replacement stage is only needed if the distorted variables are to be used jointly with other non-distorted variables. Mapping consists of ranking the distorted series and the original series in the same order and replacing each element of the original series with the corresponding distorted element.

It must be stressed here that the approach described in (Liew, Choi and Liew, 1985) was for one variable at a time. One could imagine a generalization of the method using multivariate density functions. However such a generalization: i) is not trivial, because it requires multivariate ranking-mapping; and ii) can lead to very poor fitting.

**Example** A distribution fitting software (Crystal.Ball, 2004) has been used on the original (ranked) data set 186, 693, 830, 1177, 1219, 1428, 1902, 1903, 2496, 3406. Continuous distributions tried were normal, triangular, exponential, lognormal, Weibull, uniform, beta, gamma, logistic, Pareto and extreme value; discrete distributions tried were binomial, Poisson, geometric and hypergeometric. The software allowed for three fitting criteria to be used: Kolmogorov-Smirnov, $\chi^2$ and Anderson-Darling. According to the first criterion, the best fit happened for the extreme value distribution with modal and scale parameters 1105.78 and 732.43, respectively; the Kolmogorov statistic for this fit was 0.1138. Using the fitted distribution, the following (ranked) dataset was generated and used to replace the original one: 425.60, 660.97, 843.43, 855.76, 880.68, 895.73, 1086.25, 1102.57, 1485.37, 2035.34.

#### 3.4.7.2 Synthetic data by multiple imputation

Rubin (1993) suggested creating an entirely synthetic dataset based on the original survey data and multiple imputations. Rubin's proposal was more completely developed in Raghunathan, Reiter, and Rubin (2003). A simulation study of it was given in Reiter (2002). In Reiter (2005) inference on synthetic data is discussed and in Reiter (2005b) an application is given.

We next sketch the operation of the original proposal by Rubin. Consider an original microdata set $X$ of size $n$ records drawn from a much larger population of $N$ individuals, where there are background variables $A$, non-confidential variables $B$ and confidential variables $C$. Background variables are observed and available for all $N$ individuals in the population, whereas $B$ and $C$ are only available for the $n$ records in the sample $X$. The first step is to construct from $X$ a multiply-imputed population of $N$ individuals. This population consists of the $n$ records in $X$ and $M$ (the number of multiple imputations,

typically between 3 and 10) matrices of $(B, C)$ data for the $N-n$ non-sampled individuals. The variability in the imputed values ensures, theoretically, that valid inferences can be obtained on the multiply-imputed population. A model for predicting $(B, C)$ from $A$ is used to multiply-impute $(B, C)$ in the population. The choice of the model is a nontrivial matter. Once the multiply-imputed population is available, a sample $Z$ of $n'$ records can be drawn from it whose structure looks like the one a sample of $n'$ records drawn from the original population. This can be done $M$ times to create $M$ replicates of $(B, C)$ values. The results are $M$ multiply-imputed synthetic datasets. To make sure no original data are in the synthetic datasets, it is wise to draw the samples from the multiply-imputed population excluding the $n$ original records from it.

#### 3.4.7.3 Synthetic data by bootstrap

⚠ Expert level

Fienberg (1994) proposed generating synthetic microdata by using bootstrap methods. Later, in Fienberg, Makov and Steele (1998), this approach was used for categorical data.
The bootstrap approach bears some similarity to the data distortion by probability distribution and the multiple-imputation methods described above. Given an original microdata set $X$ with $p$ variables, the data protector computes its empirical $p$-variate cumulative distribution function (c.d.f.) $F$. Now, rather than distorting the original data to obtain masked data, the data protector alters (or "smoothes") the c.d.f. $F$ to derive a similar c.d.f. $F'$. Finally, $F'$ is sampled to obtain a synthetic microdata set $Z$.

#### 3.4.7.4 Synthetic data by Latin Hypercube Sampling

⚠ Expert level

Latin Hypercube Sampling (LHS) appears in the literature as another method for generating multivariate synthetic datasets. In Huntington and Lyrintzis (1998), the LHS updated technique of Florian (1992) was improved, but the proposed scheme is still time-intensive even for a moderate number of records. In Dandekar, Cohen and Kirkendall (2002) LHS is used along with a rank correlation refinement to reproduce both the univariate (*i.e.* mean and covariance) and multivariate structure (in the sense of rank correlation) of the original dataset. In a nutshell, LHS-based methods rely on iterative refinement, are time-intensive and their running time does not only depend on the number of values to be reproduced, but on the starting values as well.

#### 3.4.7.5 Partially synthetic data by Cholesky decomposition

Expert level

Generating plausible synthetic values for all variables in a database may be difficult in practice. Thus, several authors have considered mixing actual and synthetic data. In Burridge (2004) a family of methods known as IPSO (Information Preserving Statistical Obfuscation) is proposed for generation of partially synthetic data. It consists of three methods that are described next.

*Method A: The basic IPSO procedure*

The basic form of IPSO will be called here Method A. Informally, suppose two sets of variables $X$ and $Y$, where the former are the confidential outcome variables and the latter are quasi-identifier variables. Then $X$ are taken as independent and $Y$ as dependent variables. A multiple regression of $Y$ on $X$ is computed and fitted $Y'_A$ variables are computed. Finally, variables $X$ and $Y'_A$ are released in place of $X$ and $Y$.

More formally, let $y$and $x$ be two data matrices, with rows representing respondents and columns representing variables; the row vectors $y_i$ and $x_i$ will represent the data for the $i$-th respondent, for $i = 1, \cdots, n$. The column vector $u_j$ will represent the quasi-identifier variable $j$, for $j = 1, \cdots, p$; in other words, the $u_j$ are the columns of quasi-identifier matrix $Y$. Conditionally on the specific values for confidential variables, quasi-identifier variables for different respondents are assumed to be independent. Conditional on the specific confidential variables $x_i$, the quasi-identifier variables $Y_i$ are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where $B$ is an $m \times p$ matrix with columns $\beta_j$. Thus a separate univariate normal multiple regression model is assumed for each column of $Y$ with regression parameter equal to the corresponding column of $B$, that is, $U_j \sim N\left(x\beta_j, \sigma_{jj} I\right)$.

Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of $B$ and $\Sigma$ derived from the complete dataset $(y, x)$. These estimates are a pair of sufficient statistics for the regression model. We denote in what follows the vectors of fitted values and residuals for $u_j$ as $\hat{\mu}_j$ and $\hat{r}_j$, respectively. Thus, $\hat{\mu}$, $\hat{r}$ and $\hat{\Sigma}$ will denote the matrices $x\hat{B}$, $y - x\hat{B}$ and $n^{-1}\hat{r}^t\hat{r}$, respectively.

The output of IPSO Method A is $y'_A = x\hat{B}$.

*Method B: IPSO preserving $\hat{B}$*

If a user fits a multiple regression model to $(y'_A, x)$, she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data $(y, x)$.

IPSO Method B modifies $y'_A$ into $y'_B$ in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from $(y'_B, x)$ satisfies $\hat{B}_B = \hat{B}$.

Suppose that $\tilde{y}$ is a new, artificial, set of quasi-identifier values. These can be any set of numbers initially, *e.g.* an i.i.d. normal random sample or a deterministically

chosen set. For each component new residuals $\tilde{r}_j$ are calculated by fitting the above multivariate multiple regression to the new "data" $\tilde{y}$. Define

$$y'_B = \hat{\mu} + \tilde{r} = x\hat{B} + \tilde{r}$$

The following information preservation result holds for IPSO-B.

**Lemma 3.3.7.1.** *Regardless of the initial choice $\tilde{y}$, $(y'_B, x)$ preserves the sufficient statistic $\hat{B}$.*

**Proof:** We have that

$$y'_B = x\hat{B} + \tilde{r} = x\hat{B} + \left(\tilde{y} - x\tilde{B}\right) \tag{3.7}$$

where $\hat{B}$ is the MLE estimate of $B$ obtained from $(\tilde{y}, x)$. Now, the expressions of $\hat{B}$ and $\tilde{B}$ are, respectively,

$$\hat{B} = \left(x^t x\right)^{-1} x^t y$$

and

$$\tilde{B} = \left(x^t x\right)^{-1} x^t \tilde{y}$$

Analogously, the expression of the MLE estimate of $\hat{B}_B$ obtained from $(y'_B, x)$ is

$$\hat{B}_B = \left(x^t x\right)^{-1} x^t y'_B$$

Substituting expression (3.7) for $y'_B$ in the equation above, we get

$$\hat{B}_B = \left(x^t x\right)^{-1} \left(x^t x\right) \hat{B} + \left(x^t x\right)^{-1} x^t (\tilde{y} - x\tilde{B}) = \hat{B} + \tilde{B} - \tilde{B} = \hat{B}$$

*Method C: IPSO preserving $\hat{B}$ and $\hat{\Sigma}$*

A more ambitious goal is to come up with a data matrix $y'_C$ such that, when a multivariate multiple regression model is fitted to $(y'_C, x)$, *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data $(y, x)$ are preserved.

The algorithm proposed in Burridge (2004) to get $y'_C$ is as follows

1. Generate provisional new "data" $\tilde{y}$ (this will be an $n \times p$ matrix).
2. Calculate provisional new residuals $\tilde{r}$ by fitting the multiple regression model to each column of $\tilde{y}$.
3. Define new residuals $\tilde{r}'$ as a transformation of $\tilde{r}$ so that $\tilde{r}^t\tilde{r}' = n\hat{\Sigma}$. This is easily done as follows:
   a) Let $L$ and $L_O$ be the lower triangular matrices in the Cholesky factorizations $n\hat{\Sigma} = LL^t$ and $\tilde{r}^t\tilde{r} = L_O L_O^t$.
   b) Define $\tilde{r}' = \tilde{r}\left(L_O^{-1}\right)^t L^t$. It is easily verified that $(\tilde{r}')^t\tilde{r}' = n\hat{\Sigma}$.

Information preservation in IPSO-C is as follows.

Define

$$y'_C = x\hat{B} + \tilde{r}'$$

**Lemma 3.3.7.2.** *$(y'_C, x)$ preserves the sufficient statistics $\hat{B}$ and $\hat{\Sigma}$.*
**Proof:** The expression of the MLE estimate of $\hat{B}$ obtained from $(y'_C, x)$ is

$$\hat{B}_C = (x^t x)^{-1} x^t y'_C = (x^t x)^{-1} x^t \left(x\hat{B} + \tilde{r}'\right) \tag{3.8}$$
$$= \hat{B} + (x^t x)^{-1} x^t \tilde{r} L_O^t L^t = \hat{B} + (x^t x)^{-1} x^t \left(\tilde{y} - x\tilde{B}\right) L_O^t L^t \tag{3.9}$$
$$= \hat{B} + \left(\tilde{B} - \tilde{B}\right) L_O^t L^t = \hat{B} \tag{3.10}$$
$$\tag{3.11}$$

Using that $\hat{B}_C = \hat{B}$, the expression of the MLE estimate of $\hat{\Sigma}$ obtained from $(y'_C, x)$ is

$$\hat{\Sigma}_C = \frac{\left(y'_C, x\hat{B}\right)^t \left(y'_C, x\hat{B}\right)}{n} \tag{3.12}$$
$$= \frac{\left(x\hat{B} + \tilde{r}' - x\hat{B}\right)^t \left(x\hat{B} + \tilde{r}' - x\hat{B}\right)}{n} \tag{3.13}$$
$$= \frac{\tilde{r}^t \tilde{r}'}{n} \tag{3.14}$$
$$= \hat{\Sigma} \tag{3.15}$$

where in the last equality we have used the property required on $\tilde{r}'$.
*Using IPSO to get entirely synthetic microdata*
In Mateo-Sanz, Martínez-Ballesté and Domingo-Ferrer (2004), a non-iterative method for generating entirely synthetic continuous microdata through Cholesky decomposition is proposed. This can be viewed as a special case of IPSO. In a single step of computation, the method exactly reproduces the means and the covariance matrix of the original dataset. The running time grows linearly with the number of records. Exact preservation of the original covariance matrix implies that variances and Pearson correlations are also exactly preserved in the synthetic dataset.
The idea of the method is as follows. A dataset $X$ is viewed as a $n \times m$ matrix, where rows are records and columns are variables. First, the covariance matrix $C$ of $X$ is computed (covariance is defined between variables, *i.e.* between columns). Then, a random $n \times m$ matrix $A$ is generated, whose covariance matrix is the identity matrix. Next, the Cholesky decomposition of $C$ is computed, *i.e.*, an upper triangular matrix $U$ is found such that $C = U^t U$. Finally, the synthetic microdata set $Z$ is an $n \times m$ matrix $Z = AU$.

#### 3.4.7.6 Other partially synthetic microdata approaches

⚠ Expert level

The multiple imputation approach described in Rubin (1993) for creating entirely synthetic microdata can be extended for partially synthetic microdata. As a result, multiply-imputed, partially synthetic datasets are obtained that contain a mix of actual and imputed (synthetic) values. The idea is to multiply-impute confidential values and release non-confidential values without perturbation. This approach was first applied to protect the US Survey of Consumer Finances (Kennickell, 1999), (Kennickell, 1999b). In Abowd and Woodcock (2001) and Abowd and Woodcock (2004), this technique was adopted to protect longitudinal linked data, that is, microdata that contain observations from two or more related time periods (successive years, etc.). Methods for valid inference on this kind of partial synthetic data were developed in Reiter (2003) and a non-parametric method was presented in Reiter (2003b) to generate multiply-imputed, partially synthetic data.
Closely related to multiply imputed, partially synthetic microdata is model-based disclosure protection (Franconi and Stander, 2002), (Polettini, Franconi, and Stander, 2002). In this approach, a set of confidential continuous outcome variables is regressed on a disjoint set non-confidential variables; then the fitted values are released for the confidential variables instead of the original values.

#### 3.4.7.7 Muralidhar-Sarathy hybrid generator

⚠ Expert level

Hybrid data are a mixture of original data and synthetic data. Let $V$ an original data set whose attributes are numerical and fall into confidential attributes $X(= X_1 ... X_L)$ and non-confidential attributes $Y(= Y_1 ... Y_M)$. Let $V'$ be a hybrid data set obtained from $V$, whose attributes are $X(= X'_1 ... X'_L)$ (hybrid versions of $X$) and $Y$.
Muralidhar and Sarathy (2008) proposed a procedure (called MS in the sequel) for generating hybrid data as follows

$$X'_j = \gamma + X_j \alpha^t + Y_j \beta^t + e_i, \quad j = 1, ..., n$$

MS can yield hybrid data preserving the means and covariances of original data. To that end, the following equalities must be satisfied:

$$\beta^t = \Sigma_{YY}^{-1}\Sigma_{YX}(I - \alpha^t) \tag{3.16}$$

$$\gamma = (I - \alpha)\bar{X} - \beta\bar{Y} \tag{3.17}$$

$$\Sigma_{ee} = (\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}) - \alpha(\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX})\alpha^t \tag{3.18}$$

where $I$ is the identity matrix and $\Sigma_{ee}$ is the covariance matrix of the noise terms $e$. Thus, $\alpha$ completely specifies the procedure. The authors of MS admit that $\alpha$ must be selected carefully to ensure that $\Sigma_{ee}$ is positive semidefinite. They consider three options for specifying the $\alpha$ matrix:

1. Take $\alpha$ as a diagonal matrix with all values in the diagonal being equal. In this case, $\Sigma_{ee}$ is positive semidefinite and the value of the hybrid attribute $X'_i$ depends only on $X_i$, but not on $X_j$ for $j \neq i$. All confidential attributes $X_i$ are perturbed at the same level.
2. Take $\alpha$ as a diagonal matrix, with values in the diagonal being not all equal. In this case, $X'_i$ still depends only on $X_i$, but not on $X_j$ for $j \neq i$. The differences are that the confidential attributes are perturbed at different levels and there is no guarantee that $\Sigma_{ee}$ is positive semidefinite, so it may be necessary to try several values of $\alpha$ until positive semidefiniteness is achieved.
3. Taking $\alpha$ as a non-diagonal matrix does not guarantee positive semidefiniteness either and the authors of MS do not see any advantage in it, although it would be the only way to have $X'_i$ depend on several attributes among $(X_1 \ldots X_L)$. With ℝ-Microhybrid, the dependence of $X'_i$ on the original confidential attributes is the one provided by the underlying IPSO method.

#### 3.4.7.8 Microaggregation-based hybrid data

⚠ Expert level

In (Domingo-Ferrer and González-Nicolás, 2009) an alternative procedure to generate hybrid data based on microaggregation was proposed. Let $V$ be an original data set consisting of $n$ records. On input an integer parameter $k \in \{1, \ldots, n\}$, the procedure described in this section generates a hybrid data set $V'$. The greater $k$, the more synthetic is $V'$. Extreme cases are: i) $k = 1$, which yields $V' = V$ (the output data are exactly the original input data); and ii) $k = n$, which yields a completely synthetic output data set $V'$.
The procedure calls two algorithms:

- A generic synthetic data generator $S(C, C', \text{parms})$, that is, an algorithm which, given an original data (sub)set $C$, generates a synthetic data (sub)set $C'$ pre-

serving the statistics or parameters or models of $C$ specified in parms.
- A microaggregation heuristic, which, on input of a set of $n$ records and parameter $k$, partitions the set of records into clusters containing between $k$ and $2k-1$ records. Cluster creation attempts to maximize intra-cluster homogeneity.

**Procedure 1 (*Microhybrid* ($V$,$V'$, parms, $k$))**

1. Call microaggregation($V$, $k$). Let $C_1, \ldots, C_k$ for some $k$ be the resulting clusters of records.
2. For $i = 1, \ldots, k$ call $S(C_i, C'_i, \text{parms})$.
3. Output a hybrid dataset $V'$ whose records are those in the clusters $C'_1, \ldots, C'_k$ .

At Step 1 of procedure Microhybrid above, clusters containing between $k$ and $2k-1$ records are created. Then at Step 2, a synthetic version of each cluster is generated. At Step 3, the original records in each cluster are replaced by the records in the corresponding synthetic cluster (instead of replacing them with the average record of the cluster, as done in conventional microaggregation).
The Microhybrid procedure bears some resemblance to the condensation approach proposed by (Aggarwal and Yu, 2004); however, Microhybrid is more general because:

- It can be applied to any data type (condensation is designed for numerical data only);
- Clusters do not need to be all of size $k$ (their sizes can vary between $k$ and $2k-1$);
- Any synthetic data generator (chosen to preserve certain pre-selected statistics or models) can be used by Microhybrid;
- Instead of using an ad hoc clustering heuristic like condensation, Microhybrid can use any of the best microaggregation heuristics cited above, which should yield higher within-cluster homogeneity and thus less information loss.

*Role of parameter k*
We justify here the role of parameter $k$ in Microhybrid:

- If $k = 1$, and parms include preserving the mean of each attribute in the original clusters, the output is the same original data set, because the procedure creates $n$ clusters (as many as the number of original records). With $k = 1$, even variable-size heuristics will yield all clusters of size 1, because the maximum intra-cluster similarity is obtained when clusters consist all of a single record.
- If $k = n$, the output is a single synthetic cluster: the procedure is equivalent to calling the synthetic data generator $S$ once for the entire data set.
- For intermediate values of $k$, several clusters are obtained at Step 1, whose parameters parms are preserved by the synthetic clusters generated at Step 2. As $k$ decreases, the number of clusters (whose parameters are preserved in the

data output at Step 3) increases, which causes the output data to look more and more like the original data. Each cluster can be regarded as a constraint on the synthetic data generation: the more constraints, the less freedom there is for generating synthetic data, and the output resembles more the original data. This is why the output data can be called hybrid.

It must be noted here that, depending on the synthetic generator used, there may be a lower bound for $k$ higher than 1. For example, if using IPSO (see Section 3.4.7.5) with $|X|$ confidential attributes and $|Y|$ non-confidential attributes, it turns out that $k$ must be at least $2|X| + |Y| + 1$; otherwise there are not enough degrees of freedom for the generator to work.
Note that the choice of parameter $k$ is more straightforward than the choice of $\alpha$ in the MS procedure above. Also, for the case of numerical microdata, Microhybrid can offer, in addition to mean and covariance exact preservation, approximate preservation of third-order and fourth-order moments, and also approximate preservation of all moments up to order four in randomly chosen subdomains of the dataset. Details are given in the above-referenced paper describing Microhybrid.

#### 3.4.7.9 Other hybrid microdata approaches

Expert level

A different approach called hybrid masking was proposed in Dandekar, Domingo-Ferrer and Sebé (2002). The idea is to compute masked data as a combination of original and synthetic data. Such a combination allows better control than purely synthetic data over the individual characteristics of masked records. For hybrid masking to be feasible, a rule must be used to pair one original data record with one synthetic data record. An option suggested in Dandekar, Domingo-Ferrer and Sebé (2002) is to go through all original data records and pair each original record with the nearest synthetic record according to some distance. Once records have been paired, Dandekar, Domingo-Ferrer, and Sebé (2002) suggest two possible ways for combining one original record $X$ with one synthetic record $X_S$: additive combination and multiplicative combination. Additive combination yields

$$Z = \alpha X + (1 - \alpha) X_S$$

and multiplicative combination yields

$$Z = X^{\alpha} \cdot X_s^{(1-\alpha)}$$

where $\alpha$ is an input parameter in $[0, 1]$ and $Z$ is the hybrid record. The authors present empirical results comparing the hybrid approach with rank swapping and

microaggregation masking (the synthetic component of hybrid data is generated using Latin Hypercube Sampling by Dandekar, Cohen, and Kirkendall, 2002).
Post-masking optimization is another approach to combining original and synthetic microdata is proposed in Sebé *et al.* (2002). The idea here is to first mask an original dataset using a masking method. Then a hill-climbing optimization heuristic is run which seeks to modify the masked data to preserve the first and second-order moments of the original dataset as much as possible without increasing the disclosure risk with respect to the initial masked data. The optimization heuristic can be modified to preserve higher-order moments, but this significantly increases computation. Also, the optimization heuristic can use take as initial dataset a random dataset instead of a masked dataset; in this case, the output dataset is purely synthetic.

#### 3.4.7.10 Pros and cons of synthetic and hybrid microdata

Synthetic data are appealing in that, at a first glance, they seem to circumvent the re-identification problem: since published records are invented and do not derive from any original record, it might be concluded that no individual can complain from having been re-identified. At a closer look this advantage is less clear. If, by chance, a published synthetic record matches a particular citizen's non-confidential variables (age, marital status, place of residence, etc.) and confidential variables (salary, mortgage, etc.), re-identification using the non-confidential variables is easy and that citizen may feel that his confidential variables have been unduly revealed. In that case, the citizen is unlikely to be happy with or even understand the explanation that the record was synthetically generated.

On the other hand, limited data utility is another problem of synthetic data. Only the statistical properties explicitly captured by the model used by the data protector are preserved. A logical question at this point is why not directly publish the statistics one wants to preserve rather than release a synthetic microdata set.

One possible justification for synthetic microdata would be if valid analyses could be obtained on a number of subdomains, *i.e.* similar results were obtained in a number of subsets of the original dataset and the corresponding subsets of the synthetic dataset. Partially synthetic or hybrid microdata are more likely to succeed in staying useful for subdomain analysis. However, when using partially synthetic or hybrid microdata, we lose the attractive feature of purely synthetic data that the number of records in the protected (synthetic) dataset is independent from the number of records in the original dataset.

#### 3.4.7.11 References

Abowd, J. M., and Woodcock, S. D. (2001). *Disclosure limitation in longitudinal linked tables.* In P. Doyle, J. I. Lane, J. J. Theeuwes, and L. V. Zayatz, editors, Confidentiality,

Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pages 215–278, Amsterdam, 2001. North-Holland.

Abowd, J. M. and Woodcock, S. D. (2004). *Multiply-imputing confidential characteristics and file links in longitudinal linked data.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of LNCS, pages 290–297, Berlin Heidelberg, 2004. Springer.

Aggarwal, C. C., and Yu, P. S. (2004). *A condensation approach to privacy preserving data mining.* In E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, and E. Ferrari, editors, Advances in Database Technology - EDBT 2004, volume 2992 of Lecture Notes in Computer Science, pages 183–199, Berlin Heidelberg, 2004.

Burridge, J. (2004). *Information preserving statistical obfuscation.* Statistics and Computing, 13:321–327, 2003.

Crystal.Ball. http://www.cbpro.com/.

Dandekar, R., Cohen, M., and Kirkendall, N. (2002). *Sensitive micro data protection using latin hypercube sampling technique.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 245–253, Berlin Heidelberg, Springer.

Dandekar, R., Domingo-Ferrer, J., and Sebé, F. (2002). *LHS-based hybrid microdata vs. rank swapping and microaggregation for numeric microdata protection.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 153–162, Berlin Heidelberg. Springer.

Domingo-Ferrer, J., and González-Nicolás, Ú. (2009). *Hybrid Microdata Using Microaggregation.* Manuscript.

Fienberg, S. E. (1994). *A radical proposal for the provision of micro-data samples and the preservation of confidentiality.* Technical Report 611, Carnegie Mellon University Department of Statistics.

Fienberg, S.E., Makov, U. E., and Steele, R. J. (1998). *Disclosure limitation using perturbation and related methods for categorical data.* Journal of Official Statistics, 14(4):485–502.

Florian, A. (1992). *An efficient sampling scheme: updated latin hypercube sampling.* Probabilistic Engineering Mechanics, 7(2):123–130.

Franconi, L., and Stander, J. (2002). *A model based method for disclosure limitation of business microdata.* Journal of the Royal Statistical Society D - Statistician, 51:1–11.

Huntington, D. E., and Lyrintzis, C. S. (1998). *Improvements to and limitations of latin hypercube sampling.* Probabilistic Engineering Mechanics, 13(4):245–253.

Kennickell, A. B. (1999). *Multiple imputation and disclosure control: the case of the 1995 survey of consumer finances.* In Record Linkage Techniques, pages 248–267, Washington DC, 1999. National Academy Press.

Kennickell, A. B. (1999b). *Multiple imputation and disclosure protection: the case of the 1995 survey of consumer finances.* In J. Domingo-Ferrer, editor, Statistical Data Protection, pages 248–267, Luxemburg, 1999. Office for Official Publications of the European Communities.

Liew, C. K., Choi, U. J., and Liew, C. J. (1985). *A data distortion by probability distribution.* ACM Transactions on Database Systems, 10:395–411, 1985.

Mateo-Sanz, J. M., Martínez-Ballesté, A., and Domingo-Ferrer, J. (2004). *Fast generation of accurate synthetic microdata.* In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of LNCS, pages 298–306, Berlin Heidelberg, Springer.

Muralidhar, K, and Sarathy, R, (2008). *Generating sufficiency-based nonsynthetic perturbed data.* Transactions on Data Privacy, 1(1):17–33, 2008. http://www.tdp.cat/issues/tdp.a005a08.pdf.

Polettini, S., Franconi, L., and Stander, J. (2002). *Model based disclosure protection.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 83–96, Berlin Heidelberg. Springer.

Raghunathan, T. J., Reiter, J. P., and Rubin, D. (2003). *Multiple imputation for statistical disclosure limitation.* Journal of Official Statistics, 19(1):1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–188.

Reiter, J. P. (2003b). *Using CART to generate partially synthetic public use microdata, 2003.* Duke University working paper.

Reiter, J. P. (2005). *Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study.* Journal of the Royal Statistical Society, Series A, 168:185–205.

Reiter, J. P. (2005b). *Significance tests for multi-component estimands from multiply-imputed, synthetic microdata.* Journal of Statistical Planning and Inference, 131(2):365–377.

Rubin, D. E. (1993). *Discussion of statistical disclosure limitation.* Journal of Official Statistics, 9(2):461–468.

Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J. M. and Torra, V. (2002). *Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata*

*sets.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 163–171, Berlin Heidelberg, Springer.

## 3.5 Measurement of information loss

### 3.5.1 Concepts and types of information loss and its measures

The application of SDC methods entails the loss of some information. It arises as a result e.g. from gaps occurring in data when non-perturbative SDC methods are used, or perturbations when perturbative SDC tools are used. Because of this loss the analytical worth of the disclosed data for the user decreases, which means there is a possibility that results of computations and analyses based on such data will be inadequate (e.g. the precision of estimation could be much worse).

A strict evaluation of information loss must be based on the data uses to be supported by the protected data. The greater the differences between the results obtained on original and protected data for those uses, the higher the loss of information. However, very often microdata protection cannot be performed in a data use specific manner, for the following reasons:

- Potential data uses are very diverse and it may be even hard to identify them all at the moment of data release by the data protector.
- Even if all data uses can be identified, issuing several versions of the same original dataset so that the $i$-th version has an information loss optimized for the $i$-th data use may result in unexpected disclosure by combining the differently protected datasets.

Since that data often must be protected with no specific data use in mind, generic information loss measures are desirable to guide the data protector in assessing how much harm is being inflicted to the data by a particular SDC technique.

Defining what a generic information loss measure is can be a tricky issue. Roughly speaking, it should capture the amount of information loss for a reasonable range of data uses. We will say there is little information loss if the protected dataset is analytically valid and interesting according to the following definitions by Winkler (1998):

- A protected microdata set is an *analytically valid* microdata set if it approximately preserves the following with respect to the original data (some conditions apply only to continuous variables):
  - Means and covariances on a small set of subdomains (subsets of records and/or variables)
  - Marginal values for a few tabulations of the data (the information loss in this approach concerns mainly tables created on the basis of microdata and therefore it will be discussed in Chapter 4 and Chapter 5)

  - At least one distributional characteristic

- A microdata set is an *analytically interesting microdata set*, if six variables on important subdomains are provided that can be validly analyzed.
  More precise conditions of analytical validity and analytical interest cannot be stated without taking specific data uses into account. As imprecise as they may be, the above definitions suggest some possible measures:

  - Compare raw records in the original and the protected dataset. The more similar the SDC method to the identity function, the less the impact (but the higher the disclosure risk! ). This requires pairing records in the original dataset and records in the protected dataset. For masking methods, each record in the protected dataset is naturally paired to the record in the original dataset it originates from. For synthetic protected datasets, pairing is more artificial. Dandekar, Domingo-Ferrer and Sebé (2002) proposed to pair a synthetic record to the nearest original record according to some distance.
  - Compare some statistics computed on the original and the protected datasets. The above definitions list some statistics which should be preserved as much as possible by an SDC method.

Taking the aforementioned premises into account, for microdata the information loss can concern the differences in distributions, in diversification and in shape and power of connections between various features. Therefore, the following types of measures of information loss are distinguished:

1. Measures of distribution disturbance – measures based on distances between original and perturbed values of variables (e.g. mean, mean of relative distances, complex distances, etc.),
2. Measures of impact on variance of estimation – computed using distances between variances for averages of continuous variables before and after SDC or multi-factor ANOVA for a selected dependent variable in relation to selected independent categorical variables (in this case, the measure of information loss involves a comparison of components of coefficients of determination $R^2$ - in terms of within-group and inter-group variance - for relevant models based on original and perturbed values (cf. Hundepool et al. (2012))),
3. Measures of impact on the intensity of connections – comparisons of measures of direction and intensity of connections between original continuous variables and between relevant perturbed ones; such measures can be e.g. correlation coefficients or test of independence.

### 3.5.2 Information loss measures for categorical data

Straightforward computation of measures based on basic arithmetic operations like addition, subtraction, multiplication and division on categorical data is not possible. Neither

is the use of most descriptive statistics like Euclidean distance, mean variance, correlation, etc. The following alternatives are considered in Domingo-Ferrer and Torra (2001):

- Direct comparison of categorical values
- Comparison of contingency tables
- Entropy-based measures

Below we will describe examples for each of such types of measures.

#### 3.5.2.1 Direct comparison of categorical values

⚠ Expert level

Comparison of matrices $X$ and $X'$ for categorical data requires the definition of a distance for categorical variables. Definitions consider only the distances between pairs of categories that can appear when comparing an original record and its protected version (see discussion above on pairing original and protected records).
For a nominal variable $V$ (a categorical variable taking values over an unordered set), the only permitted operation is comparison for equality. This leads to the following distance definition:

$$d_V(c, c') = \begin{cases} 0, & \text{if } c = c' \\ 1, & \text{if } c \neq c' \end{cases}$$

where $c$ is a category in an original record and $c'$ is the category which has replaced $c$ in the corresponding protected record.
For an ordinal variable $V$ (a categorical variable taking values over a totally ordered set), let $\leq V$ be the total order operator over the range $D(V)$ of $V$. Define the distance between categories $c$ and $c'$ as the number of categories between the minimum and the maximum of $c$ and $c'$ divided by the cardinality of the range:

$$\mathrm{dc}\,(c, c') = \frac{|c'' : \min\,(c, c') \leq c'' < \max\,(c, c')|}{|D(V)|} \tag{3.19}$$

#### 3.5.2.2 Comparison of contingency tables

⚠ Expert level

An alternative to directly comparing the values of categorical variables is to compare their contingency tables. Given two datasets $F$ and $G$ (the original and the protected set, respectively) and their corresponding $t$-dimensional contingency tables for $t \leq K$, we can define a contingency table-based information loss measure $CTBIL$ for a subset

$W$ of variables as follows:

$$CTBIL(F,G;W,K) = \sum_{\substack{\{V_{ji}\cdots V_{jt}\}f\subseteq W \\ |\{V_{ji}\cdots V_{jt}\}|\leq K}} \sum_{i_1\cdots i_t} |x^F_{i_1\cdots i_t} - x^G_{i_1\cdots i_t}| \tag{3.20}$$

where $x^{\text{file}}_{\text{subscripts}}$ is the entry of the contingency table of file at position given by subscripts.
Because the number of contingency tables to be considered depends on the number of variables $|W|$, the number of categories for each variable, and the dimension $K$, a normalized version of (3.20) may be desirable. This can be obtained by dividing expression (3.20) by the total number of cells in all considered tables.
Distance between contingency tables generalizes some of the information loss measures used in the literature. For example, the $\mu$-ARGUS software (Hundepool et al., 2005) measures information loss for local suppression by counting the number of suppressions. The distance between two contingency tables of dimension one returns twice the number of suppressions. This is because, when category $A$ is suppressed for one record, two entries of the contingency table are changed: the count of records with category $A$ decreases and the count of records with the "missing" category increases.

#### 3.5.2.3 Entropy-based measures

⚠ Expert level

In De Waal and Willenborg (1999), Kooiman, Willenborg and Gouweleeuw (1998) and Willenborg and De Waal (2001), the use of Shannon's entropy to measure information loss is discussed for the following methods: local suppression, global recoding and PRAM. Entropy is an information-theoretic measure, but can be used in SDC if the protection process is modelled as the noise that would be added to the original dataset in the event of it being transmitted over a noisy channel.
As noted earlier, PRAM is a method that generalizes noise addition, suppression and recoding methods. Therefore, our description of the use of entropy will be limited to PRAM.
Let $V$ be a variable in the original dataset and $V'$ be the corresponding variable in the PRAM-protected dataset. Let $\mathbf{P}_{V,V'} = \{\mathbb{P}(V' = j \mid V = i)\}$ be the PRAM Markov matrix. Then, the conditional uncertainty of $V$ given that $V' = j$ is:

$$H(V \mid V' = j) = -\sum_{i=1}^{n} \mathbb{P}(V = i \mid V' = j) \log \mathbb{P}(V = i \mid V' = j) \tag{3.21}$$

The probabilities in (3.21) can be derived from $\mathbf{P}_{V,V'}$ using Bayes' formula. Finally,

the entropy-based information loss measure $EBIL$ is obtained by accumulating expression (3.21) for all individuals $r$ in the protected dataset $G$

$$EBIL\left(\mathbf{P}_{V,V'}, G\right) = \sum_{r \in G} H\left(V \mid V' = j_r\right)$$

where $j_r$ is the value taken by $V'$ in record $r$.
The above measure can be generalized for multivariate datasets if $V$ and $V'$ are taken as being multidimensional variables (*i.e.* representing several one-dimensional variables).
While using entropy to measure information loss is attractive from a theoretical point of view, its interpretation in terms of data utility loss is less obvious than for the previously discussed measures.

### 3.5.3 Information loss measures for continuous data

Assume a microdata set with $n$ individuals (records) $I_1, I_2, \cdots, I_n$ and $p$ continuous variables $Z_1, Z_2, \cdots, Z_p$. Let $X$ be the matrix representing the original microdata set (rows are records and columns are variables). Let $X^{'}$ be the matrix representing the protected microdata set. The following tools are useful to characterize the information contained in the dataset:

- Covariance matrices $V$ (on $X$) and $V^{'}$ (on $X^{'}$).
- Correlation matrices $R$ and $R^{'}$.
- Correlation matrices $RF$ and $RF^{'}$ between the $p$ variables and the $p$ factors principal components $PC_1, PC_2, \cdots, PC_p$ obtained through principal components analysis.
- Communality between each of the $p$ variables and the first principal component $PC_1$ (or other principal components $PC_i$'s). Communality is the percent of each variable that is explained by $PC_1$ (or $PC_i$). Let $C$ be the vector of communalities for $X$ and $C^{'}$ the corresponding vector for $X^{'}$.
- Matrices $F$ and $F^{'}$ containing the loadings of each variable in $X$ on each principal component. The $i$-th variable in $X$ can be expressed as a linear combination of the principal components plus a residual variation, where the $j$-th principal component is multiplied by the loading in $F$ relating the $i$-th variable and the $j$-th principal component (Chatfield and Collins, 1980). $F^{'}$ is the corresponding matrix for $X^{'}$.

There does not seem to be a single quantitative measure which completely reflects those structural differences. Therefore, we proposed in Domingo-Ferrer, Mateo-Sanz, and Torra (2001) and Domingo-Ferrer and Torra (2001) to measure information loss through the discrepancies between matrices $X$, $V$, $R$, $RF$, $C$ and $F$ obtained on the original data and the corresponding $X^{'}$, $V^{'}$, $R^{'}$, $RF^{'}$, $C^{'}$ and $F^{'}$ obtained on the protected dataset.

In particular, discrepancy between correlations is related to the information loss for data uses such as regressions and cross tabulations.

Matrix discrepancy can be measured in at least three ways:

**Mean square error** Sum of squared componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

**Mean absolute error** Sum of absolute componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

**Mean variation** Sum of absolute percent variation of components in the matrix computed on protected data with respect to components in the matrix computed on original data, divided by the number of cells in either matrix. This approach has the advantage of not being affected by scale changes of variables.

Table 3.9 summarizes the measures proposed in Domingo-Ferrer, Mateo-Sanz and Torra (2001) and Domingo-Ferrer and V. Torra (2001). In this table, $p$ is the number of variables, $n$ the number of records, and components of matrices are represented by the corresponding lowercase letters (*e.g.* $x_{\mathrm{ij}}$ is a component of matrix $X$). Regarding $X - X^{'}$ measures, it makes also sense to compute those on the averages of variables rather than on all data (call this variant $\overline{X} - \overline{X'}$). Similarly, for $V - V^{'}$ measures, it would also be sensible to use them to compare only the variances of the variables, *i.e.* to compare the diagonals of the covariance matrices rather than the whole matrices (call this variant $S - S^{'}$).

Table 3.9: Information loss measures for continuous microdata. Source: Domingo-Ferrer, Mateo-Sanz and Torra (2001).

| | Mean square error | Mean abs. error | Mean variation |
|---|---|---|---|
| $X - X'$ | $\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$ | $\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$ | $\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$ |
| $V - V'$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{p(p+1)/2}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{p(p+1)/2}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{p(p+1)/2}$ |
| $R - R'$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{p(p-1)/2}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{p(p-1)/2}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{p(p-1)/2}$ |
| $RF - RF'$ | $\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$ | $\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$ | $\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$ |
| $C - C'$ | $\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$ | $\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$ | $\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$ |

| | Mean square error | Mean abs. error | Mean variation |
|---|---|---|---|
| $F - F'$ | $\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p} (f_{ij} - f'_{ij})^2}{p^2}$ | $\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p} \lvert f_{ij} - f'_{ij} \rvert}{p^2}$ | $\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p} \frac{\lvert f_{ij} - f'_{ij} \rvert}{\lvert f_{ij} \rvert}}{p^2}$ |

In Yancey, Winkler and Creecy (2002), it is observed that dividing by $x_{\mathrm{ij}}$ causes the $X - X^{'}$ mean variation to rise sharply when the original value $x_{\mathrm{ij}}$ is close to 0. This dependency on the particular original value being undesirable in an information loss measure, Yancey, Winkler and Creecy (2002) propose to replace the mean variation of $X - X^{'}$ by the more stable measure IL1 given by

$$\frac{1}{np} \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{|x_{ij} - x'_{ij}|}{\sqrt{2} S_j}$$

where $S_j$ is the standard deviation of the $j$-th variable in the original dataset. This measure was incorporated into the sdcMicro R package. The IL1 measure, in turn, is highly sensitive to small disturbances and weak differentiation of feature values - it may take too high values for variables with low differentiation, and too low - when the differentiation is significant. In practice, if $S_j$ is very close to zero, we obtain as a results INF (infinity). In this case, the measure becomes really useless, because it will not allow to compare the loss of information in several microdata sets with statistical confidentiality protected in various ways - if for each of such sets the IL1 measure will be equal to INF.

Trottini (2003) argues that, since information loss is to be traded off for disclosure risk and the latter is bounded —there is no risk higher than 100%—, upper bounds should be enforced for information loss measures. In practice, the proposal in Trottini (2003) is to limit those measures in Table 3.9 based on the mean variation to a predefined maximum value.

Młodak (2020) proposed a new measure of information loss for continuous variables in terms of assesment of impact on the intensity of connections, which was slighthly improved by Młodak, Pietrzak and Józefowski (2022). It is based on on diagonal entries of inversed correlation matrices for continuous variables in the original ($R^{-1}$) and perturbed ($R'^{-1}$) data sets, i.e. $\rho_{jj}^{(-1)}$ and $\rho'_{jj}{}^{(-1)}$, $j = 1, 2, \ldots, m_c$ (where $m_c$ is the number of continuous variables):

$$\gamma = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{m_c} \left( \frac{\rho_{jj}^{(-1)}}{\sqrt{\sum_{l=1}^{m} \left(\rho_{ll}^{(-1)}\right)^2}} - \frac{\rho'_{jj}{}^{(-1)}}{\sqrt{\sum_{l=1}^{m} \left(\rho'_{ll}{}^{(-1)}\right)^2}} \right)^2} \in [0, 1]. \tag{3.22}$$

Values of (3.22) are also easily interpretable - it can be understood as the expected loss of information about connections between variables. As one can easily see, the result can

be expressed in %. Of course, both matrices - $R$ and $R'$ - must be based on the same correlation coefficient. The most obvious choice in this respect is the Pearson's index. However, when tau-Kendall correlation matrix is used, one can also apply it to ordinal variables. The method will be not applicable if the correlation matrix is singular. The main advantage of the measure $\gamma$ is that it treats all variables as an inseparable whole and takes all connections between analysed variables, even those hard to observe, into account. $\gamma$ can be computed in the sdcMicro R package using the function `IL_correl()`.

### 3.5.4 Complex measures of information loss

⚠ Expert level

The above presented concepts of information loss prompt the question whether it is possible to construct complex measure of information loss taking variables of all measurement scales into account. The relevant proposal was formulated by Młodak (2020) and applied by Młodak, Pietrzak and Józefowski (2022) to the case of microdata from the Polish survey of accidents at work. For categorical variables it is based on the approaches 3.5.2.1 and 3.5.2.2 , i.e. if the variable $X_j$ is nominal, then (treating NA as a separate level)

$$d(x'_{ij}, x_{ij}) = \begin{cases} 1 & \text{if } x'_{ij} = x_{ij}, \\ 0 & \text{if } x'_{ij} \neq x_{ij}. \end{cases} \tag{3.23}$$

If $X_j$ is ordinal (assuming for simplification and without loss of generality that categories are numbered from 1 to $\mathfrak{r}_j$, where $\mathfrak{r}_j$ is the number of categories), then (NA is treated as a separate, lowest category)

$$d(x'_{ij}, x_{ij}) = \frac{\mathfrak{r}(x'_{ij}, x_{ij})}{\mathfrak{r}_j - 1}, \tag{3.24}$$

where $\mathfrak{r}(x'_{ij}, x_{ij})$ is the absolute difference in categories between $x'_{ij}$ and $x_{ij}$. These partial distances take always values from [0,1]. There are, however, some problems with using them, especially if recoding is applied. The number of categories of a recoded variable in the original set and in the set after SDC will be different. Therefore, in the first place, it should be ensured that the numbers of the categories left unchanged are identical in both variants. For example, if before recoding the variable $X_j$ had $\mathfrak{r}_j$=8 categories marked as 1,2,3,4,5,6,7,8 and as a result of recoding categories 2 and 3 and 6 and 7 were combined, then the new categories should have respectively numbers 1,2,4,5,6,8. Then the option (3.24) for categorical variables applies in this case as well.
Much more complicated situation occurs for continuous variables. Młodak (2020) proposed several options is this respect, e.g. normalized absolute value or normalized

square of difference between ${x_{ij}}'$ and $x_{ij}$, i.e.

$$d(x'_{ij}, x_{ij}) = |x'_{ij} - x_{ij}| / \max_{k=1,2,\dots,n} |x'_{kj} - x_{kj}|, \tag{3.25}$$

or

$$d(x'_{ij}, x_{ij}) = (x'_{ij} - x_{ij})^2 / \max_{k=1,2,\dots,n} (x'_{kj} - x_{kj})^2, \tag{3.26}$$

$i = 1, 2, \dots, n$, $j = 1, 2, \dots, m_c$, where $n$ is the number of records and $m_c$ - the number of continuous variables.
Measures (3.23) and (3.24) also have another significant weakness. The measure of information loss should be an increasing function due to individual partial information losses. This means that, for example, if for some $i \in \{1, 2, \dots, n\}$ the value $|x'_{ij} - x_{ij}|$ will increase and all $|x'_{hj} - x_{hj}|$ for $h \neq i$ remain the same, the value of the distance should increase. Meanwhile, in the case of formulas (3.25) and (3.26), this will not be the case. If, for the same, the indicated absolute difference (or the square of the difference, respectively) between the original value and the value after SDC reaches a maximum, then the partial loss of information for $i$ will remain unchanged - it will be 1, and for the others it will turn out to be smaller. As a result, we get a smaller metric value, while the information loss actually increased.
Taking the aforementioned observations into account Młodak (2020) proposed in the discussed case the distance of the form:

$$d(x'_{ij}, x_{ij}) = \frac{2}{\pi} \arctan |x'_{ij} - x_{ij}|. \tag{3.27}$$

The arcus tangens (arctan) function was used to ensure that the distance between original and perturbed values takes values from $[0, 1]$. To achieve this, an ascending function bounded on both sides (both from the top and from the bottom) should be applied. The arctan seems to be a good solution and is also easy to compute. Of course – like any function of this type – it is not perfect: for larger absolute differences between original and perturbed values it tends to be close to $\frac{\pi}{2}$ (and, in consequence, $d(x'_{ij}, x_{ij})$ to be close to 1). On the other hand, owing to this property it exhibits more clearly all information losses due to perturbation.
The complex measure of distribution disturbance is given by (cf. Młodak, Pietrzak and Józefowski (2022)):

$$\lambda = \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{d(x'_{ij}, x_{ij})}{mn} \in [0, 1], \tag{3.28}$$

where $d(\cdot, \cdot) \in [0, 1]$ is measure of distance according to the formulas (3.23), (3.24) or (3.27) according to the measurement scale of a given value.
Authors of the aforementioned paper indicated also than one can measure the contribution of particular variables $X_j$ to total information loss as follows

$$\lambda_j = \sum_{i=1}^{n} \frac{d(x'_{ij}, x_{ij})}{n} \in [0,1], \tag{3.29}$$

$j = 1, 2, \ldots, m$.
An additional problem occurs if non-perturbative SDC tools are used. In this case the original values are either suppressed or remained unchanged. How to proceed in this case during computation of the measures (3.26 and (3.27) also depends on the measurement scale of the variables. If the used $X_j$ is nominal, then if $x'_{ij}$ is hidden then one should assume $d(x'_{ij}, x_{ij}) = 1$; if $X_j$ is ordinal, then we assign $x'_{ij} := 1$ if $x_{ij}$ is closer to $\mathfrak{r}_j$ or $x'_{ij} := \mathfrak{r}_j$ if $X_j$ is closer to 1; if $X_j$ is continuous, then

$$x'_{ij} := \begin{cases} \max\limits_{h=1,2,\ldots,n} x_{hj} & \text{if} \quad x_{ij} \leq \operatorname{med}\limits_{h=1,2,\ldots,n} x_{hj}, \\ \min\limits_{h=1,2,\ldots,n} x_{hj} & \text{if} \quad x_{ij} > \operatorname{med}\limits_{h=1,2,\ldots,n} x_{hj}. \end{cases}$$

The measures (3.28) and (3.29) can be expressed as a percentages and show total information loss and contribution of particular variables to it, respectively. The greater the value of $\lambda/\lambda_j$, the bigger the loss/contribution. In this way users obtain clear and easily understandable information about expected information loss owing to the application of SDC. These measures were implemented to the sdcMicro R package and are computed by the function `IL_variables`.

### 3.5.5 Practical realization of trade-off between safety and utility of microdata

Achieving the optimal balance between minimization of dislosure risk and minimization of the information loss is not easy. It is very hard (if even possible) to take all aspects deciding on level of these quantities (especially in the case of risk) into account. Moreover, both risk and information loss can be assessed from various point of views. Thus, first one should establish the possible factors which may decide on the type and level of dislosure risk and the most preferred direction of data use by the user. In the case of risk, one should assess not only internal risk (including different types of variables and their relationships) but also assess what alternative data sources the interested data user could have access to due to his place of employment and position held (such information is usually provided in official data access request). The priorities in measurement of information loss preferred by the user should be a basis for establishment of used measure in this context. For instance, if the users prefers comparison of distributions of some phenomena, then the measures of distribution disturbance should have much higher priority than others. On the other hand, if the subject of interest of an user are connections between some features, then for categorical variables the information loss should be assessed using the measures for contingency tables (as they are in fact frequency tables, this problem is discussed in Chapter 5). For continuous variables the aforementioned measures of impact on the intensity of connections can be, of course, applied.

Similarly as e.g. in the case of significance and loss in testing of statistical hypotheses, the most obvious and easy approach to obtain reasonable compromise between these two expectations is to apply one of two following ways:

- establishing arbitrarily maximum allowable level of disclosure risk and minimize the information loss in this situation - it defends, first of all, the data confidentiality and trust to data holder in terms of privacy protection,
- establishing arbitrarily maximum allowable level of information loss and minimize the disclosure risk in this situation - it defends, first of all, the data utility for users and data provider as a source of reliable, creadible and useful data.

In practice, the data holder (e.g. official statistics) prefers rather the first approach as the strict protection of data privacy is usually an obligation imposed by valid law regulations. So, assurance of the safety of confidential information is very important.

### 3.5.6 Example

The manner of assessing disclosure risk and information loss owing to the application of SDC methods was demonstrated using data from a case study published on the website of International Household Survey Network (IHSN)[2] Statistical Disclosure Control for Microdata: A Practice Guide - Case Study Data and R Script, being a supplement to the book by Benschop, Machingauta and Welch (2022). Use was made of part of the code from the first study of this type, in which the authors applied SDC measures to a set of farms using the sdcMicro package.

The following categorical variables were selected as key variables: REGION, URBRUR (area of residence), HHSIZE (household size), OWNAGLAND (agricultural land ownership), RELIG (religion of household head). The authors of the case study applied local data suppression to these variables.

SDC was also applied to quantitative variables concerning 1) expenditure: TFOODEXP (total food expenditure), TALCHEXP (total alcohol expenditure), TCLTHEXP (total expenditure on clothing and footwear), THOUSEXP (total expenditure on housing), TFURNEXP (total expenditure on furnishing ), THLTHEXP (total expenditure on health), TTRANSEXP (total expenditure on transport), TCOMMEXP (total expenditure on communications), TRECEXP (total expenditure on recreation), TEDUEXP (total expenditure on education), TRESTHOTEXP (total expenditure on restaurants and hotel ), TMISCEXP (total miscellaneous expenditure); 2) income: INCTOTGROSSHH (total gross household income – annual), INCRMT (total amount of remittances received from remittance sending members), INCWAGE (wage and salaries – annual), INCFARMBSN (gross income from household farm businesses – annual), INCNFARMBSN (gross income from household non-farm businesses – annual), INCRENT (rental income – annual), INCFIN (financial income from savings, loans, tax refunds, maturity payments on insurance),

[2] http://www.ihsn.org/software/disclosure-control-toolbox

INCPENSN (pension and other social assistance – annual), INCOTHER (other income – annual), and 3) land size: LANDSIZEHA (land size owned by household in ha). 1% noise was added to the variables relating to all components of expenditure and income; 5% noise was added to outliers. Values of the LANDSIZEHA variable were rounded (1 digit for plots smaller than 1 and to no digits for plots larger than 1) and grouped (values in intervals 5-19 to 13, and values in intervals 20-39 to 30, values larger than 40 to 40).

In the case study, the PRAM method was applied to variables describing apartment equipment: ROOF (roof type), WATER (main source of water), TOILET (main toilet facility), ELECTCON (electricity), FUELCOOK (main cooking fuel), OWNMOTORCYCLE (ownership of motorcycle), CAR (ownership of car), TV (ownership of television), LIVESTOCK (number of large-sized livestock owned). The data were stratified by REGION variable making sure that variants of the transformed variables were not modified in 80% of cases.

The set of data anonymised in the manner described above was used as the starting point for the assessment of the risk of disclosure and information loss. Tables Table 3.10 and Table 3.11 shows descriptive statistics for the risk of disclosure in the case of key variables before and after applying local suppression. While the risk was significantly reduced, one must bear in mind that the risk of disclosure was already relatively low in the original dataset. The maximum value of individual risk dropped from 5.5% in the original dataset to 0.3% after applying local suppression. The global risk in the original set was on average equal to 0.05%, which means that the expected number of disclosed units was 0.99; after applying local suppression, the global risk dropped to less than 0.02, which means that the expected number of disclosed units was 0.35.

As regards the assessment of disclosure risk for quantitative variables, an interval of [0.0%, 83.5%] was chosen, where the upper limit represents the worst case scenario in which the intruder is sure that each nearest neighbour is in fact the correct linkage.

Several of the measures mentioned above have been developed to assess the loss of information. Based on distances between values of variables that were to be anonymised in the original set and the their values in the anonymised set, $\lambda$ measures were calculated. Table Table 3.12 shows the general value of $\lambda$ and its values for individual variables ($\lambda_k$). The overall loss of information for the anonymised variables is 14.3%. The greatest loss is observed for quantitative variables to which noise was added; in the case of INCTOTGROSSHH, the loss of information measured by $\lambda$ reaches 83.4%. The loss of information was much lower in the case of key variables subjected to local suppression and those modified with the PRAM method: the maximum loss was 9.7% and 9.4%, respectively.

Overall information loss was determined using two measures described above: $IL1$ and $\lambda$. $IL1$ was equal to 79.4, which indicates relatively large standard deviations of anonymised values of quantitative variables from standard deviations of the original variables. The value of the second measure, which is based on correlation coefficients, is 0.6%, which indicates a slight loss of information regarding correlations between the quantitative variables. Nevertheless, it should be stressed that as a result of to numerous cases of non-response

in the quantitative variables, the value of $\lambda$ was calculated on the basis of only 111 observations, i.e. less than 6% of all units.

The above assessment was conducted using the R sdcMicro package. Because some of the information loss measures described above are not implemented in this package, they were not used in the assessment.

Table 3.10: Descriptive statistics of individual risk measures for quantitative variables

| **Statistic** | **Original values** | **Values after anonymisation** |
|---|---|---|
| Min | 0.0007 | 0.0007 |
| Q1 | 0.0021 | 0.0021 |
| Me | 0.0067 | 0.0059 |
| Q3 | 0.0213 | 0.0161 |
| Max | 5.5434 | 0.3225 |
| Mean | 0.0502 | 0.0176 |

Table 3.11: Global risk measures for quantitative variables

| **Statistic** | **Original values** | **Values after anonymisation** |
|---|---|---|
| Risk % | 0.0502 | 0.0176 |
| Expected number of disclosures | 0.9895 | 0.3476 |

Table 3.12: Loss of information due to anonymisation, overall and for individual variables

| **Variable** | $\lambda$ (%) |
|---|---|
| **OVERALL** | **14.3** |
| URBRUR | 0.5 |
| REGION | 0.2 |
| OWNAGLAND | 2.5 |
| RELIG | 1.1 |
| LANDSIZEHA | 9.7 |
| TANHHEXP | 50.7 |
| TFOODEXP | 38.3 |
| TALCHEXP | 12.6 |
| TCLTHEXP | 8.4 |
| THOUSEXP | 14.6 |
| TFURNEXP | 6.0 |

| **Variable** | $\lambda$ (%) |
|---|---|
| THLTHEXP | 12.3 |
| TTRANSEXP | 18.5 |
| TCOMMEXP | 9.2 |
| TRECEXP | 4.5 |
| TEDUEXP | 41.4 |
| TRESTHOTEXP | 16.6 |
| TMISCEXP | 6.4 |
| INCTOTGROSSHH | 73.6 |
| INCRMT | 32.1 |
| INCWAGE | 71.0 |
| INCFARMBSN | 15.1 |
| INCNFARMBSN | 24.0 |
| INCRENT | 10.4 |
| INCFIN | 1.3 |
| INCPENSN | 17.1 |
| INCOTHER | 17.7 |
| ROOF | 6.0 |
| TOILET | 7.6 |
| WATER | 9.4 |
| ELECTCON | 1.7 |
| FUELCOOK | 4.3 |
| OWNMOTORCYCLE | 3.3 |
| CAR | 1.5 |
| TV | 7.3 |
| LIVESTOCK | 1.3 |

### 3.5.7 References

Benschop, T., Machingauta, C., and Welch, M. (2022). *Statistical Disclosure Control: A Practice Guide*, https://readthedocs.org/projects/sdcpractice/downloads/pdf/latest/

Chatfield, C., and Collins, A. J., (1980). *Introduction to Multivariate Analysis*, Chapman and Hall, London, 1980.

Dandekar, R., Domingo-Ferrer, J., and Sebé, F., (2002). *LHS-based hybrid microdata vs. rank swapping and microaggregation for numeric microdata protection.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 153–162, Berlin Heidelberg, 2002. Springer.

De Waal, A. G., and Willenborg, L. C. R. J. (1999). *Information loss through global recoding and local suppression.* Netherlands Official Statistics, 14:17–20, 1999. special issue on SDC.

Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001). *Comparing sdc methods for microdata on the basis of information loss and disclosure risk*. In Pre-proceedings of ETK-NTTS'2001 (vol. 2), pages 807–826, Luxemburg, 2001. Eurostat.

Domingo-Ferrer, J., and Torra, V. (2001). *Disclosure protection methods and information loss for microdata.* In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pages 91–110, Amsterdam, 2001. North-Holland. http://vneumann.etse.urv.es/publications/bcpi.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R:, and Giessing, S. (2005). *μ-ARGUS version 4.0 Software and User's Manual*. Statistics Netherlands, Voorburg NL, may 2005. https://research.cbs.nl/casc.

Hundepool, A., Domingo–Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P. (2012). *Statistical Disclosure Control*. John Wiley & Sons, Ltd.

Kooiman, P. L., Willenborg, L. and Gouweleeuw, J. (1998). *PRAM: A method for disclosure limitation of microdata.* Technical report, Statistics Netherlands (Voorburg, NL), 1998.

Młodak, A. (2020). Information loss resulting from statistical disclosure control of output data. *Wiadomości Statystyczne. The Polish Statistician*, 65 (9), 7–27. (in Polish)

Młodak, A., Pietrzak, M., & Józefowski, T. (2022). The trade–off between the risk of disclosure and data utility in SDC: A case of data from a survey of accidents at work. *Statistical Journal of the IAOS*, 38 (4), 1503–1511.

Trottini, M. (2003) . *Decision models for data disclosure limitation*. PhD thesis, Carnegie Mellon University, 2003. http://www.niss.org/dgii/TR/Thesis-Trottini-final.pdf.

Willenborg, L., and De Waal, T., (2001). *Elements of Statistical Disclosure Control*. Springer-Verlag, New York, 2001.

Winkler, W. E. (1998). *Re-identification methods for evaluating the confidentiality of analytically valid microdata.* In J. Domingo-Ferrer, editor, Statistical Data Protection, Luxemburg, 1999. Office for Official Publications of the European Communities. (Journal version in Research in Official Statistics, vol. 1, no. 2, pp. 50-69, 1998).

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). *Disclosure risk assessment in perturbative microdata protection.* In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 135–152, Berlin Heidelberg, 2002. Springer.

## 3.6 Software

### 3.6.1 $\mu$-ARGUS

The $\mu$-ARGUS software has been developed to facilitate statisticians, mainly in the NSI's, to apply the SDC-methods described above to create safe micro data files. It is a tool to apply the SDC-methodology, not a black-box that will create a safe file without knowing the background of the SDC-methodology. The development of $\mu$-ARGUS has started at Statistics Netherlands by implementing the Dutch methods and rules. With this software as a starting point many other methods have been added. Several of these methods have been developed and/or actually implemented during the CASC-project.

In this section we will just a short overview of $\mu$-ARGUS, as an extensive manual is available, fully describing the software.

The starting point of $\mu$-ARGUS has been implementation of the threshold rules for identifying unsafe records and procedures for global recoding and local suppression.

**Data:** $\mu$-ARGUS can both protect fixed and free format ASCII files.

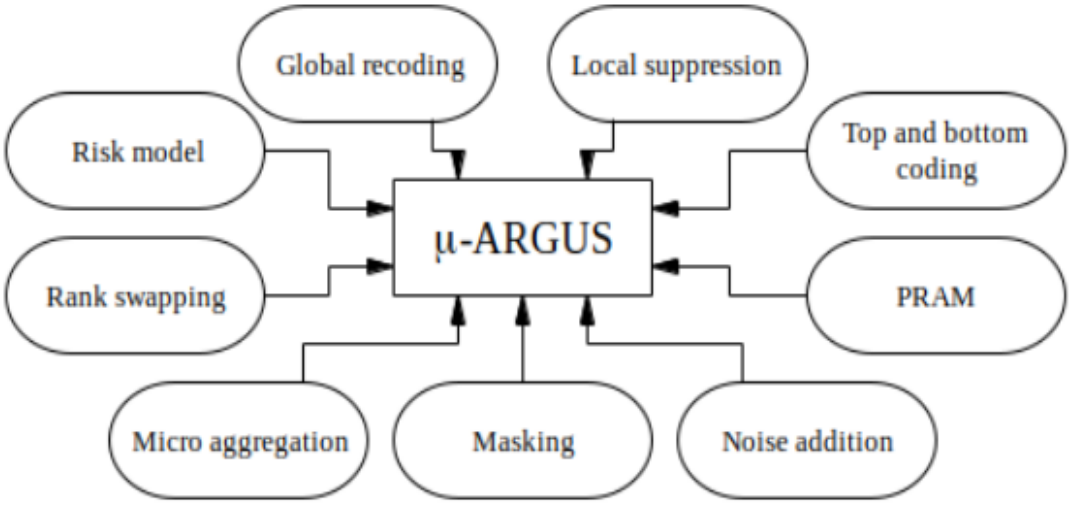


Figure 3.2: Overview of $\mu$-ARGUS

Many of the methods described previously in the methodology section can be applied with $\mu$-ARGUS to a dataset. It is our aim to include other methods as well in the near future, if time permits

$\mu$-ARGUS is a flexible interactive program that will guide you through the process of data protection. In a typically $\mu$-ARGUS run you will typically go through the following steps, given that the microdata set is ready.

1. **Meta data**. $\mu$-ARGUS needs to know the structure of the data set. Not only the general aspects but also additional SDC-specific information. As there is till now no suitable flexible standard for meta data allowing us to specify also the SDC-specific parts of the meta data, we have to rely on the ARGUS meta data format.
   This can be prepared (partially) externally or it can be specified interactively during a $\mu$-ARGUS session.
2. **Threshold-rule/risk models.** Selection and computation of frequency tables on which several SDC-methods (like risk models, threshold rule) are based
3. **Global recoding.** Selection of possible recodings and inspection of the results.
4. Selection and application of **other protection methods** like:
   - Microaggregation (3.4.2.3)
   - PRAM (3.4.6)
   - Rounding (3.4.2.5)
   - Top and bottom coding (3.4.3.3)
   - Rank swapping (3.4.2.4)
   - Noise addition (3.4.2.1)
5. **Risk model:** selection of the risk-level
6. **Generate the safe micro file.** During this process all data transformations specified above. This is also the moment that all remaining unsafe combinations will be protected by local suppressions. Also an extensive report will be generated.

When the above scheme has been followed a safe microdata file has been generated. $\mu$-ARGUS is capable of handling very large datasets. Only during the first phase, when the datafile is explored and the frequency tables are computed some heavy computations are performed. This might take some time depending on the size of the datafile. However all the real SDC-work (global recoding and the other methods named under 4 and 5 above) are done at the level of the information prepared during this first phase. This will be done very quickly. Only in the final phase when the protected datafile is made, the operation might be time consuming depending on the size of the datafile.

This architecture of $\mu$-ARGUS has the advantage that all real SDC-work, that will be done interactively, will have a very quick response time. Inspecting the results of various recodings is easy and simple.

### 3.6.2 sdcMicro

sdcMicro is an R package implementing almost all methods discussed in Section 3.4. The required steps to use the package are essentially the same as outlined in Section 3.6.1 and are quickly summarized below as well.

1. **Definition of a problem** The first step is always to create an object that defines the current sdc problem. This task can be achieved by calling function `createSdcObj()`. In this function, quite a few parameters can be set. The most important ones are:

- **Data**: the input data set needs to be a `data.frame` / `data.table` but it should be noted that any functionality from `R` can be used to create such objects from a variety of files exported or generated from other tools such as SAS, SPSS or STATA among using plain text-files (such as `.csv`) or other structured formats like `.json` or `.xml` as long those can be converted to a rectangular data structures. It is of course also possible to use queries to database systems in order to create suitable input objects.
- **Key variables for risk assessment**: the user is required to specify a set of categorical key variables. These variables are automatically used when computing risk measures (see also Section 3.3.3).
- **Numerical key variables**: It is also possible (but optional) to specify a set of numerical variables that are deemed important. Such variables can (automatically) be used to apply suitable perturbation methods (such as e.g masking by noise) to it.
- **Weights**: In case the underlying microdata step from a survey sample, a variable holding suitable weights can be specified. This is required in order to make sure that risk measures are computed correctly.
- **Strata**: Sometimes it is useful if a specific anonymization approach is applied independently to specific strata of the underlying population. In `sdcMicro` this can be achieved by defining a variable that holds different values for different groups of the population.
- **Ghost-variables**: This allows to link variables to (categorical key) variables in a sense that modifications to the relevant key-variable (e.g suppression) are transferred and applied to the dependent variables that are referred to as *"ghost"* variables
- **Excluding direct identifiers**: In statistical practice microdata files often contain direct identifiers which can be identified already on creation of an input object. If such variables have been defined, they will be removed prior to any computations.

It should be noted that while it is very convenient to work with an object created with `createSdcObj()`, it is perfectly possible to apply all implemented methods of the package also to simpler data-structures like a `data.frame`.

2. **Application of SDC-methods** Once a problem instance has been created, some helpful summary statistics such as the number of observations violating $k$-anonymity or (global) risk measures such as the expected number of re-identifications given the defined risk-scenario, are readily available and are shown by simply printing out the object.

   The next step is then to interactively apply SDC techniques to the object and re-assess the impact of its application both on risk-measures as well as on data-utility. If the application yields unexpected or bad results, the implemented `undo()`-method can be used to revert to the state before application of the specific methods. This

allows to quickly try out different parameter settings and makes the process of applying SDC methods quite interactive.

The package allows to (for example) add stochastic noise to numerical variables (3.4.2.1) using `addNoise()`, post-randomize values in categorically scaled variables (3.4.6) with function `pram()`, create synthetic microdata (3.4.7) with method `dataGen()` or perform global recoding (3.4.3.2) by using `globalRecode()`. Furthermore it is possible to apply rank-swapping (3.4.2.4) with function `rankSwap()`, compute SUDA-scores (3.3.7) using `suda2()`, compute individual risk estimates (3.3.5) with `indivRisk()` and `freqCalc()` as well as make a set of categorical key variables fulfill $k$-anonymity using `kAnon()`. In the current versions of the package, TRS (targeted record swapping, 5.6) is implemented and can be called using the `recordSwap()` function. A detailed discussion and overview is available in a custom vignette.

3. **Exporting Results** Once the interactive process has been finished, the package allows to quickly write out a safe dataset that contains all applied techniques also respecting any settings defined when initializing the object itself such as *"ghost-variables"* using function `writeSafeFile()`.

   Further more there is a `report()` functionality available that can be applied to an sdc-object at any time. This method can be called to either generate an internal or external report summarizing the process. The difference between the internal and the external report is the level of detail. While the external report is targeted for public consumption and does not contain any (sensitive) values such as specific parameter settings, the internal report lists any techniques that have been applied to protect the microdata in great detail. Both variants result in a `html` file that can easily be shared.

4. **Graphical User-Interface** Creating safe, protected microdata files is often a challenging task. Also having to dive into `R` and write code to perform several steps of the procedure can be a hurdle for non-experts in `R`. In order to mitigate this problem and to facilitate the use of `sdcMicro`, the package comes with an interactive, shiny-based graphical user-interface (Meindl, 2019). The interface can be started using the `sdcApp` function and its functionality is explained in detail in a custom vignette

## 3.7 Introductory example: rules at Statistics Netherlands

As has been shown in the previous sections there are many sophisticated ways of making a safe protected microdata set. And it is far from a simple straightforward task to select the most appropriate method for the Disclosure Protection of a microdata set. This requires a solid knowledge of the survey in question as well as a good overview of all the methods described in the previous sections.

However as an introduction we will describe here a method/set of rules inspired by those currently applied at Statistics Netherlands for making both microdata files for researchers as well as public use files. This approach can be easily applied, as it is readily available in $\mu$-ARGUS. These rules are based on the ARGUS threshold-rule in combination with global recoding and local suppression (see Section 3.4.3.2 and 3.4.3.4). This rule only concentrates on the identifying variables or key-variables, as these are the starting point for an intrusion. There rules have primarily been developed for microdata about persons.

*Microdata for researchers*
For the microdata for researchers one could use the following set of rules:

1. Direct identifiers should not be released and therefore should be removed from the microdata set.
2. The indirect identifiers are subdivided into extremely identifying variables, very identifying variables and identifying variables. Only direct regional variables are considered to be extremely identifying. Very identifying variable are very visible variables like gender, ethnicity etc. Each combination of values of an extremely identifying variable, a very identifying variable and an identifying variable should occur at least 100 times in the population.
3. The maximum level of detail for occupation, firm and level of education is determined by the most detailed direct regional variable. This rule does not replace rule 2, but is instead a practical extension of that rule.
4. A region that can be distinguished in the microdata should contain at least 10 000 inhabitants.
5. If the microdata concern panel data direct regional data should not be released. This rule prevents the disclosure of individual information by using the panel character of the microdata.

If these rules are violated, global recoding and local suppression are applied to achieve a safe file. Both global recoding and local suppression lead to information loss, because either less detailed information is provided or some information is not given at all. A balance between global recoding and local suppression should always be found in order to make the information loss due to the statistical disclosure control measures as low as possible. It is recommended to start by recoding some variables globally until the number of unsafe combinations that has to be protected is sufficiently low. Then the remaining unsafe combinations have to be protected by local suppressions.

For business microdata these rules are not appropriate. Opposite to personal microdata business data tends to be much more skewed. Each business is much more visible in a microdata set. This makes it very hard to make a safe business micro dataset.

*Microdata for the general public*
The software package $\mu$-ARGUS (Hundepool et al, 2005) is also of help in producing public use microdata files. For public use microdata files one could use the following set of rules:

1. The microdata must be at least one year old before they may be released.
2. Direct identifiers should not be released. Also direct regional variables, nationality, country of birth and ethnicity should not be released.
3. Only one kind of indirect regional variables (e.g. the size class of the place of residence) may be released. The combinations of values of the indirect regional variables should be sufficiently scattered, i.e. each area that can be distinguished should contain at least 200 000 persons in the target population and, moreover, should consist of municipalities from at least six of the twelve provinces in the Netherlands. The number of inhabitants of a municipality in an area that can be distinguished should be less than 50 % of the total number of inhabitants in that area.
4. The number of identifying variables in the microdata is at most 15.
5. Sensitive variables should not be released.
6. It should be impossible to derive additional identifying information from the sampling weights.
7. At least 200 000 persons in the population should score on each value of an identifying variable.
8. At least 1 000 persons in the population should score on each value of the crossing of two identifying variables.
9. For each household from which more than one person participated in the survey we demand that the total number of households that correspond to any particular combination of values of household variables is at least five in the microdata.
10. The records of the microdata should be released in random order.

According to this set of rules the public use files are protected much more severely than the microdata for research. Note that for the microdata for research it is necessary to check certain trivariate combinations of values of identifying variables and for the public use files it is sufficient to check bivariate combinations, but the thresholds are much higher. However, for public use files it is not allowed to release direct regional variables. When no direct regional variable is released in a microdata set for research, then only some bivariate combinations of values of identifying variables should be checked according to the statistical disclosure control rules. For the corresponding public use files all the bivariate combinations of values of identifying variables should be checked.

## 3.8 Further examples

In this section we provide examples based on real surveys of the various steps described in the previous sections in order to describe a possible process of microdata anonymisation in practice. The two surveys analysed are one from the social domain, the Labour Force Survey (LFS), and one on business data, the Community Innovation Survey (CIS). Notice how the complexity of the reasoning on business data may rise sharply as compared to social microdata.

### 3.8.1 Labour Force Survey

The Labour Force Survey[3] is one of the surveys subject to Regulation EC 831/2002 on access to microdata for scientific purposes. The Labour Force Survey (LFS) is the main data source for the analysis of the labour market, employment, unemployment as well as the conditions and the level of involvement at the labour market. Some of the main observed variables are:

1. demographic variable: (gender, year of birth, marital status, relationship to reference person, place of residence);
2. labour status: (labour status during the reference week, etc.);
3. employment characteristics of the main job: (professional status, economic activity of local unit, country of place of work, etc.);
4. education and training;
5. income;
6. technical item relating to the interview.

The sampling designs in the EU-LFS[4] are extremely varied. Most NSIs employ some kind of multistage stratified random sampling design, especially those that do not have central population register available.

In the document Eurostat (2004) a proposal for the anonymisation of the LFS is presented. Here we show a possible approach to disclosure scenario definition that leads to the definition of the identifying variables.

*Disclosure scenarios*
A spontaneous identification (see Section 3.3.2) can happen when an intruder has a direct knowledge of some statistical units belonging to the sample and, whether such units assume extremely particular values for some variables or for some combinations. In the Labour Force data set there are several variables that may lead to a spontaneous identification. Some of these variables are: professional status, number of persons working at local unit, income, economic activity, etc. To avoid a possible spontaneous identification such variables are usually checked to see whether there are unusual patterns or very rare keys and, if necessary some recoding or suppression may be suggested.

Also the external register scenario could be considered for the LFS data. The two external archives taken as example in the Italian study are: (i) the Electoral roll for the Individual archive scenario, and (ii) the Population register for the Household archive scenario (see Section 3.3.2). The *Electoral roll* is a register containing information on people having electoral rights, mainly demographic variables (gender, age, place of residence and birth) and sometimes variables such as, marital status, professional and/or educational information.

[3]The European Union Labour Force Survey (EU LFS)? is conducted in the 25 Member States of the European Union and 3 countries of the European Free Trade Association (EFTA) in accordance with Council Regulation (EEC) No. 577/98 of 9 March 98: http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1913,47567825,1913_47568351&_dad=portal&_schema=PORTAL

[4]http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-CC-06-007/EN/KS-CC-06-007-EN.PDF

The key variables considered as reliable for re-identification attempts under this scenario are: gender, age and residence. Place of birth is removed from the MF and the others are not considered as their quality was not deemed sufficient for re-identification purposes. The *Population register* maybe a public register containing demographic information at individual and household level. Particularly, the set of key variables considered for the household archive scenario comprises the variables: gender, age, place of residence and marital status as individual information, the household size and parental relationship as household information.

*Risk assessment*
The definition of the identifying variables allows for a definition of risk assessment. This can be performed as in Eurostat (2004) following the reasoning of Section 3.7 or other risk measures can be used. If the survey employs a complex multi-staged stratified random sample design possibly with calibration, then the ARGUS individual risk may be used especially when hierarchical information on the household need to be released.

*Protection*
The risk assessment procedure will show the keys at risk and based on this information a strategy for microdata protection needs to be adopted. If the number of keys at risk is very large then some variables are too detailed and some global recoding is advisable in order to avoid the application of a high percentage of local suppressions. If the keys at risk are particularly concentrated on certain values of an identifying variable a local recoding of such variable could be sufficient to solve the problem.

### 3.8.2 Community Innovation Survey

The Community Innovation Survey is one of the surveys subject to Regulation CE 831/2002 on access to microdata for scientific purpose. A lot of effort has been put in anonymising this microdata set, see for example Eurostat (2006).

In this section we propose a study of disclosure scenarios to define identifying variables and a risk assessment analysis to single out the records at risk for the Community Innovation Survey (CIS) based on Ichim (2006). A protection stage is then outlined giving different choices. The interested reader is referred to that paper for more information on the whole process.

CIS provides information on the characteristics of innovation activity at enterprise level[5].

[5]Some of the main observed variables in the CIS3 are : principal economic activity, geographical information, number of employees in 1998 and 2000, turnover in 1998 and 2000, exports in 1998 and 2000, gross investment in tangible goods: 2000, number of valid patents at end of 2000, number of employees with higher education (in line with the number of employees in 2000), expenditure in intramural RD (in line with the turnover in 2000), expenditure in extramural RD (in line with the turnover in 2000), expenditure in acquisition of machinery (in line with the turnover in 2000), expenditure in other external knowledge (in line with the turnover in 2000), expenditure in training, market (in line with the turnover in 2000), total innovation expenditure (in line with the turnover in 2000), number of persons involved in intra RD (in line with the number of employees in 2000).

The CIS statistical population is determined by the size of the enterprise (all enterprises with 10 or more employees) and its principal economic activity.

*Disclosure scenario*
Since generally business registers are publicly available, it is supposed that an intruder could use such information to identify an enterprise. Public business registers report general information on name, address, turnover (TURN), number of employees (EMP), principal activity of an enterprise (NACE), region (NUTS). Therefore, the identifying variables of the hypothesized disclosure scenario are: economic classification (NACE), region (NUTS), number of employees (EMP) and turnover (TURN). The information content of these variables must be somehow reduced in order to increase the intruder uncertainty. An initial coding performed on such variables was: NACE at 2 digits, Nuts recoded at national level (no regional breakdown) and three enterprise size classes.

Additionally, in the CIS data set there are several confidential variables that may be subject to spontaneous identification. Some examples are total expenditure on innovation (RTOT), exports, number of persons involved in intra RD, etc. Such variables are never published in an external register, but they can assume extremely particular values on some units. Mere additional information would then clearly identify an enterprise. Special attention must be paid on these variables. A check performed by the *survey experts* is generally suggested. These assessments must be performed with respect to each combination of categorical identifying variables to be released. The analysis by the survey expert suggested to remove from the data to be released the variable Country of head office. With the given details on NACE, size class and NUTS all the other continuous variables were not deemed sufficiently spread to lead to a spontaneous identification of a unit. For this reason it maybe suggested to let them unchanged.

*Risk assessment*
A unit is considered at risk if it is 'recognisable' either in the external register scenario or in the spontaneous identification scenario. It is assumed that an intruder may confuse a unit U with others when there is a sufficient number of units in a well-defined (and not too large) neighbourhood of U. The anonymisation proposal developed in Ichim (2006) is based on the idea that similarity based on clusters and confusion both express the same concept, although in different frameworks. : When a unit belongs to a cluster, it belongs to a high density (sufficient number of close units) subset of data. Hence the unit may be considered as being confused with others. The algorithms taking into account these two features (distance from each other and number of neighbours) are called density based algorithms and Ichim (2006) uses one of these algorithms to identify isolated units i.e. units at risk with respect to the identifying variables.

*Protection by perturbation*
Once the units at risk have been identified, protection should be applied. Several different proposals in the field of data perturbation methods are possible. The proposal by Eurostat protection is achieved by the application to the main continuous variables in the data set of individual ranking and some local suppression of particular values. This microaggregation

would be applied to the whole file irrespective to different economic classifications or size classes and without taking into account possible relationships between variables (for example turnover needs to be greater than export or expenditures). This strategy is perfectly acceptable if a slight modification of the data is deemed sufficient.

An alternative could be to apply a perturbation only to these records at risk (mainly the large size enterprises in single NACE 2 digits) whereas the rest of the file is released unchanged. Ichim (2006) suggests different perturbations of the points at risk whereas these are in the middle of the distribution of points (nearest cluster imputation) or if they are in the tail (microaggregation). A further adjustment is proposed in order to preserve turnover totals for each combination of categorical identifying variables. This is deemed important by users who need to compare results with published tables. A study of the information loss of this approach is presented in Ichim (2006).

### 3.8.3 References

Eurostat (2004). *Proposal on anonymised LFS microdata.* CSC 2004/B5/ item 2.2.2.

Eurostat (2006). *The CIS4. An amended version of the micro-data anonymisation method.* Doc. Eurostat/F4/STI/CIS/M2/8.

Ichim, D. (2006). *Microdata anonymisation of the Community Innovation Survey: a density based clustering approach for risk assessment.* Contribution Istat. Shortly available from http://www.istat.it/dati/pubbsci/contributi/

Trottini, M., Franconi, L. and Polettini, S. (2006). *Italian Household Expenditure Survey: A proposal for Data Dissemination.* In Domingo Ferrer, J and Franconi, L. (eds) Privacy in Statistical Databases, CENEX-SDC Project International Conference, Rome, Italy, December 2006, 318-333.

# 4 Magnitude tabular data

## 4.1 Introduction

Statistical magnitude tables display sums of observations of a quantitative variable where each sum relates to a group of observations defined by categorical variables observed for a set of respondents.

Respondents are typically companies but can also be individuals or households, etc. Grouping variables typically give information on geography or economic activity or size, etc. of the respondents. The *"cells"* of a table are defined by cross-combinations of the grouping variables.

Each *"table cell"* presents a sum of a quantitative variable such as income, turnover, expenditure, sales, number of employees, number of animals owned by farms, etc. These sums are the *"cell values"* (sometimes also referred to as *"cell totals"*) of a magnitude table. The individual observations of the variable (for each individual respondent) are the *"contributions"* to the cell value.

| | Industry A | Industry B | ··· | Total |
|---|---|---|---|---|
| Region 1 | 540 (12) | 231 (15) | ··· | |
| Region 2 | 48 (2) | 125 (8) | ··· | |
| ... | ... | ... | ... | |

Table 4.1: Example: Turnover (number of respondents) by Region and Industry

The *"dimension"* of a table is given by the number of grouping variables used to specify the table. We say that a table contains *"margins"* or *"marginal cells"*, if not all cells of a table are specified by the same number of grouping variables. The smaller the number of grouping variables, the higher the *"level"* of a marginal cell. A two-dimensional table of some business survey may for instance provide sums of observations grouped by economic activity and company size classes. At the same time it may also display the sums of observations grouped by only economic activity or by only size classes. These are then margins/marginal cells of this table. If a sum across all observations is provided, we refer to it as the *"total"* or *"overall total"*.

At first sight, one might find it difficult to understand how the kind of summary information published in magnitude tables presents a disclosure risk at all. However, it often

occurs that cells of a table relate to a single or to only a few respondents. The number of this kind of small cells in a table will increase, the more grouping variables are used to specify the table, the higher the amount of detail provided by the grouping variables, and the more uneven the distributions of respondents over the categories of the grouping variables.

If a table cell relates to a small group of respondents (or even only one), then publication of the cell value may imply a disclosure risk. This is the case if these respondents could be identified by an intruder using information displayed in the table.

> **Example 1** Let a table cell display the turnover of companies in the mining sector for a particular region X. Let us assume that company A is the only mining company in this region. This is a fact that will be known to certain intruders (think, for instance, of another mining company B in a neighbouring region Y). So, if that table cell is published, company B would be able to disclose the turnover of company A.

In order to establish if a disclosure risk is connected to the publication of a cell value in a table, and in order to protect against this risk, data providers (like, e.g. National Statistical Institutes) should apply tabular data protection methods. In many countries this is a legal obligation to official statistical agencies. It may also be regarded as a necessary requirement in order to maintain the trust of respondents: After all, if in the instance above company A realizes that company B might, by looking into the published table, disclose the value of turnover it has reported, and if it considers this value as confidential information, it may refuse to respond to that survey in the next period, or (if the survey is compulsory) it may choose to provide incorrect or inaccurate information.

Especially for business statistics, the most popular method for tabular data protection is *cell suppression* . In tables protected by cell suppression, all values of cells for which a disclosure risk has been established are eliminated from the publication. Alternatively, other methods based on cell perturbation etc. may also be used to protect tabular data. While we focus in this chapter on cell suppression, we will also mention alternatives.

Section 4.2 introduces into the methodological concepts of tabular data protection. In Section 4.2.1 we present the most common methods for disclosure risk assessment for each individual cell of a table (and for special combinations of individual cells). These methods are called *"primary"* disclosure control methods. Table cells for which a disclosure risk has been established are called *"sensitive"*, *"confidential"*, or *"unsafe"* cells. Primary disclosure risk is usually assessed by applying certain sensitivity rules. Section 4.2.1 explains the concept of sensitivity rules and the most common rules.

While detailed tabulated summary information also on smaller groups of statistical objects (companies by subgroups of subsectors at the district level, households by size and income at the neighbourhood level, etc.) might be of interest to certain user groups, it is also a responsibility (maybe the most important one) of official statistics to provide summary

information at a high aggregate level by producing summary statistics on large groups of a population (e.g. for *all* companies of an economy sector). Because of this, it is not enough to have methodologies to protect individual cells. It implies a need for the so-called *"secondary"* tabular data protection methodologies.

Assume that a table displays the sum of a variable "Production" by three subsectors of an economy sector. Assume that this sum is sensitive for one of the subsectors and that the table is protected by cell suppression, meaning that the confidential cell value is suppressed.

**Example 1a** Production (in mill. Euro)

| Sector | Subsector I | Subsector II | Subsector III |
|---|---|---|---|
| 56,600 | suppressed (sensitive) | 47,600 (non-sensitive) | 8,002 (non-sensitive) |

With respect to the total production for this sector we distinguish two cases: Either it is foreseen to be published – we then consider it as a cell of the table (e.g. the "total"). If the cell values of the two non-sensitive subsectors and the "total" are displayed, then users of the publication can disclose the cell value for the sensitive subsector by taking the difference between the "total" and the subsector values for the two non-sensitive sectors (56,600 – 47,600 – 8,002 = 998). In order to avoid this, a secondary protection measure for this table has to be taken, e.g. selecting one of the two non-sensitive subsector cells and suppressing it as well. This would be called a "secondary suppression".
The other option is that the "total" is not foreseen to be displayed / published. Then no secondary protection measure would be needed. In this instance, because the production for one subsector is suppressed, interested users of the table cannot compute the production for the sector on their own – and so the sector-level information is completely lost!

From a general perspective, the purpose of secondary tabular data protection methodologies is chiefly to avoid undesirable effects such as this, ensuring that – while some "small", primary confidential cells within detailed tables may have to be protected (by cell suppression or by some perturbative method) – sums for larger groups, i.e. the margins of those detailed tables, are preserved to some extent. For cell suppression this means that suppression of marginal cells should be avoided as far as possible. For perturbative methods it means that high level margins should try to be preserved exactly (more or less).

Considering this as the basic idea of secondary protection, after Section 4.2.1 we assume for the remainder of the chapter margins and overall totals always to be part of a table.

In Section 4.2.2 we introduce the concepts of secondary tabular data protection methodologies. The focus will be on cell suppression, but we will also mention other methodologies.

The software package $\tau$-ARGUS (see Hundepool et al., 2005) provides software tools for disclosure protection methods for tabular data. Section 4.3 is concerned with the practical implementation of secondary cell suppression as offered by $\tau$-ARGUS. In Section 4.3.1 we discuss information loss concepts as well as table structures considered by $\tau$-ARGUS. We compare the performance of different algorithms for secondary cell suppression in Section 4.3.2 and give software recommendations. In Section 4.3.3 we explain how to set up procedures for tabular data protection in a practical way and give an introductive example in Section 4.3.4. In Section 4.4 we briefly introduce the methodological concepts of the secondary cell suppression algorithms provided by $\tau$-ARGUS. The chapter ends with Section 4.5, introducing Controlled Tabular Adjustment as new emerging protection technique for magnitude tables which could become an alternative to the well-established cell suppression methodologies.

## 4.2 Disclosure Control Concepts for Magnitude Tabular Data

In this section we explain the main concepts for assessment and control of disclosure risk for magnitude tables.

Section 4.2.1 is concerned with disclosure risk for each individual cell of tables presenting summaries of quantitative variables and will introduce the most common methods used to assess this risk.

In order to preserve the margins of tables to some extent while protecting individual sensitive cells, special disclosure control methodologies have been developed. Section 4.2.2 presents basic concepts of these methodologies, focusing on secondary cell suppression as the most prominent instance. We finish section methods with a brief, comparative overview of alternative methods for tabular data protection.

### 4.2.1 Sensitive Cells in Magnitude Tables

We begin this section by describing intruder scenarios typically considered by statistical agencies in the context of disclosure control for magnitude tables. Considering these intruder scenarios statistical agencies have developed some *'safety rules'* as measures to assess disclosure risks. This section will also introduce the most popular rules using some illustrative examples. Finally, we compare rules and give guidance on making a decision between alternative rules.

***Intruder scenarios***
If a table cell relates to a small group (or even only one) respondent, then publication of the cell value may imply a disclosure risk. This is the case, if these respondents could be identified by an intruder using information displayed in the table. In example 1 of Section 4.1, for the intruder (company B) it is enough to know that the cell value reports

the turnover of mining companies in region X. In that example company B is assumed to be able to identify company A as the only mining company in region X. Hence, publication of the cell value implies a disclosure risk: if company B looks into the publication they will be able to disclose the turnover of company A.

But what if a cell value does not relate to one, but to two respondents?

**Example 1b** Let us assume this time that both companies (A and B) are located in region X, and are the only mining companies there. Let us further assume that they both are aware of this fact. Then again publication of the cell value implies a disclosure risk (this time to both companies): if any of the two companies look into the publication and subtract their own contribution to the cell value (i.e. the turnover they reported) from the cell value, they will be able to disclose the turnover of the other company.

**Example 1c** Assume now that the table cell relates to more than two respondents. Imagine this time that four companies (A, B, C and D) are located in region X. Then theoretically three of them (B, C and D, say) could form a *coalition* to disclose the turnover of company A. Such a coalition might be a rather theoretical construct. An equivalent but perhaps more likely scenario could be that of another party who knows the contributions of companies B, C and D (perhaps a financial advisor working for all three companies) who would then be able to disclose also the turnover of company A by subtracting the contributions of B, C and D from the cell value.

The examples above are based on the intruder scenario typical for business data: it is usually assumed, that the "intruders", those who might be interested in disclosing individual respondent data, may be "other players in the field", e.g. competitors of the respondent or other parties who are generally well informed on the situation in the part of the economy to which the particular cell relates. Such intruder scenarios make sense, because, unlike microdata files for researchers, tabular data released by official statistics are accessible to everybody – which means they are accessible in particular to those well informed parties.

In the scenarios of example 1 there is a risk that magnitude information is disclosed exactly. But how about approximate disclosure?

**Example 1d** Let us reconsider the example once more. Assume this time that in region X there are 51 companies that belong to the mining sector, e.g. company A and 50 very small companies $S_1$ to $S_{50}$. Assume further that 99 % of the turnover in mining in region X is contributed by company A. In that scenario, the cell value (turnover in the mining sector for region X) is a pretty close approximation of the turnover of company A. And even though the potential intruder (in our example mining company B of the neighbour region Y) may not be able to identify all 51 mining companies of region X, it is very likely that they will know that there is one very big company in region X and which company that is.

***Sensitivity of variables***

The presumption of the sensitivity of a variable often matters in the choice of a particular protection method. For example, especially in the case of tables presenting business magnitude information many agencies decide that this kind of information must be protected also against the kind of approximate disclosure illustrated by example 1d above, because it is so sensitive.

Considering the above explained intruder scenarios statistical agencies have developed some *'safety rules'* (also referred to as *'sensitivity rules'* or *'sensitivity measures'*), measures to assess disclosure risks. We will now introduce the most popular rules, starting with an overview presenting formal representation of these rules in Table 4.3. After that, the rules (or rather, classes of rules) will be discussed in detail. We explain in which situations it may make sense to use those rules, using simple examples for illustration where necessary.

***Sensitivity rules***

Table 4.3 briefly presents the most common sensitivity rules. Throughout this chapter we denote $x_1 \geq x_2 \geq \cdots \geq x_N$ the ordered contributions by respondents $1, 2, \ldots, N$, respectively, to a cell with cell total (or cell value) $X = \sum_{i=1}^{N} x_i$

Table 4.3: Sensitivity rules

| **Rule** | **Definition** |
|---|---|
| | **A cell is considered *unsafe*, when ...** |
| **Minimum frequency rule** | the cell frequency is less than a pre-specified minimum frequency $n$ (the common choice is $n = 3$). |
| $(n, k)$-**dominance rule** | the sum of the $n$ largest contributions exceeds $k$% of the cell total, *i.e.* $$x_1 + \ldots + x_n > \frac{k}{100} X \quad (4.1)$$ |

**$p$%-rule** the cell total minus the 2 largest contributions $x_1$ and $x_2$ is less than $p$% of the largest contribution, *i.e.*[6]

$$X - x_2 - x_1 < \frac{p}{100} x_1 \tag{4.2}$$

---

Note that both the dominance rule and the $p$%-rule are meaningful only when all contributions are non-negative. Moreover, the dominance rule does not make sense for $k = 100$ and neither does the $p$%-rule for $p = 0$. Both rules are asymptotically equal to minimum frequency rules for $k \to 100$, or $p \to 0$ respectively.

Both, the dominance rule and the $p$%-rule belong to a class of rules which are referred to as *"concentration rules"* below.

When cell suppression is used to protect the table, any aggregate (or: cell in a table) that is indeed *'unsafe'*, or *'sensitive'* according to the sensitivity rule employed, is subject to what is called *'primary suppression'*.

Choice of a particular sensitivity rule is usually based on certain intruder scenarios involving assumptions about additional knowledge available in public or to particular users of the data, and on some (intuitive) notion on the sensitivity of the variable involved.

***Minimum frequency rule***

When the disseminating agency thinks it is enough to prevent exact disclosure, all cells with at least as many respondents as a certain, fixed minimum frequency $n$ are considered safe. Example 1 of Section 4.1 (cell value referring to one company in the mining sector of a region) illustrates the disclosure risk for cells with frequency 1. Example 1b above (on two mining companies) shows that there is a similar risk for cells with frequency 2.

Normally the minimum frequency $n$ will be set to 3.

An exception is the case when for some $n_0$ larger than 3 the agency thinks it is realistic to assume that a coalition of $n_0 - 2$ respondents contributing to the same cell may pool their data to disclose the contribution of another respondent. In such a case we set $n$ to $n_0$. Example 1c above provides an instance for this case with $n_0 = 5$. (The intruder knows the pooled data of $5 - 2 = 3$ companies (e.g. B, C and D)).

It should be stressed here that a minimum frequency larger than 3 normally does not make much sense, even though in example 1c, for some cells of a table it may happen that such a 'pooled data' situation actually occurs. Usually there is no way for an agency to know for which cell which size to assume for the 'pool'. Let us assume, for instance, that in a cell with 100 respondents 99 are in the stock market and are therefore obliged to publish data

[6] $X - x_n - \cdots - x_2 - x_1 < \frac{p}{100} x_1$ for the case of coalitions of $n - 1$ respondents, where $n > 2$

which are also their contributions to that cell value. Then we should consider 99 as the size of the pool, because anybody could add up (*i.e.* pool) these 99 published data values in order to disclose the (confidential) contribution of company 100 who is not in the stock market. But should the agency really consider all cells with less than 101 respondents as unsafe?

***Concentration rules***

A published cell total is of course always an upper bound for each individual contribution to that cell. This bound is the closer to an individual contribution, the larger the size of the contribution. This fact is the mathematical foundation of the well-known concentration rules. Concentration rules like the dominance and $p$%-rule make sense only if it is assumed specifically that the intruders are able to identify the largest contributors to a cell. The commonly applied sensitivity rules differ in the particular kind and precision of additional knowledge assumed to be around.

When a particular variable is deemed strongly confidential, preventing only exact disclosure may be judged inadequate. In such a case a concentration rule should be specified. For reasons that will be explained below, we recommend use of the so called $p$% -rule. Another, well known concentration rule is the '$n$ respondent, $k$ percent' dominance rule. Note, that it is absolutely essential to keep the parameters of a concentration rule confidential!

Traditionally, some agencies use a combination of a minimum frequency rule together with a $(1, k)$-dominance rule. This approach, however, is inadequate, because it ignores the problem that in some cases the contributor with the second largest contribution to a cell which is non-sensitive according to this rule is able to derive a close upper estimate for the contribution of the largest one by subtracting her own contribution from the aggregate total. Example 1 provides an instance.

**Example 1** Application of the (1,90)-rule. Let the total value of a table cell be $X = 100,000$, let the two largest contributions be $x_1 = 50,000$ and $x_2 = 49{,}000$. Since 50,000 $< \frac{90}{100}$100,000 the cell is safe according to the $(1, 90)$-rule: there seems to be no risk of disclosure. But the second largest contributor is able to derive an upper estimate $\hat{x}_1 = 100{,}000 - 49{,}000 = 51{,}000$ for the largest contribution which overestimates the true value of $50,000$ by 2% only: quite a good estimate!

Unlike the $(1, k)$-dominance rule, both the $(2, k)$-dominance rule and $p$%-rule take the additional knowledge of the second largest contributor into account properly. Of the two, the $p$%-rule should be preferred, because the $(2, k)$-dominance rule has a certain tendency for overprotection, as we will see in the following.

***$p$%-rule and dominance-rule***

We will show in the following that, according to both types of concentration rules, an

aggregate total (*i.e.* cell value) $X$ is considered as sensitive, if it provides an upper estimate for one of the individual contributions that is relatively close to this contribution.

⚠ Expert level

Assume that there are no coalitions of respondents, *i.e.* there are no intruders knowing more than one of the contributions. Then the closest upper estimate of any other contribution can be obtained by the second largest contributor, when it subtracts its own contribution $x_2$ from the aggregate total (*i.e.* cell value) $X$ to estimate the largest contribution ($\hat{x}_1 = X - x_2$) as seen in example 1. All other scenarios of a contributor subtracting its own value from the total, to estimate another, result in larger relative error. In the, rather unlikely, scenario that $n-1$ (for $n > 2$), respondents pool their data in order to disclose the contribution of another, the closest upper estimate of any other contribution can be obtained by the coalition of respondents $2, 3, \ldots, n$ when they estimate $\hat{x}_1 = X - x_2 - x_3 - \ldots - x_n$.
The question is now, how to determine whether such an estimate is 'relatively close'. Application of the $p$%-rule yields that the upper estimate $\hat{x}_1$ will overestimate the true value by at least $p$% for any non-sensitive cell, *i.e.* $\hat{x}_1 - x_1 \geq \frac{p}{100} x_1$ . That is, the $p$%-rule sets the difference between estimate and true value of the largest contribution in relation to the value of the largest contribution itself.
When we adapt relation (4.1) in table 1 (see definition of the $(n,k)$-rule) to the case of $n = 2$, subtract both sides from $X$ and then divide by $X$ the result is

$$(X - x_2) - x_1 < \frac{100 - k}{100} X \tag{4.3}$$

In this formulation, the $(2,k)$-rule looks very similar to the formulation of the $p$%-rule given by (4.2). Both rules define an aggregate to be sensitive, when the estimate $\hat{x}_1 = X - x_2$ does not overestimate the true value of $x_1$ 'sufficiently'. The difference between both rules is in how they determine this 'sufficiency'. According to the $p$%-rule, it is expressed as a rate (i.e. $p$%) of the true value of the largest contribution $x_1$, while according to the $(2,k)$-rule, it is expressed as a rate (i.e. $(100-k)$%) of the aggregate total $X$. Considering this, the concept of the $p$%-rule seems to be more natural than that of the $(2,k)$-rule.
$(2,k)$-rules correspond to $p$%-rules in the following way:
If $k$ is set to $100\frac{100}{100+p}$ then

A) any aggregate, which is safe according to the $(2,k)$-rule, is also safe according to the $p$%-rule (this will be proven below), but

B) not any aggregate, which is safe according to the $p$%-rule, is also safe according to this $(2,k)$-rule. An example is given below (example 2). In these cases the aggregate could be published according to the $p$%-rule, but would have to be suppressed according to the $(2,k)$-rule.

Based on the above explained idea, that the concept of the $p$%-rule is more natural than that of the $(2,k)$-rule, we interpret this as a tendency for over-protection in the $(2,k)$-rule. Example 2 below is an instance for this kind of over-protection.
We therefore recommend use of the $p$%-rule instead of a $(2,k)$-dominance rule.

***How to obtain the $p$ parameter?***
When we replace a $(2,k)$-dominance rule, by a $p$%-rule, the natural choice is to derive the parameter $p$ from $k = 100\frac{100}{100+p}$ , e.g. to set $p = 100\frac{100-k}{k}$
Thus, a $(2,80)$-dominance rule would be replaced by a $p$%-rule with $p = 25$, a $(2,95)$-dominance rule by a $p$%-rule with $p = 5.26$ .
If we also derive $p$ from this formula, when replacing a $(1,k)$-dominance rule, we will obtain a much larger number of sensitive cells. In addition to the cells which are unsafe according to the $(1,k)$-dominance rule which will then also be unsafe according to the $p$%-rule, there will be cells which were safe according to the $(1,k)$-dominance rule, but are not safe according to the $p$%-rule, because the rule correctly considers the insider knowledge of a large second largest contributor. We could then put up with this increase in the number of sensitive cells. Alternatively, we could consider the number of sensitive cells that we used to assign (with the $(1,k)$-dominance rule) as a kind of a maximum-prize we are prepared to 'pay' for data protection. In that case we will reduce the parameter $p$. The effect will be that some of the cells we used to consider as sensitive according to the $(1,k)$-dominance rule will now not be sensitive. But this would be justified because those cells are less sensitive as the cells which are unsafe according to the $p$%-rule, but are not according to the former $(1,k)$-dominance rule, as illustrated above by Example 1.

**Example 2** Let $p = 10$, then $k = 100\frac{100}{100+p} = 90.9$, let the total value of a table cell be $X = 110{,}000$, let the largest two contributions be $x_1 = 52{,}000$, and $x_2 = 50{,}000$.
Then

$$\hat{x_1} = X - x_2 = 110{,}000 - 50{,}000 = 60{,}000$$

and

$$100\frac{\hat{x_1} - x_1}{x_1} = 100\frac{60{,}000 - 52{,}000}{52{,}000} = 15.4$$

*i.e.* the upper estimate $\hat{x_1} = X - x_2$ will overestimate the true value by 15.4%. So the aggregate is safe according to the $p$%-rule at $p = 10$.
On the other hand the two largest contributions are $x_1 + x_2 = 102{,}000$. As $102{,}000 > \frac{100}{100+p}X = 100{,}000$ the aggregate is not safe according to the $(2,k)$-rule.

*Proof of A.*

For an aggregate which is safe according to the $(2,k)$-rule with $k = 100 \cdot \frac{100}{100+p}$ the following will hold:

$$x_1 \leq \frac{k}{100} \cdot X = \frac{100}{100+p} X \tag{4.4}$$

and

$$\frac{(X - x_2) - x_1}{X} \geq 1 - \frac{k}{100} = 1 - \frac{100}{100+p} = \frac{p}{100+p} \tag{4.5}$$

(c.f. (4.3) ).
This is equivalent to

$$\frac{(X - x_2) - x_1}{x_1} \geq \frac{p}{100+p} \frac{X}{x1} \tag{4.6}$$

From (4.4) it follows that

$$\frac{p}{100+p} \frac{X}{x_1} \geq \frac{p}{100}$$

And hence from (4.6) that

$$\frac{(X - x_2) - x_1}{x_1} \geq \frac{p}{100+p} \cdot \frac{X}{x_1} \geq \frac{p}{100}$$

***The*** $(p,q)$***-rule***
A well known extension of the $p$%-rule is the so called prior-posterior $(p,q)$-rule. With the extended rule, one can formally account for general knowledge about individual contributions assumed to be around *prior* to the publication, in particular that the second largest contributor can estimate the smaller contributions $X_R = \sum_{i>2} x_1$ to within $q$%. An aggregate is then considered unsafe when the second largest respondent could estimate the largest contribution $x_1$ to within $p$ percent of $x_1$ , by subtracting her own contribution and this estimate $\hat{X}_R$ from the cell total, *i.e.* when $|\,(X - x_2) - x_1 - \hat{X}_R| < \frac{p}{100} x_1$. Because $(X - x_2) - x_1 = X_R$, the left hand side is assumed to be less than $\frac{q}{100} X_R$. So the aggregate is considered to be sensitive, if $X_R < \frac{p}{q} x_1$. Evidently, it is actually the ratio $\frac{p}{q}$ which determines which cells are considered safe, or unsafe. Therefore, any $(p,q)$-rule with $q < 100$ can also be expressed as $(p^*,q^*)$-rule, with $q^* = 100$ and

$$p^* := 100 \frac{p}{q} \tag{4.7}$$

Of course we can also adapt the $(n,k)$-dominance rule to account for $q$% relative a priori bounds: Let e.g. $n = 2$. According to (4.3) above, an aggregate should then be considered unsafe when the second largest respondent could estimate the largest contribution $x_1$ to within $(100 - k)$ percent of $X$ , by subtracting her own contribution and the estimate $\hat{X}_R$ from the cell total, *i.e.* when $|\,(X - x_2) - x_1 - \hat{X}_R| < (100 - k)/100X$. Just as in the case of the $p$%-rule, we see that the aggregate is sensitive, when $X_R < \frac{100-k}{q} X$, and that for

given parameters $k$ and $q$ the parameter $k^*$ corresponding to $q^* = 100$ should be chosen such that

$$\frac{100 - k^*}{100} = \frac{100 - k}{q} \tag{4.8}$$

For a more analytical discussion of sensitivity rules the interested reader is referred to (Cox, 2001), for more generalized formulations considering coalitions to (Loeve, 2001).

***Negative contributions***

When disclosure risk has to be assessed for a variable that can take not only positive, but also negative values, we suggest to reduce the value of $p$ (or increase $k$, for the dominance-rule, resp.). It may even be adequate to take that reduction even to the extent of replacing a concentration rule by a minimum frequency rule. This recommendation is motivated by the following consideration. Above, we have explained that the $p$%-rule is equivalent to a $(p, q)$-rule with $q = 100$. When contributions may take negative, as well as positive values, it makes sense to assume that the bound $q_-$ for the relative deviation of *a priori* estimates $\hat{X}_R$ exceeds 100%. This can be expressed as $q_- = 100f$, with $f > 1$. According to (4.7) the $p$ parameter $p_f$ for the case of negative/positive data should be chosen as $p_f = 100\frac{p}{q_-} = 100\frac{p}{100f} = \frac{p}{f} < p$. That means particularly that for large $f$ the $p$%-rule with corresponding parameter $p_f$ is asymptotically equal to the minimum frequency rule (c.f. the remark just below Table 4.3).

In case of the dominance rule, because of (4.8), $k_f$ can be determined by $\frac{100-k_f}{100} = \frac{100-k}{100f}$, and because $f > 1$ this means that $k_f > k$. For large $f$ the dominance rule with parameter $k_f$ will be asymptotically equal to the minimum frequency rule.

***Waivers***

Sometimes, respondents authorize publication of an aggregate even if this publication might cause a risk of disclosure for their contributions. Such authorizations are also referred to as *'waivers'*. When $s$ is the largest respondent from which no such waiver has been obtained, and $r$ is the largest respondent except for $s$ then for any pair $(i, j)$, $i \neq j$ of respondents, where no waiver has been obtained from $j$, it holds $x_i + x_j \leq x_r + x_s$. Therefore, we will be able to deal with such a situation properly, if we reformulate the concentration rules as

$X - x_s - x_r < \frac{p}{100}x_s$, for the $p$%-rule,

and

$x_r + x_s > \frac{k}{100}X$, for the $(2, k)$-dominance rule.

***Foreign trade rules***

In foreign trade statistics traditionally a special rule is applied. Only for those enterprises that have actively asked for protection of their information special sensitivity rules are applied. This implies that if the largest contributor to a cell asks for protection and contributes over a certain percentage to the cell total, that cell is considered a primarily unsafe cell.

Given the normal level of detail in the foreign trade tables the application of the standard sensitivity rules would imply that a very large proportion of the table should have been suppressed. This is considered not to be a desirable situation and for a long time there have not been serious complaints. This has led to this special rule. And this rule has a legal basis in the European regulation 638/2004.

In $\tau$-ARGUS special options (*the request rule*) have been introduced to apply this rule. The secondary cell suppression will be done in the standard way.

***Holdings***
In many datasets, especially economic datasets, the reporting unit is different from the entities we want to protect. Often companies have several branches in various regions. So the marginal cell may have several contributions but from one company only. Therefore it might be a mistake to think that such a cell is safe if we look only at the largest two contributors, while the largest three might belong to only one company.

We need to group the contributions from one company together to one contribution before we can apply the sensitivity rules.

There is no problem in making the cell totals because we are only interested in the sum of the contributions.

In $\tau$-ARGUS it is possible to take this into account by using the holding option.

***Sampling weights***
Often tables are created from datasets based on sample surveys and NSIs collect a lot of their information this way. There are two reasons why disclosure risk assessment may also be necessary for a sample survey: especially in business surveys the common approach is to sample with unequal probabilities. The first is that large companies are often sampled with probability 1. Typically it will be the case for some of the aggregates foreseen for a publication that all respondents have been sampled with probability 1. In the absence of non-response, for such an aggregate, the survey data are data for the full population. Secondly even, if sampling probabilities are smaller than 1, if an aggregate relates to a strongly skewed population, and the sampling error is small, then the probability is high that the survey estimate for the aggregate total may also be a close estimate for the largest unit in that population. It makes sense therefore to assess the disclosure risk for survey sample estimates, if the sampling errors are small but the variables are strongly sensitive. The question is then, how to determine technically, if a sample survey aggregate should be considered safe, or unsafe. The following procedure is used in $\tau$-ARGUS:

For sample survey data each record has a sampling weight associated with it. These sampling weights are constructed such that the tables produced from this datasets will resemble as much as possible the population, as if the table had been based on a full census.

Making a table one has to take into account these sampling weights. If the data file has a sampling weight, specified in the metadata, the table can be computed taking these weights into account. For making a cell total this is a straightforward procedure, however the sensitivity rules have to be adapted to the situation where we do not know the units of the population with the largest values.

One option is the following approximation:

**Example** A cell with two contributions:
*100, weight 4*
*10, weight 7*
The weighted cell value then equals $(4 \times 100) + (7 \times 10) = 470$. Without considering the weights there are only two contributors to the cell: one contributing 100 and another contributing 10. However by taking account of the sampling weights the individual contributions can be approximated as 100, 100, 100, 100, 10, 10, 10, 10, 10, 10 and 10. The largest two contributors are now 100 and 100. These are regarded as the largest two values for application of the safety rules. If the weights are not integers, a simple extension is applied.

*Note:*
This procedure cannot be applied in combination with the holding concept, because naturally for the approximate contributions it is unknown which holding they belong to.

It should also be noted here that with the above procedure it may happen that a cell is considered safe, even though the sample estimate of the cell total provides a very close estimate of the unit in the population with the largest value. Assume *e.g.* the sequence of values in the full population to be 900, 100, 100, and seven times 5. Then the population total would be 1135. Assume that we sampled two units, both with a value of 100, but with different probabilities so that the sampling weights for the two are 1, and 9, respectively, yielding a population estimate of 1000 which will not be sensitive according to the above procedure. If the second largest unit as an intruder subtracts her own contribution (= 100) from the population estimate she will obtain an estimate of the value of the largest unit which exactly matches the true value of 900. But there is of course a sampling error connected to this estimate and this fact should be considered by the intruder.

On the other hand, there is a certain tendency for overprotection associated to the above procedure, *i.e.* there is a chance that according to the above procedure an aggregate is unsafe, when in fact it is not. Assume *e.g.* the sequence of values in the full population to be 100, 50, 50, 1, 1. Assume we have sampled (with unequal probabilities) the unit with value 100 with a sampling weight of 2 associated to it, and one of the units with value 1

with a sampling weight of 3. According to the above procedure two values of 100 would be considered as largest contribution for an estimated cell total of 203 which according to for instance a $p$%-rule would be considered as sensitive for any $p > 3$ . But in fact, even the second largest unit in the population with a value of 50 will overestimate the value 100 of the largest unit by about 50% when it subtracts her own contribution from the population estimate of 203.

This tendency for overprotection can be avoided, when, instead of the above procedure, in the case of sample surveys we replace for instance for the $p$%-rule the formulation (4.2) of Table 4.3 by the following

$$\hat{X} - x_2^s - x_1^s < \frac{p}{100} x_1^s \tag{4.9}$$

where $\hat{X}$ denote the estimated population total, and $x_i^s$, $i = 1, 2$, the largest two observations from the sample.

The difference between this, and the above procedure is that with (4.9) we consider as intruder the respondent with the second largest contribution observed in the sample, whereas according to the above procedure, whenever the sampling weight of the largest respondent is 2 or more, an 'artificial' intruder is assumed who contributes as much as the largest observation from the sample.

Considering formulation (4.3) of the dominance-rule, it is straightforward to see that we can adapt (4.9) to the case of dominance-rules by

$$x_1^s + ... + x_n^s > \frac{k}{100} \hat{X} \tag{4.10}$$

It is also important to note in this context that sampling weight should be kept confidential, because otherwise we must replace (4.9) by

$$\hat{X} - w_2 x_2^s - w_1 x_1^s < \frac{p}{100} x_1^s \tag{4.11}$$

where $w_i$, $i = 1, 2$, denote the sampling weights. Obviously, according to (4.11) more aggregates will be sensitive.

Considering this as the basic idea of secondary protection, after Section 4.2.1 we assume for the remainder of the chapter margins and overall totals always to be part of a table.

In Section 4.2.2 we introduce into the concepts of secondary tabular data protection methodologies.

### 4.2.2 Secondary tabular data protection methods

For reasons explained in the introduction of this chapter, we assume here that tables always include margins and overall totals along with their 'inner' cells. Thus, there is

always a linear relationship between cells of a table. (Consider for instance example 1a of the introduction to this chapter: The sector result (say: $X_T$) is the sum of the three sub-sector results ($X_1$ to $X_3$), i.e. the linear relationship between these four cells is given by the relation $X_T = X_1 + X_2 + X_3$. As we have seen in example 1a, if it has been established that a disclosure risk is connected to the release of certain cells of a table, then it is not enough to prevent publication of these cells. Other cells must be suppressed (so called 'complementary' or 'secondary' suppressions), or be otherwise manipulated in order to prevent the value of the protected sensitive cell being recalculated through f.i. differencing.

Within this section we explain the methodological background of secondary tabular data protection methods.

At the end of the section we give a brief comparison of secondary cell suppression, partial suppression and controlled tabular adjustment as alternative disclosure limitation techniques.

⚠ Expert level

When a table is protected by cell suppression, by making use of the linear relation between published and suppressed cell values in a table with suppressed entries, it is always possible for any particular suppressed cell of a table to derive upper and lower bounds for its true value. This holds for either tables with non-negative values, and those tables containing negative values as well, when it is assumed that instead of zero, some other (possibly tight) lower bound for any cell is available to data users in advance of publication. The interval given by these bounds is called the *'feasibility interval'*.

**Example 3** This example (Geurts, 1992, Table 10, p 20) illustrates the computation of the feasibility interval in the case of a simple two-dimensional table where all cells may only assume non-negative values:

Table 4.4: Example 3

| **Example** | **1** | **2** | **Total** |
|---|---|---|---|
| **1** | $X_{11}$ | $X_{12}$ | 7 |
| **2** | $X_{21}$ | $X_{22}$ | 3 |
| **3** | 3 | 3 | 6 |
| **Total** | 9 | 7 | 16 |

For this table the following linear relations hold:

$$X_{11} + X_{12} = 7 \quad \text{(R1)} \tag{4.12}$$
$$X_{21} + X_{22} = 3 \quad \text{(R2)} \tag{4.13}$$
$$X_{11} + X_{21} = 6 \quad \text{(C1)} \tag{4.14}$$
$$X_{12} + X_{22} = 4 \quad \text{(C2)} \tag{4.15}$$
$$\tag{4.16}$$

with $X_{ij} \geq 0, \forall(i,j)$. Using linear programming methodology, it is possible to derive systematically for any suppressed cell in a table an upper bound $(X^{\max})$ and a lower bound $(X^{\min})$ for the set of feasible values. In the example above, for cell (1,1) these bounds are $(X_{11}^{\min}) = 3$ and $X_{11}^{\max} = 6$ . In this simple instance, however, we do not need linear programming technology to derive this result: Because of the first column relation (C1) $X_{11}$ must be less or equal 6, and because of the second row relation (R2) $X_{21}$ must be less or equal 3. Therefore, and because of the first column relation (C1) $X_{11}$ must be at least 3.

A general mathematical statement for the linear programming problem to compute upper and lower bounds for the suppressed entries of a table is given in Fischetti and Salazar (2000).

In principle, a set of suppressions (the *'suppression pattern'*) should only be considered valid, if the bounds for the feasibility interval of any sensitive cell cannot be used to deduce bounds on an individual respondent contribution contributing to that cell that are too close according to the sensitivity rule employed. For a mathematical statement of that condition, we determine safety bounds for primary suppressions. We call the deviation between those safety bounds and the true cell value *'upper and*

*lower protection levels'*. The formulas of Table 4.5 can be used to compute upper protection levels. Out of symmetry considerations the lower protection level is often set identical to the upper protection level.

Table 4.5: Upper protection levels

| Sensitivity rule | Upper protection level |
|---|---|
| $(1, k)$-rule | $\frac{100}{k} x_1 - X$ |
| $(n, k)$-rule | $\frac{100}{k} \cdot (x_1 + x_2 + \cdots + x_n) - X$ |
| $p\%$-rule | $\frac{p}{100} \cdot x_1 - (X - x_1 - x_2)$ |
| $(p, q)$-rule | $\frac{p}{q} \cdot x_1 - (X - x_1 - x_2)$ |

Note that we recommend using protection levels of zero when instead of a concentration rule only a minimum frequency rule has been used for primary disclosure risk assessment. As explained in 4.2.1, minimum frequency rules (instead of concentration rules) should only be used, if it is enough to prevent exact disclosure only. And in such a case, a protection level of zero should be enough. Using $\tau$-ARGUS this can be achieved by setting the parameter 'minimum frequency range' to zero.
If the distance between upper bound of the feasibility interval and true value of a sensitive cell is below the upper protection level computed according to the formulas of Table 4.5, then this upper bound could be used to derive an estimate for individual contributions of the sensitive cell that is too close according to the safety rule employed, which can easily be proven along the lines of Cox (1981).

**Example 4** Cell $X = 330$ with 3 contributions of distinct respondents $x_1 = 300,\ x_2 = 20,\ x_3 = 10$ is confidential (or 'unsafe') according to the $(1, 85)$-dominance rule. If the feasible upper bound $X^{\max}$ for this confidential cell value is less than $\frac{100}{85} \cdot x_1 = 352.94$ , then it will provide an upper estimate for the largest contribution $x_1$ which is too close according to the $(1, 85)$-dominance rule.

Example 5 proves the formula given in Table 4.5 for the case of the $p$%-rule:

**Example 5** Let $X + U$ be the upper bound of the feasibility interval for a cell with cell value $X$ . The second largest respondent can then deduce an upper bound $x_1^U$ by subtracting its own contribution from that upper bound: $x_1^U = X + U - x_2$. According to the definition of the $p$%-rule proper protection means that $x_1^U \geq \left(1 + \frac{p}{100}\right) \cdot x_1$. So, if the feasibility interval provides upper protection, it must hold that $U \geq \left(1 + \frac{p}{100}\right) \cdot x_1 - X + x_2 = p/100 \cdot x_1 - (X - x_1 - x_2)$.

The distance between upper bound of the feasibility interval and true value of a sensitive cell must exceed the upper protection level; otherwise the sensitive cell is not properly protected. This safety criterion is a necessary, but not always sufficient criterion for proper protection! It is a sufficient criterion, when the largest respondent makes the same contribution also within the combination of suppressed cells within the same aggregation (a row or column relation of the table, for instance), and when no individual contribution of any respondent (or coalition of respondents) to such a combination of suppressed cells is larger than the second largest respondent's (or coalition's) contribution.
Cases where the criterion is not sufficient arise typically, when the only other suppressed cell within the same aggregation is a sensitive cell too (and not the marginal cell of the aggregation), or when the same respondent can contribute to more than one cell within the same aggregation. In these cases, it may turn out that the in-

dividual contribution of a respondent may still be disclosed, even though the upper bound of the feasibility interval is well away from the value of the sensitive cell to which this respondent contributes. The most prominent case is that of two cells with only one single contributor (a.k.a. *'singletons'*) within the same aggregation. No matter how large the cell values, because they both know their own contribution of course, both can use this additional knowledge to disclose the other's contribution. For an analytical discussion of these issues see Cox(2001). In principle, a suppression pattern should not be considered as valid, when certain respondents can use their insider knowledge (on their own contribution to a cell) to disclose individual contributions of other respondents.

The problem of finding an optimum set of suppressions known as the 'secondary cell suppression problem' is to find a valid set of secondary suppressions with a minimum loss of information connected to it. For a mathematical statement of the secondary cell suppression problem see e.g. Fischetti and Salazar (2000).

However, cell suppression is not the only option to protect magnitude tables. As one alternative to cell suppression, *'Partial Suppression'* has been suggested in Salazar (2003). The partial suppression problem is to find a valid suppression pattern, where the size of the feasibility intervals is minimal. The idea of the partial suppression method is to publish the feasibility intervals of the resulting suppression pattern. Because there is a perception that the majority of users of the publications prefer the statistical agencies to provide actual figures rather than intervals, and partial suppression tends to affect more cells than cell suppression, this method has not yet been implemented.

Another alternative method for magnitude table protection is known as *'Controlled Tabular Adjustment'* (CTA) suggested for instance in Cox and Dandekar (2002), Castro (2003), or Castro and Giessing (2006). CTA methods attempt to find the closest table consistent with the constraints imposed by the set of table equations and by the protection levels for the sensitive cells, of course taking into account also bounds on cell values available to data users in advance of publication, like non-negativity. This adjusted table would then be released instead of the original table.

## 4.3 The $\tau$-ARGUS Implementation of Cell Suppression

The software package $\tau$-ARGUS provides software tools for disclosure protection methods for tabular data. This can be achieved by modifying the table so that it contains less, or less detailed, information. $\tau$-ARGUS allows for several modifications of a table: a table can be redesigned, meaning that rows and columns can be combined; sensitive cells can be suppressed and additional cells to protect these can be found in some optimal way (secondary cell suppression). Several alternative algorithms for the selection of secondary suppressions are available in $\tau$-ARGUS, *e.g.* the Hypercube/GHMiter method, a Modular

and a Full optimal solution, and a method based on a network flow algorithm. Instead of cell suppression one could also use other methods for disclosure control of tables. One of these alternative methods is Controlled Rounding. However, use of Controlled Rounding is more common for frequency tables.

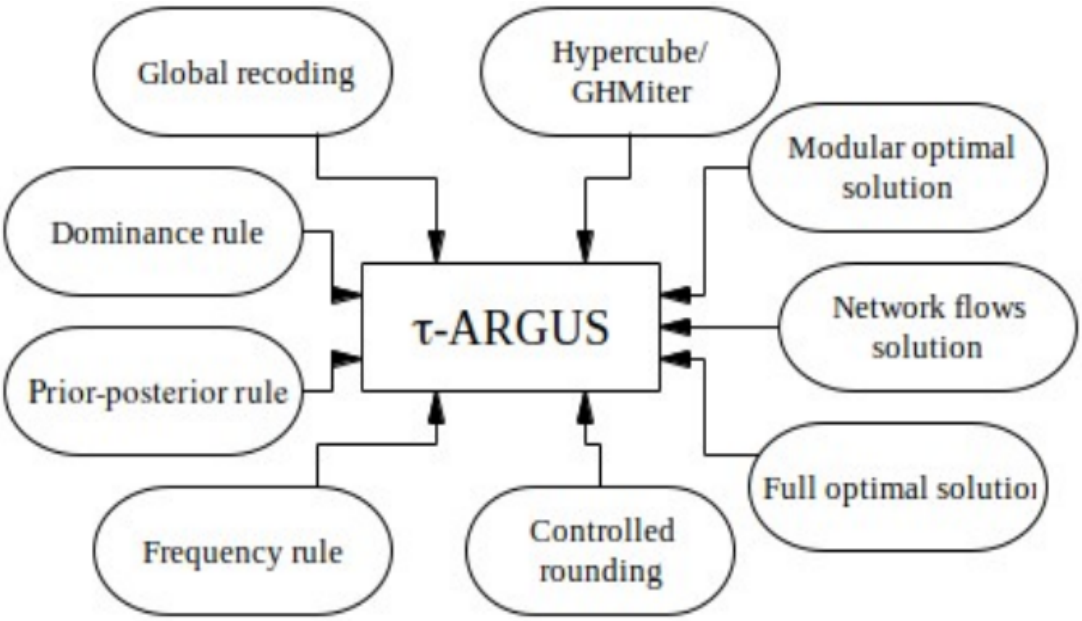


Figure 4.1: Overview of the $\tau$-ARGUS architecture

We start this section by introducing the basic issues related to the question of how to properly set up a tabular data protection problem for a given set of tables that are intended to be published, discussing tabular data structures, and particularly tables with hierarchical structures. Section 4.3.2 provides an evaluation of algorithms for secondary cell suppression offered by $\tau$-ARGUS, including a recommendation on which tool to use in which situation. After that, in Section 4.3.3 we give some guidance on processing table protection efficiently, especially in the case of linked tables. We finish the section with an introductive example in Section 4.3.4.

### 4.3.1 Setting up a Tabular Data Protection Problem in Practice

At the beginning of this section we explain the basic concept of tables in the context of tabular data protection, and complex structures like hierarchical and linked tables. In order to find a good balance between protection of individual response data and provision of information – in other words, to take control of the loss of information that obviously cannot be avoided completely because of the requirements of disclosure control, it is necessary to somehow rate the information loss connected to a cell that is suppressed. We

therefore complete this section by explaining the concept of cell costs to rate information loss.

***Table specification***

For setting up the mathematical formulation of the protection problems connected to the release of tabulations for a certain data set, all the linear relations between the cells of those tabulations have to be considered. This leads us to a crucial question: What is a table anyway? In the absence of confidentiality concerns, a statistician creates a table in order to show certain properties of a data set, or to enhance comparison between different variables. So a single table might literally mix apples and oranges. Secondly, statisticians may wish to present a number of those 'properties', publishing multiple tables from a particular data set. Where does one table end and the next start? Is the ideal table one that fits nicely on a standard-size sheet of paper? With respect to tabular data protection, we have to think of tables in a different way:

Magnitude tables display sums of observations of a quantitative variable, the so-called '*response variable*'. The sums are displayed across all observations and/or within groups of observations. These groups can be formed by grouping the respondents according to certain criteria such as their economic activity, region, size class of turnover, legal form, etc. In that case the grouping of observations is defined by categorical variables observed for each individual respondent, the 'explanatory variables'. It also occurs that observations are grouped by categories of the response variable, for instance fruit production into apples, pears, cherries, etc.

The "*dimension*" of a table is given by the number of grouping variables used to specify the table. "*Margins*" or "*marginal cells*" of a table are those cells, which are specified by a smaller number of grouping variables. The smaller the number of grouping variables, the higher the "*level*" of a marginal cell. A two-dimensional table of some business survey may for instance provide sums of observations grouped by economic activity and company size classes. At the same time it may also display the sums of observations grouped by only economic activity or by only size classes. These are then margins/marginal cells of this table. If a sum across all observations is provided, we refer to it as the "*total*" or "*overall total*".

Note that tables presenting ratios or indexes typically do not define a tabular data protection problem, because there are no additive structures between the cells of such a table, and neither between a cell value, and the values of the contributing units. Think, for instance, of mean wages, where the cell values would be a sum of wages divided by a sum over the number of employees of several companies. However, there may be exceptions: if, for instance, it is realistic to assume that the denominators (say, the number of employees) can be estimated quite closely, both on the company level and on the cell level, then it might indeed be adequate to apply disclosure control methods based, for example, on the enumerator variable (in our instance: the wages).

***Hierachical, linked tables***
Data collected within government statistical systems must meet the requirements of many users, who differ widely in the particular interest they take in the data. Some may need community-level data, while others need detailed data on a particular branch of the economy but no regional detail. As statisticians, we try to cope with this range of interest in our data by providing the data at several levels of detail. We usually combine explanatory variables in multiple ways, when creating tables for publication. If two tables presenting data on the same response variable share some categories of at least one explanatory variable, there will be cells which are presented in both tables – those tables are said to be *linked* by the cells they have in common. In order to offer a range of statistical detail, we use elaborate classification schemes to categorize respondents. Thus, a respondent will often belong to various categories of the same classification scheme - for instance a particular community within a particular county within a particular state - and may thus fall into three categories of the regional classification.

The structure between the categories of hierarchical variables also implies sub-structure for the table. When, in the following, we talk about sub-tables without substructure, we mean a table constructed in the following way:

For any explanatory variable we pick one particular non-bottom-level category (the 'food production sector' for instance). Then we construct a 'sub-variable'. This sub-variable consists only of the category picked in the first step and those categories of the level immediately below belonging to this category (bakers, butchers, etc.). After doing that for each explanatory variable the table specified through a set of these sub-variables is free from substructure then, and is a sub-table of the original one. Any cell within the sub-table does also belong to the original table. Many cells of the original table will appear in more than one sub-table: The sub-tables are linked. Example 6 provides a simple instance of a one-dimensional table with hierarchical structure.

**Example 6** Assume the categories A, AA, AB, AB1, AB2, and AB3 of some variable EXP resemble a hierarchical structure as depicted in Figure 4.2 below:

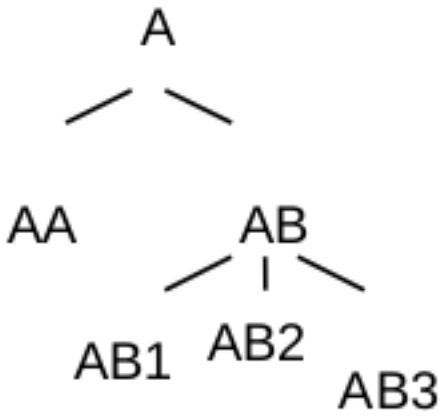


Figure 4.2: Hierarchical structure

Let the full table present turnover by the categories of EXP. Then the two subtables of this table will be:

| EXP | Turnover |
|---|---|
| **A** | |
| **AA** | |
| **AB** | |

| EXP | Turnover |
|---|---|
| **AB** | |
| **AB1** | |
| **AB2** | |

Note that cell **AB** appears in both subtables

Of course, when using cell suppression methods to protect a set of linked tables, or subtables, it is not enough to select secondary suppressions for each table (or sub-table) separately. Otherwise it might for instance happen that the same cell is suppressed in one table because it is used as secondary suppression, while within another table it remains unsuppressed. A user comparing the two tables would then be able to disclose confidential cells in the first table. A common approach is to protect tables separately, but note any complementary suppression belonging also to one of the other tables; suppress it in this table as well, and repeat the cell suppression procedure for this table. This approach is called a '*backtracking procedure*'. Although within a backtracking process for a hierarchical table the cell-suppression procedure will usually be repeated several times for each subtable, the number of computations required for the process will be much smaller than

when the entire table is protected all at once.

It must however be stressed, that a backtracking procedure is not global according to the denotation in Cox (2001). See Cox (2001) for discussion of problems related to non-global methods for secondary cell suppression.

For mathematical formulation of linked tables structures see de Wolf (2007) and de Wolf, Giessing (2008).

***Cell costs***

The challenge of tabular data protection is to preserve as much information in the table as possible, while creating the required uncertainty about the true values of the sensitive cells, as explained in the previous section. It is necessary to somehow rate the information content of data, in order to be able to express the task of selecting an 'optimal set' of secondary suppressions, or, alternatively, of adjusting a table in an optimal way, as mathematical programming problem. Information loss is expressed in these mathematical models as the sum of costs associated to the secondary suppressions, or non-sensitive cells subject to adjustment.

For cell suppression, the idea of equating a minimum loss of information with the smallest number of suppressions is probably the most natural concept. This would be implemented technically by assigning identical costs to each cell. Yet experience has shown that this concept often yields a suppression pattern in which many large cells are suppressed, which is undesirable. In practice other cost functions, based on cell values, or power transformations thereof, or cell frequencies yield better results. Note that several criteria, other than the numeric value, may also have an impact on a user's perception of a particular cells importance, such as its situation within the table (totals and sub-totals are often rated as highly important), or its category (certain categories of variables are often considered to be of secondary importance). $\tau$-ARGUS offers special facilities to choose a cost function and to carry out power transformations of the cost function:

***Cost function and $\lambda$***

The cost function indicates the relative importance of a cell. In principle the user could attach a separate cost to each individual cell, but that will become a rather cumbersome operation. Therefore in $\tau$-ARGUS a few standard cost functions are available[7]:

- Unity, i.e. all cells have equal weight (just minimising the number of suppressions)

[7] One of the secondary cell suppression methods of $\tau$-ARGUS, the hypercube method, uses an elaborate internal weighting scheme in order to avoid to some extent the suppression of totals and subtotals which is essential for the performance of the method. In order to avoid that this internal weighting scheme is thrown out of balance, it has been decided to make the cost function facilities of t-ARGUS inoperational for this method.

- The number of contributors (minimising the number of contributors to be hidden)
- The cell value itself (preserving as much as possible the largest – usually most important – cells)
- The cell value of another variable in the dataset (this will be the indication of the important cells).

Especially in the latter two cases the cell value could give too much preference to the larger cells. Is a 10 times bigger cell really 10 times as important? Sometimes there is a need to be a bit more moderate. Some transformation of the cost function is desired. Popular transformations are square-root or a log-transformation. Box and Cox (1960) have proposed a system of power transformations:

$$y = \frac{x^\lambda - 1}{\lambda}$$

In this formula $\lambda = 0$ yields to a log-transformation and the $-1$ is needed for making this formula continuous in $\lambda$.

This last aspect and the fact that this could lead to negative values we have introduced a simplified version of the lamda-transformation in $\tau$-ARGUS

$$y = \begin{cases} x^\lambda & \text{for } \lambda \neq 0 \\ \log(x) & \text{for } \lambda = 0 \end{cases}$$

So choosing $\lambda = \frac{1}{2}$ will give the square-root.

### 4.3.2 Evaluation of Secondary Cell Suppression algorithms offered by $\tau$-ARGUS

The software package $\tau-$ARGUS offers a variety of algorithms to assign secondary suppressions which will be described in detail in Section 4.4. While some of those algorithms are not yet fully developed, or not yet fully integrated into the package, respectively, two are indeed powerful tools for automated tabular data protection. In this section we focus on those latter algorithms. We compare them with respect to quality aspects of the protected tables, while considering also some practical issues. Based on this comparison, we finally give some recommendation on the use of those algorithms.

We begin this section by introduction of certain minimum protection standards PS1, and PS2. We explain why it is a 'must' for an algorithm to satisfy those standards in order to be useable in practice. For those algorithms that qualify 'useable' according to these criteria, we report some results of studies comparing algorithm performance with regard to information loss, and disclosure risk.

In Section 4.2.2 we have defined criteria for a safe suppression pattern. It is of course possible to design algorithms that will ensure that no suppression pattern is considered feasible unless all of these criteria are satisfied. In practice, however, such an algorithm would either require too much computer resource (in particular: CPU time) when applied to the large, detailed tabulations of official statistics, or the user would have to put up with a strong tendency for oversuppression. In the following, we therefore define minimum standards for a suppression pattern that is safe w.r.t. the safety rule adopted except for a residual disclosure risk that might be acceptable to an agency. It should be noted that those standards are definitely superior, or at least equal to what could be achieved by the traditional manual procedures for cell suppression, even if carried out with utmost care! For the evaluation, we consider only those algorithms that ensure validity of the suppression pattern at least with respect to those relaxed standards.

***Minimum protection standards***

For a table with hierarchical substructure, feasibility intervals computed on basis of the set of equations for the full table normally tend to be much closer than those computed on basis of separate sets of equations corresponding to sub-tables without any substructure. Moreover, making use of the additional knowledge of a respondent, who is the single respondent to a cell (a so called *'singleton'*), it is possible to derive intervals that are much closer than without this knowledge[8].

Based on the assumption of a simple but probably more realistic disclosure scenario where intruders will deduce feasibility intervals separately for each sub-table, rather than taking the effort to consider the full table, and that single respondents, using their special additional knowledge, are more likely attempting to reveal other suppressed cell values when they are within the same row or column (more general: relation), the following minimum protection standards (PS) make sense:

(PS1) *Protection against exact external disclosure*:
With a valid suppression pattern, it is not possible to disclose the value of a sensitive cell exactly, if no additional knowledge (like that of a singleton) is considered, and if subsets of table equations are considered separately, i.e. when feasibility intervals are computed separately for each sub-table.

(PS2) *Protection against singleton disclosure*:
A suppression pattern, with only two suppressed cells within a row (or column) of a table is not valid, if each of the two corresponds to a single respondent who are not identical.

(PS1*) extension of (PS1) for inferential (instead of exact) disclosure,

[8] Consider for instance the table of example 3 in Section 4.2.2. Without additional knowledge the bounds for cell (1,1) are $X_{11}^{\min} = 3$ and $X_{11}\max = 6$. However, if cell (1,2) were a singleton-cell (a cell with only a single contributor) then this single contributor could of course exactly compute the cell value of cell (1,1) by subtracting her own contribution from the row total of row 1.

(PS2*) extension of (PS2), covering the more general case where a single respondent can disclose another single respondent cell, not necessarily located within the same row (or column) of the table.

$\tau$-ARGUS offers basically two algorithms, Modular and Hypercube, both satisfying the above-mentioned minimum standards regarding disclosure risk. Both take a backtracking approach to protect hierarchical tables within an iterative procedure: they subdivide hierarchical tables into sets of linked, unstructured tables. The cell suppression problem is solved for each subtable separately. Secondary suppressions are coordinated for overlapping subtables.

The Modular approach, a.k.a. HiTaS algorithm (see de Wolf, 2002), is based on the optimal solution as implemented by Fischetti and Salazar-González. By breaking down the hierarchical table in smaller non-structured subtables, protecting these subtables and linking the subtables together by a backtracking procedure the whole table is protected. For a description of the modular approach see **?@sec-modular**, the optimal Fischetti/Salazar-González method is described in section Section 4.4.1. See also Fischetti and Salazar-González (2000).

The Hypercube is based on a hypercube heuristic implemented by Repsilber (2002). See also the descriptions of this method in Section 4.4.4.

Note that in the current implementation of $\tau$-ARGUS, use of the optimization tools requires a license for additional commercial software (LP-solver), whereas use of the hypercube method is free. While Modular is only available for the Windows platform, of Hypercube there are also versions for Unix, IBM (OS 390) and SNI (BS2000).

Both Modular and Hypercube provide sufficient protection according to standards PS1* (protection against inferential disclosure) and PS2 above. Regarding singleton disclosure, Hypercube even satisfies the extended standard PS2*. However, simplifications of the heuristic approach of Hypercube cause a certain tendency for oversuppression. This evaluation therefore includes a relaxed variant that, instead of PS1*, satisfies only the reduced standard PS1, i.e. zero protection against inferential disclosure. We therefore refer to this method as Hyper0 in the following. Hyper0 processing can be achieved simply by deactivating the option “Protection against inferential disclosure required” when running Hypercube out of $\tau$-ARGUS.

| Algorithm | | Modular | Hypercube | Hyper0 |
|---|---|---|---|---|
| Procedure for secondary suppression | | Fischetti/Salazar optimization | Hypercube heuristic | |
| Protection standard | Interval/Exact discolsure | PS1* | PS1* | PS1 |
| | Singleton | PS2 | PS2* | PS2 |

Table 4.8: Algorithms for practical use

In the following, we briefly report results of two evaluations studies, in the following referred to as study A, and study B. For further detail on these studies see Giessing (2004) for study A, and Giessing et al. (2006) for study B. Both studies are based on tabulations of a strongly skewed magnitude variable.

For study A, we used a variety of hierarchical tables generated from the synthetic micro-data set supplied as CASC deliverable 3-D3. In particular, we generated 2- and 3-dimensional tables, where one of the dimensions had a hierarchical structure. Manipulation of the depth of this hierarchy resulted in 7 different tables, with a total number of cells varying between 460 and 150,000 cells. Note that the Hyper0 method was not included in study A.

For study B, which was carried out on behalf of the German federal and state statistical offices, we used 2-dimensional tables of the German turnover tax statistics. Results have been derived for tables with 7-level hierarchical structure (given by the NACE economy classification) of the first dimension, and 4-level structure of the second dimension given by the variable Region.

We compare the loss of information due to secondary suppression in terms of the number and cell values of secondary suppression, and analyze the disclosure risk of the protected tables. Table 4.9 reports the results regarding the loss of information obtained in study A.

| **Table Hier. levels** | | **No. Cells** | **No. Suppressions (%)** | | **Added Value of Suppressions (%)** | |
|---|---|---|---|---|---|---|
| | | | **Hypercube** | **Modular** | **Hypercube** | **Modular** |
| | | | 2-dimensional tables | | | |
| 1 | 3 | 460 | 6.96 | 4.35 | 0.18 | 0.05 |
| 2 | 4 | 1,050 | 10.95 | 8.29 | 0.98 | 0.62 |
| 3 | 6 | 8,230 | 14.92 | 11.48 | 6.78 | 1.51 |
| 4 | 7 | 16,530 | 14.97 | 11.13 | 8.24 | 2.12 |
| | | | 3-dimensional tables | | | |
| 5 | 3 | 8,280 | 14.63 | 10.72 | 6.92 | 1.41 |
| 6 | 4 | 18,900 | 17.31 | 15.41 | 12.57 | 3.55 |
| 7 | 6 | 148,140 | 15.99 | 10.63 | 23.16 | 3.91 |

Table 4.9: Results of study A - Number and value of secondary suppressions

Table 4.10 presents information loss results from study B on the two topmost levels of the second dimension variable Region, e.g. on the national, and state level.

| Algorithm for secondary suppression | Number | | | | Added Value overall |
|---|---|---|---|---|---|
| | State level | | National level | | |
| | abs | % | abs | % | % |
| Modular | 1,675 | 10.0 | 7 | 0.6 | 2.73 |
| Hyper0 | 2,369 | 14.1 | 8 | 0.7 | 2.40 |
| Hypercube | 2,930 | 17.4 | 22 | 1.9 | 5.69 |

Table 4.10: Results of study B - Number and value of secondary suppressions

Both studies show that best results are achieved by the modular method, while using either variant of the hypercube method leads to an often quite considerable increase in the amount of secondary suppression compared to the result of the modular method. The size of this increase varies between hierarchical levels of the aggregation: on the higher levels of a table the increase tends to be even larger than on the lower levels. Considering the results of study B presented in Table 4.10, on the national level we observe an increase of about 214% for Hypercube (14% for Hyper0), compared to an increase of 75% (41%) on the state level. In Giessing et al. (2006) we report also results on the district level. On that level the increase was about 28% on the higher NACE levels for Hypercube (9% for Hyper0), and 20 % for Hypercube (14% for Hyper0) on the lower levels. In study A we observed increases mostly around 30% for Hypercube, for the smallest 2-dimensional, and the largest 3-dimensional instance even of 50% and 60%, respectively. In particular, our impression is that the hypercube method tends to behave especially badly (compared to Modular), when tabulations involve many short equations, with only a few positions.

A first conclusion from the results of study B reported above was, because of the massive oversuppression, to exclude the Hypercube method in the variant that prevents inferential disclosure from any further experimentation. Although clearly only second-best performer, testing with Hyper0 was to be continued, because of technical and cost advantages mentioned above.

In experiments with state tables where the variable Region was taken down to the community level, we found that this additional detail in the table caused an increase in suppression at the district level of about 70 – 80% with Hyper0 and about 10 – 30% with Modular.

***Disclosure risk***

As explained in Section 4.2.2, in principle, there is a risk of inferential disclosure, when the bounds of the feasibility interval of a sensitive cell could be used to deduce bounds on an individual respondent's contribution that are too close according to the method employed for primary disclosure risk assessment. However, for large, complex structured tables, this risk is rather a risk potential, not comparable to the disclosure risk that would result from a publication of that cell. After all, the effort connected to the computation of feasibility intervals based on the full set of table equations is quite considerable. Moreover, in a part of the disclosure risk cases only insiders (other respondents) would actually be able to disclose the individual contribution.

Anyway, with this definition, and considering the full set of table equations, in study B we found between about 4% (protection by Modular) and about 6% (protection by Hyper0) of sensitive cells at risk in an audit step where we computed the feasibility intervals for a tabulation by NACE and state, and for two more detailed tables (down to the district level) audited separately for two of the states. In study A, we found up to about 5.5% (1.3%) cells at risk for the 2-dimensional tables protected by Modular (Hypercube, resp.), and for the 3-dimensional tables up to 2.6% (Modular) and 5.6% (Hypercube).

Of course, the risk potential is much higher for at-risk cells when the audit is carried out separately for each subtable without substructure, because the effort for this kind of analysis is much lower. Because Modular protects according to PS1* standard, there are no such cells in tables protected by Modular. For Hyper0, in study B we found about 0.4% cells at risk, when computing feasibility intervals for several subtables of a detailed tabulation.

For those subtables, when protected by Modular, for about 0.08% of the singletons, it turned out that another singleton would be able to disclose its contribution. Hyper0 satisfies PS2* standard, therefore of course no such case was found, when the table was protected by Hyper0.

***Recommendation***
The methods Modular and Hypercube of $\tau$-ARGUS are powerful tools for secondary cell suppression. Concerning protection against external disclosure both satisfy the same protection standard. However, Modular gives much better results regarding information loss. Even compared to a variant of Hypercube (Hyper0) with relaxed protection standard, Modular performs clearly better. Although longer computation times for this method (compared to the fast hypercube method) can be a nuisance in practice, the results clearly justify this additional effort – after all, it is possible, for instance, to protect a table with detailed hierarchical structure in two dimensions and more than 800,000 cells within just about an hour. Considering the total costs involved in processing a survey, it would neither be justified to use Hypercube (or Hyper0) only in order to avoid the costs for the commercial license which is necessary to run the optimization tools of Modular.

We also recommend use of Modular to assign secondary suppressions in 3-dimensional tables. However, when those tables are given by huge classifications, long computation times may become an actual obstacle. It took us, for instance, about 11 hours to run a 823 x 18 x 10 table (first dimension hierarchical). The same table was protected by Hyper0 within about 2 minutes. Unfortunately, replacing Modular by Hyper0 in that instance leads to about 28% increase in the number of secondary suppressions.

According to our findings, we would expect the Hypercube method (i.e. Hyper0) to be good alternative to Modular, if

- Tabulations involve more than 2 dimensions and are very large, and

- the majority of table equations (e.g. rows, columns, …) are long, involving many positions, and
- the distribution of the variable which is tabulated is rather even, not skewed, resulting in a lower rate of sensitive cells, and
- protection against inferential disclosure is not an important issue, for instance when a minimum frequency rule instead of a concentration rule is employed to assess the primary disclosure risk.

In the next section we will give some hints on what to do, if according to these conditions use of the Modular method seems to not be an option, because table sizes are too large, in order to avoid being confined to the use of Hyper or Hyper0.

Concerning protection against singleton (insider) disclosure, the Hypercube method fulfils a higher standard. However, our empirical results confirm that in practice there are very few singleton cells left unprotected by Modular against this kind of disclosure risk. We are therefore confident that the problem can be fixed easily by a specialized audit step that would not even involve linear programming methods.

### 4.3.3 Processing table protection efficiently

Within this section we give some guidance for how to facilitate disclosure analysis, and how to use the methods of $\tau$-ARGUS in an efficient way. For more tips and tricks to improve practical performance, see van der Meijden (2006).

Especially in large tables with detailed hierarchical structure we have observed that the choice of the parameter lambda ($\lambda$) may substantially affect the suppression pattern – and the results may be unexpected. Normally we would expect that constant cell costs will lead to a solution with a (relatively) small number of secondary suppressions. In hierarchical tables, however, it may turn out that given this cost function the Modular method tends to suppress cells in the margins of subtables rather than inner cells which in turn require additional suppressions in a (linked) subtable. This effect may be so strong that in the end, a $\lambda$ close to zero yielding approximately constant costs may lead to a larger number of suppressions as if a larger $\lambda$ (say: 0.5) had been chosen. For larger tables it makes therefore sense to test the outcome with different values of $\lambda$.

***Table design***
In Section 4.3.1 we already discussed how to set up tables for tabular data protection. If the concept for the publication is not yet fixed, some table-redesign may help to reduce the complexity of the problem.

***Table redesign***
If a large number of sensitive cells are present in a table, it might be an indication that

the spanning variables are too detailed. In that case one could consider combining certain rows and columns in the table. (This might not always be possible because of publication policy.) Otherwise the number of secondary cell suppressions might just be too enormous. It is a property of the sensitivity rules that a joint cell is safer than any of the individual cells. So as a result of this operation the number of unsafe cells is reduced. One can try to eliminate all unsafe combinations in this way, but that might lead to an unacceptably high information loss. Instead, one could stop at some point, and eliminate the remaining unsafe combinations by using other techniques such as cell suppression.

In practice, it may also often be considered impractical and also not adequate to do the disclosure analysis for each table of the entire publication separately. A well-known simplification is to focus the analysis on certain lead variables. These lead tables are protected with $\tau$-ARGUS as described above and the pattern found is then used for all the other tables. This strategy has the advantage that it is efficient and prevents the recalculation of suppressed cells by using the inherent relations between the different tables in the publication. The fact that for certain cells in the other tables not always all cells that should be considered unsafe according to the concentration rules will actually be considered unsafe, is considered a minor problem.

A compromise between the copying of the suppression pattern and protecting each table individually is the coordination of the primary suppressions only. This strategy is supported by the concept of shadow variables in $\tau$-ARGUS. The shadow variable is used for the lead-table and only for finding the primary unsafe cells. After that the pattern is used for each individual table itself and the secondary cell suppression is done in the traditional way.

***Shadow variables***

The whole theory of primary unsafe cells is based on the idea that certain larger contributors in a cell might be at risk and need protection. The largest contributors therefore play an important role in the sensitivity rules.

Often the cell value itself is a very good estimate of the size of the contributors. But sometimes the cell value of a table is not always the best indicator of the size of the company. E.g. if the cell value is the investment in some commodity, the largest companies do not always have the largest values. Simply applying the sensitivity rules could yield a situation in which a rather small company will dominate the cell and force it to become unsafe. If we assume that this smaller company is not very well known and visible there is no reason to protect this cell.

If we want to protect the real large companies it could be a good idea to apply the sensitivity rules on another table (*shadow variable*) with the same spanning variables (e.g. turnover) and use the pattern of primary unsafe cells. Copy the pattern to the current table and compute the secondary unsafe cells in the normal way.

In $\tau$-ARGUS we call the table where the primary unsafe cells are computed the shadow table.

Another reason for using a shadow variable could be the coordination of the pattern in different tables with the same spanning variables based on one dataset. See the remarks on periodical datasets below.

***Ignore response variable relations***
A commonly used strategy to simplify disclosure analysis is not to reflect the full linear relationship between published tabulation cells in the table protection procedure. A typical case, where this would be justified is the following: Assume we publish salaries of seasonal workers from foreign countries ($S_{\text{foreigners}}$), and also salaries of all seasonal workers ($S_{\text{S-W}}$), by region, for instance. Normally, for table protection we should consider the relation

$$S_{\text{S-W}} = S_{\text{foreigners}} + S_{\text{non-foreigners}} \ , \qquad \text{(R1)}$$

even if we are not interested in publishing $S_{\text{non-foreigners}}$ data, and may not even have collected them. Typically $S_{\text{S-W}}$ data would also appear in other relations, for instance together with data for permanent workers as

$$S_{\text{all}} = S_{\text{S-W}} + S_{\text{permanent}} \ , \qquad \text{(R2)}$$

so we would have to protect a set of linked tables. From a practical point of view this is of course not desirable. We assume now that $S_{\text{non-foreigners}}$ figures are for all regions much larger as $S_{\text{non-foreigners}}$ figures and that the $S_{\text{non-foreigners}}$ are not published. Then the $S_{\text{S-W}}$ figures cannot be used to estimate a sensitive $S_{\text{foreigners}}$ cell value. It is thus not necessary to coordinate the suppression pattern for $S_{\text{S-W}}$ and $S_{\text{foreigners}}$. This means that it is not necessary to carry out tabular data protection for the (R1) by Region tabulation. It is enough to protect two separate tables: $S_{\text{S-W}}$ or (R2) by Region, and $S_{\text{foreigners}}$ by Region.

This has been an instance of how to avoid a linked-tables structure. Those structures can, on the other hand, be very useful to make tabular data protection problems tractable in practice.

***Splitting detailed tables***
Assume we have collected data on labour costs, including information on NACE category, Region, and Size class of the number of employees. We may want to release a 3-dimensional tabulation of labour costs by NACE, Region and Size Class. Looking at this table however, we may realize that while for a certain section (say: $\text{NACE}_1$) of the economy this amount of detail may be adequate, for the others ($\text{NACE}_2$, $\text{NACE}_3$, ..., $\text{NACE}_n$) it is not: In those sections many of the lower-level cells are either empty, or sensitive. As pointed out at the end of the section on information loss of Section 4.3.2, more detail in a tabulation

generally leads to more suppressions on the higher levels. This kind of oversuppression can be avoided by taking a splitting approach, splitting the table into a set of linked tables: Instead of one table, we protect three: 1.) Labour costs by Region, Size Class and $\text{NACE}_1$, 2.) Labour costs by Region and NACE, and 3.) Labour costs by size class and NACE. Even though we cannot then release any labour cost data for a particular NACE category, in a particular Region and in a particular Size Class unless the NACE category belongs to $\text{NACE}_1$, the additional unsuppressed information gained by means of the splitting approach (as compared to the simple approach dealing with a single 3-dimensional table only) is usually rated much higher.

***Use of hierarchical structures***
Especially when working with the Modular method for secondary cell suppression, using hierarchically structured variables usually helps a lot to reduce computational complexity, and hence computing time. Smaller numbers of categories per equation will make the program run faster. For the hypercube method, this is less of an issue. Here it is generally good to avoid structures with very few categories within a relation. The method has been developed for large tables, where the number of categories per relation is usually less than 20.

***How to process linked tables***
We explain now a way how to process a set of linked tables, for instance $T_1$, $T_2$, $T_3$ using Modular for secondary cell suppression. This process is normally an iterative process. The method described here is referred to as "*traditional method*" in de Wolf, Giessing (2008). De Wolf, Giessing (2008) suggest an extension of the $\tau$-ARGUS Modular (referred to as "*adapted Modular*") method as a much more practical approach to deal with linked tables. Once this extension is available the method described in the following should be used only in cases where the adapted Modular method cannot be applied to the full set of linked tables (e.g. if the tables involve too many dimensions).

The question is then, which protection levels (cf. Section 4.2.2) to assign to those 'manually unsafe cells' (in the $\tau$-ARGUS terminology). A proper method would consist of using a CTA method (c.f. Section 4.2.2) to find the closest adjusted table $T_1^*$ to $T_1$ that is consistent with the suppression pattern $S_{11}$ (that is, all cell values of $T_1^*$ cells that are not in $S_{11}$ must be identical to the original $T_1$ cell values), and the constraints imposed by the protection levels for the sensitive cells. We could then compute protection levels for $S_{11}$ cells on basis of the distances between adjusted and original cell value. In practice a much simpler approach is usually taken, by assigning a fixed percentage (the 'manual safety range' $\tau$-ARGUS terminology) of the cell value as protection level. There is, however, no way to tell which percentage would be suitable: when a small cell is a secondary suppression for a large, strongly sensitive cell, a protection level even of 100% may be required. On the other hand, when a large cell is a secondary suppression for a very small sensitive cell, even 1% may cause overprotection.

After processing $T_2$ in this fashion, $T_3$ is processed. In addition to the $S_{113}$ cells, here we also have to consider as additional suppressions $S_{213}$ cells, i.e. the subset of secondary suppressions selected for Table $T_2$ during the first iteration which are identical to a cell in $T_3$.

In this way, when processing table $T_i$, $(i = 1, 2, 3)$ in iteration step $k_0$, we consider as additional suppressions $\bigcup_{k<k_0} (S_{1\mathrm{ki}} \cup S_{2\mathrm{ki}} \cup S_{3\mathrm{ki}}) \cup \bigcup_{j<i} S_{\mathrm{jk}_0 i}$, while we may try to avoid (by weighting, or by setting cell status to 'protected') that those cells that the tables have in common, i.e. the cells of $T_{i_j}$ $(j \neq i)$, and that are not already contained in the set of additional suppressions are selected as secondary suppressions. Note that we assign 0% protection levels to cells in $S_{iki}$, the set of secondary suppression which have been assigned to the current table in a previous step $k$ of the iteration. The process is continued until for each of the tables the set of additional suppressions that has to be considered when processing the table in this step is identical to the set considered in the previous step.

The necessary modifications of cell properties like cell status, or cell costs can be achieved relatively easily using the a priori file facility of $\tau$-ARGUS. The general concept of this file is outlined in the following.

***A-priori file***

Through the standard sensitivity rules and other options we can apply all kinds of SDC-methods. Nevertheless there are situations which still require 'manual' intervention. E.g. for some other outside reason a cell must be suppressed or contrarily must be excluded from secondary suppression. Examples are the exports of certain sensitive goods (military equipment) to certain countries. These cells must be suppressed. Sometimes a cell value is known already to outsiders (maybe by other publications etc.) and therefore cannot be selected as a secondary suppression.

Also the preferences for the selection of the secondary suppressions via the cost function could be used as an option to influence the secondary suppression pattern. This might be the result of the year-to-year periodical data issue (see below).

Specifying all this information manually during a $\tau$-ARGUS run for a large table can be rather cumbersome and error-prone. In the batch version of $\tau$-ARGUS it is even impossible to do this. So the a-priori file has been introduced.

In this file the user can specify for each cell (by giving the values of the spanning variables):

- cells that cannot be selected as secondary suppressions
- cells that must be suppressed
- specify a new value of the cost function for a cell; this can be either high or low. This will influence the chance of a cell as a candidate for sec. suppression.

Note that this option is not valid for the hypercube, as the hypercube has its own cost function that cannot be influenced.

Further details of this file can be found in the $\tau$-ARGUS manual.

***Periodical data***

If tables are provided regularly (say monthly, quarterly or annually) certain problems may be connected to disclosure control: Firstly, it may seem to be too much effort. Secondly, if we do disclosure analysis independently each time, suppression patterns may differ. We will have the situation that we publish a particular cell value for one period, but use it as secondary suppression in the next period. Also with a cell published in one period, but becoming a primary cell in the second period, there is a risk of a very narrow estimate based on the previous period. Especially if the observations tend to be rather constant over time the pre-period value might be a close estimate of the suppressed value for the current period, which then might be used to disclose a current-period sensitive cell.

An alternative might be simply to copy the suppression pattern of the previous period, and add some (primary and secondary – if required) suppressions for those cases where cells were not sensitive in the last period, but are sensitive in the current period. Another option would be to assign low costs to previous-period suppressions when assigning secondary suppressions to the current data tables. However, both these strategies will cause an increase in information loss, which will be stronger the more changes there are in cell sensitivity between periods.

***Tables with negative values***

In Section 4.2.1 we already discussed the problem of negative contributions, presenting an idea on how to solve the problem with respect to primary protection. However, when there are negative contributions, it may also happen that some of the table cell values turn out to be negative. The current version of $\tau$-ARGUS does not yet support the processing of such a table, but a little pre-processing may solve the problem:

If the microdata is available, one should simply add another variable to the dataset presenting the absolute values for the variable of interest. Cell suppression should then be based on this (positive) variable. As explained in Section 4.2.1, we recommend use of a minimum frequency rule, rather than a concentration rule in such a case. Hence, taking absolute values would not affect the results of primary suppression. As explained in Section 4.2.1, we recommend using protection levels of zero in the case of a minimum frequency rule. So the fact that some cell values will be larger for the shadow variable than for the original variable will have no effect on the feasibility checks within the process of secondary cell suppression. It may still have an effect on the selection of secondary suppressions, if we use this shadow variable also as cost variable, because in that way we assign too large costs to cells with many and/or large negative contributions. However, there are options to modify cell costs for individual cells in $\tau$-ARGUS.

In the more difficult case that the microdata is not available, we could make a new table file, selecting only bottom-level cells. In this new file, we could replace the original cell values by their absolutes, and then use $\tau$-ARGUS facilities to add up the table again.

### 4.3.4 Introductive Example

The German census of horticulture is a decennial survey. The main publication of 2005 data involves about 70 response variables grouped by geography and size class of horticultural area. Many of the response variables are frequencies which were not considered as sensitive. Magnitude variables are horticultural or agricultural area, classified by type of establishment, or by production type, etc., and number of labourers classified in similar ways.

Because the publication presents data on low geographic detail, the experts from agriculture statistics thought it would not be adequate to base disclosure analysis simply on a few lead variables, like horticultural area and copy the resulting suppression pattern to other variables: The argument was that, even if there are many units falling into a particular size class/geography cell, and this cell is not dominated by any large units, it might still happen, for instance, that only a single unit falls into a particular subgroup of, say, the type of production, and this unit could be identifiable by this property (given the size class and geography information). In such a case the subgroup cell should be suppressed to avoid that sensitive information would be published about this unit. Consequently, disclosure risk assessment had to be carried out for each of the variables separately.

The main approach taken to simplify the original extremely complex linked tables structure resulting from linear relationship between the response variables was to ignore response variable relations like we discussed it in 4.3.3: A typical instance is fruit production. Fruit can be produced by farm companies focussing on fruit production (which was one of the categories for 'type of production'), but also by other companies. The publication presents the overall fruit production area, and also fruit production area of farms focussing on fruit production. In order to avoid disclosure-by-differencing problems, in principal we should have processed fruit production area by both categories (e.g. 'farms focussing on fruit production' and 'other farms'). However, as the 'other farms' category corresponds to a much larger group, it seemed to be justified not to include this variable into the disclosure control process, treating 'fruit production area (by size class and geography)' and 'fruit production area of farms focussing on fruit production (by size class and geography)' as separate 2-dimensional tables.

In this way, the disclosure control procedure finally involved 23 tables. 16 were 2-dimensional, the others were 3-dimensional. Most of the tables could be processed independently, but unfortunately, 3 of the 3-dimensional tables were linked.

The explanatory variables were Geography (at 358 categories and 4 levels, down to the department level), size class (at 9 categories), and various 'type of production' classifications, the largest involving 10 categories with hierarchical substructure.

Data for the tabulations was delivered by the German state statistical offices on an aggregate level. Some SAS procedures were developed to organize data import from those 16 files, e.g. to build suitable $\tau$-ARGUS table format files, and also read the results of disclosure processing (e.g. cell status: safe/unsafe) back into the 16 files. Each line of that table format file (corresponding to a particular table cell) contained the respective codes for the spanning variables, the cell value, the cell frequency, the value of the largest and second-largest contributor, and a cell status (safe/unsafe). A special $\tau$-ARGUS facility was used to build federal-level aggregates from the 16 corresponding state-level cells.

For processing the tables we used $\tau$-ARGUS. The $p$%-rule was employed for primary protection. For secondary cell suppression we ran the Modular method of $\tau$-ARGUS. For the 20 tables that could be protected independently from each other, these were straightforward applications of the software, which all finished successfully after at most a few minutes.

The only problem was the processing of the 3 linked tables. We used some little SAS programs to organise the import of secondary suppressions selected in one table into the others in the fashion described in Section 4.3.3, making suitable a-priori files for the next $\tau$-ARGUS run. However, it turned out that this procedure did not work well for that instance: the tables were extremely detailed with few non-zero, non-sensitive cells on the lower levels. As a result, quite a lot of secondary suppressions were assigned, also on the higher levels of the table, in particular those parts that were also contained in one of the other 2 tables. The number of suppressions in the overlap sections did not decrease substantially from one iteration step to the next.

After it was explained to the experts from agricultural statistics that the tables were probably too detailed, it became clear that they had not really been interested in publishing data for those tables on the district level. What they were actually interested in for those 3 tables were 3-dimensional tabulations by size class and geography, down to the level 'Region' (just above the district level), and 2-dimensional tabulations by geography down to the district level, but without size classes. Processing the 3 linked, 3-dimensional tables down to the level 'Region' was a straightforward application of the linked tables processing described in 4.3.3, finished after one iteration step. We then copied secondary suppressions of those three 3-dimensional tables into the corresponding three 2-dimensional district level tables and again applied the linked tables processing which also ended successfully after a single iteration.

## 4.4 Methodological concepts of secondary cell suppression algorithms in $\tau$-ARGUS

Within this section we briefly explain the methodological concepts of the secondary cell suppression algorithms offered by $\tau$-ARGUS.

### 4.4.1 Optimal

The optimal approach is based on a Mathematical Programming technique which consists of solving Integer Linear Programming programs modelling the combinatorial problems under different methodologies (Cell Suppression and Controlled Rounding). The main characteristic of these models is that they share the same structure, thus based only on a 0-1 variable for each cell. In the Cell Suppression methodology, the variable is 1 if and only if the cell value must be suppressed. In the Controlled Rounding methodology, the variable is 1 if and only if the cell value must be rounded up. No other variables are necessary, so the number of variables in the model is exactly the number of cells in the table to be protected. In addition, the model also imposes the protection level requirements (upper, lower and sliding) in the same way for the different methodologies (Cell Suppression and Controlled Rounding). These requirements ask for a guarantee that an attacker will not get too narrow an interval of potential values for a sensitive cell, which he/she will compute by solving two linear programming programs (called attacker problems). Even if a first model containing this two-attacker problem would lead to a bi-level programming model, complex to be solved in practice, a Benders' decomposition approach allows us to convert the attacker problems into a set of linear inequalities. This conversion provides a second model for each methodology that can be efficiently solved by a modern cutting-plane approach. Since the variables are 0-1, a branching phase can be necessary, and the whole approach is named ”branch-and-cut algorithm”.

Branch-and-cut algorithms are modern techniques in Operations Research that provide excellent results when solving larger and complicated combinatorial problems arising in many applied fields (like routing, scheduling, planning, telecommunications, etc.).The idea is to solve a compact 0-1 model containing a large number of linear inequalities (as the ones above mentioned for the Cell Suppression and for the Controlled Rounding) through an iterative procedure that does not consider all the inequalities at the same time, but generates the important ones when needed. This dynamic procedure of dealing with large models allows the program to replace the resolution of a huge large model by a short sequence of small models, which is termed a ”decomposition approach”. The on-line generation of the linear inequalities (rows) was also extended in this work to the variables (columns), thus the algorithm can also works on tables with a large number of cells, and the overall algorithm is named ”branch-and-cut-and-price” in the Operations Research literature.

To obtain good performance, the implementation has also considered many other ingredients, standard in branch-and-cut-and-price approaches. For example, it is fundamentally the implementation of a pre-processing approach where redundant equations defining the table are eliminated, where variables associated to non-relevant cells are removed, and where dominated protection levels are detected. The pre-processing is fundamental to make the problem as small as possible before starting the optimization phase. Another fundamental ingredient is the heuristic routine, which allows the algorithm to start with an upper bound of the optimal loss of information. This heuristic routine ensures the

production of a protected pattern if the algorithm is interrupted by the user before the end. In other words, thanks to the heuristic routine, the implemented algorithm provide a near-optimal solution if the execution is cancelled before having a proof of optimality. During the implicit enumeration approach (i.e., the branch-and-cut-and-price) the heuristic routine is called several times, thus providing different protected patterns, and the best one will be the optimal solution if its loss of information is equal to the lower bound. This lower bound is computed by solving a relaxed model, which consists of removing the integrability condition on the integer model. Since the relaxed model is a linear program, a linear programming solver must be called.

We have not implemented our own linear programming solver, but used a commercial solver which is already tested by other programmers for many years. A robust linear programming solver is a guarantee that no numerical trouble will appear during the computation.

That is the reason to require either CPLEX (from ILOG) or XPRESS (from DashOptimization). Because the model to be solved can be applied to all type of table structures (2-dim, 3-dim, 4-dim, etc), including hierarchical and linked tables, we cannot use special simplex algorithm implementations, like the min-cost flow computation which would require to work with tables that can be modelled as a network (e.g., 2-dimensional tables or collections of 2-dim tables linked by one link). On this special table, ad-hoc approaches (solving network flows or short path problems) could be implemented to avoid using general linear programming solvers.

The optimal solution is the result of solving this complex optimisation problem. Although the computing power of modern PCs has been increased considerably during the past period, nevertheless this method will not be applicable for very large tables, which are not uncommon in NSIs. Another drawback of the current implementation is the lack of facilities to protect against the singleton problem. This might be included in a future release.

However this method is used very successfully in the modular approach described in the next section.

**Reference:**

Fischetti, M. and J.J. Salazar-González (1998). *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints.* Technical Paper, University of La Laguna, Tenerife.

### 4.4.2 Modular {#sec-modular}

The modular (HiTaS) solution is a heuristic approach to cell suppression in hierarchical tables. Hierarchical tables are specially linked tables: at least one of the spanning variables exhibits a hierarchical structure, *i.e.* contains (many) sub-totals.

In Fischetti and Salazar (1998) a theoretical framework is presented that should be able to deal with hierarchical and generally linked tables. In what follows, this will be called the mixed integer approach. In this framework, additional constraints to a linear programming problem are generated. The number of added constraints however, grows rapidly when dealing with hierarchical tables, since many dependencies exist between all possible (sub-)tables containing many (sub-)totals. The implemented heuristic approach (HiTaS) deals with a large set of (sub)-tables in a particular order. A non hierarchical table can be considered to be a hierarchical table with just one level. In that case, the approach reduces to the original mixed integer approach and hence provides the optimal solution. In case of a hierarchical table, the approach will provide a sub-optimal solution that minimises the information loss per sub-table, but not necessarily the global information loss of the complete set of hierarchically linked tables.

In the following section, a short description of the approach is given. For a more detailed description of the method, including some examples, see *e.g.*, De Wolf (2002).

HiTaS deals with cell suppression in hierarchical tables using a top-down approach. The first step is to determine the primary unsafe cells in the base-table consisting of all the cells that appear when crossing the hierarchical spanning variables. This way all cells, whether representing a (sub-)total or not, are checked for primary suppression. Knowing all primary unsafe cells, the secondary cell suppressions have to be found in such a way that each (sub-)table of the base-table is protected and that the different tables cannot be combined to undo the protection of any of the other (sub-)tables. The basic idea behind the top-down approach is to start with the highest levels of the variables and calculate the secondary suppressions for the resulting table. The suppressions in the interior of the protected table are then transported to the corresponding marginal cells of the tables that appear when crossing lower levels of the two variables. All marginal cells, both suppressed and not suppressed, are then 'fixed' in the calculation of the secondary suppressions of that lower level table, i.e., they are not allowed to be (secondarily) suppressed. This procedure is then repeated until the tables that are constructed by crossing the lowest levels of the spanning variables are dealt with.

A suppression pattern at a higher level only introduces restrictions on the marginal cells of lower level tables. Calculating secondary suppressions in the interior while keeping the marginal cells fixed, is then independent between the tables on that lower level, i.e., all these (sub)-tables can be dealt with independently of each other. Moreover, added primary suppressions in the interior of a lower level table are dealt with at that same level: secondary suppressions can only occur in the same interior, since the marginal cells are kept fixed.

However, when several empty cells are apparent in a low level table, it might be the case that no solution can be found if one is restricted to suppress interior cells only. Unfortunately, backtracking is then needed.

Obviously, all possible (sub)tables should be dealt with in a particular order, such that the marginal cells of the table under consideration have been protected as the interior of a

previously considered table. To that end, certain groups of tables are formed in a specific way (see De Wolf (2002)). All tables within such a group are dealt separately, using the mixed integer approach.

The number of tables within a group is determined by the number of parent-categories the variables have one level up in the hierarchy. A parent-category is defined as a category that has one or more sub-categories. Note that the total number of (sub)-tables that have to be considered thus grows rapidly.

For the singleton problem there is a procedure included in the modular solution. At least on each row/column of a (sub-)table it is guaranteed that no disclosure by the singleton is possible.

### 4.4.3 Network

The network flows package for cell suppression (NF CSP) provided by the Dept. of Statistics and Operations Research of the Universitat Politècnica de Catalunya implements a fast heuristic for the protection of statistical data in two dimensional tables with one hierarchical dimension. The method sensibly combines and improves ideas of previous approaches for the secondary cell suppression problem in two-dimensional tables. Details about the heuristic can be found in (Castro, 2003a) and (Castro, 2003b)

Network flow heuristics for secondary cell suppression build on the fact, that a suppressed cell in a two-dimensional table is safe from exact disclosure if, and only if, the cell is contained in a 'cycle', or 'alternating path' of suppressed cells, where a cycle is defined to be a sequence of non-zero cells with indices $\{(i_0, j_0), (i_1, j_0), (i_1, j_1), \ldots, (i_n, j_n), (i_0, j_n)\}$, where all $i_k$ and $j_l$ for $k = 0, 1, \ldots, n$ , and $l = 0, 1, \ldots, n$ respectively ($n$: length of the path), are different, e.g. $i_{k_l} \neq i_{k_2}$ unless $k_1 = k_2$ , and $j_{l_1} \neq j_{l_2}$ unless $l_1 = l_2$ .

The NF CSP heuristic is based on the solution of a sequence of shortest-path subproblems that guarantee a feasible pattern of suppressions (i.e., one that satisfies the protection levels of sensitive cells). Hopefully, this feasible pattern will be close to the optimal one. In the ARGUS implementation solutions of the shortest-path subproblems are computed by either of two network flows optimization solvers also implemented by Universitat Politècnica de Catalunya.

However the current implementation of the network flows do (not yet) protect against the singleton problem.

### 4.4.4 Hypercube

In order to ensure tractability also of big applications, $\tau$−ARGUS interfaces with the GHM*ITER* hypercube method of R. D. Repsilber of the Landesamt für Datenverarbeitung und Statistik in Nordrhein-Westfalen/Germany, offering a quick heuristic solution. The

method has been described in depth in (Repsilber, 1994), or (Repsilber, 2002). For a briefer description see (Giessing and Repsilber, 2002).

> The approach builds on the fact that a suppressed cell in a simple n-dimensional table without substructure cannot be disclosed exactly if that cell is contained in a pattern of suppressed, nonzero cells, forming the corner points of a hypercube.

The algorithm subdivides n-dimensional tables with hierarchical structure into a set of $n$-dimensional sub-tables without substructure. These sub-tables are then protected successively in an iterative procedure that starts from the highest level. Successively for any primary suppression in the current sub-table, all possible hypercubes with this cell as one of the corner points are constructed.

If protection against inferential disclosure is requested, for each hypercube, a lower bound is calculated for the width of the feasibility interval for the primary suppression that would result from the suppression of all corner points of the particular hypercube. To compute that bound, it is not necessary to implement the time consuming solution to the Linear Programming problem and it is possible to consider bounds on cell values that are assumed to be known to an intruder. If it turns out that the bound for the feasibility interval width is sufficiently large, the hypercube becomes a feasible solution. For any of the feasible hypercubes, the loss of information associated with the suppression of its corner points is calculated. The particular hypercube that leads to minimum information loss is selected, and all its corner points are suppressed. If several hypercubes are considered feasible, the method selects the one which requires the smallest number of additional cells to be suppressed. If there are several feasible hypercubes that would all lead to the same number of additional secondary suppressions, the method picks the one with the smallest total of weighted cell values of those additional suppressions. The weights are determined automatically, considering the hierarchical level of cells. Cells on a higher level get larger weights. The weights are determined temporarily, only the subtable which is currently processed is considered.

After all sub-tables have been protected once, the procedure is repeated in an iterative fashion. Within this procedure, when cells belonging to more than one sub-table are chosen as secondary suppressions in one of these sub-tables, in further processing they will be treated like sensitive cells in the other sub-tables they belong to. The same iterative approach is used for sets of linked tables.

The 'hypercube criterion' is a sufficient but not a necessary criterion for a 'safe' suppression pattern. Thus, for particular subtables the 'best' suppression pattern may not be a set of hypercubes – in which case, of course, the hypercube method will miss the best solution and lead to some overprotection. Other simplifications of the heuristic approach that add to this tendency for over-suppression are the following: when assessing the feasibility of a hypercube to protect specific target suppressions against inferential disclosure, the method

- is not able to consider protection maybe already provided by other cell suppressions (suppressed cells that are not corner points of this hypercube) within the same sub-table,
- does not consider the sensitivity of multi-contributor primary suppressions properly, that is, it does not consider the protection already provided to the individual respondent contributions in advance of cell suppression through aggregation of these contributions.

In the implementation offered by $\tau$-ARGUS, GHMITER makes sure that a single respondent cell will never appear to be corner point of one hypercube only, but of two hypercubes at least. Otherwise it could happen that a single respondent, who often can be reasonably assumed to know that he is the only respondent, could use his knowledge on the amount of his own contribution to recalculate the value of any other suppressed corner point of this hypercube. Because of this strategy, the method yields PS2$^*$ standard (c.f. section ).

When protection against inferential disclosure is requested, the settings of the method are determined on basis of the assumptions that an intruder could estimate each cell value to within bounds of $\pm q\%$ prior to the publication, where the percentage of $q$ for those *a priori* bounds is selected by the $\tau$-ARGUS user. Considering these *a priori* bounds, $\tau$-ARGUS determines the settings for the hypercube method in such a way, that it yields PS1$^*$ protection standard (c.f. section ). Because the method is unable to 'add' the protection given by multiple hypercubes, in certain situations it will be unable to confirm that a cell has been protected properly according to PS1$^*$ standard and will attempt to select the hypercube that provides the largest amount of protection to the target suppression in that situation.

## 4.5 Controlled Tabular Adjustment

*'Controlled Tabular Adjustment'* (CTA) suggested for instance in Cox and Dandekar (2002), Castro (2003), or Castro and Giessing (2006). CTA is a new emerging protection method for magnitude tabular data. Unlike cell suppression, it is a perturbative method.

The starting point of the cell suppression methodology presented in 4.3 is that suppressing cells in a table results in theory in replacing the confidential cell value by a feasibility interval that could principally be computed by any user of a table published with suppressions. As explained in 4.2.2., in case the proper protection is given to the cells, the feasibility interval covers the protection interval (i.e. the interval that is needed to protect the individual contribution). CTA methodology on the other hand aims at finding the closest additive table to the original table ensuring that adjusted values of all confidential cells are safely away from their original value (considering the protection intervals) and that the adjusted values are within a certain range of the real values.

Several variants of CTA have been discussed in the literature (Dandekar and Cox, 2002), (Cox et al., 2004), (Castro, 2006) etc., suggesting to obtain CTA by using (mixed integer) linear programming methodology. The main differences between those alternatives are on one hand the way in which the deviation sense of the primary suppressions is determined (heuristically vs. optimal, i.e. through solution of integer programming problems). On the other hand, the definition of constraints matters (forcing the adjusted values to be within a "certain range" of the real values). And finally there is the issue of information loss/cell costs, e.g. the distance metric used to determine what is "close" to the original table. Typically, weighted metrics are used. Implementations usually offer a choice of cost functions.

As an alternative to the above linear programming based approaches, (Cox et al., 2006) and (Ichim and Franconi, 2006) suggest methology to achieve CTA using statistical methods like Iterative Proportional Fitting or Calibration techniques.

While for cell suppression, the effect of SDC on the data is obvious to the users of the data, for CTA this is not the case. Many information loss measures proposed in the literature that should help – eventually both parties: data providers and data users – to judge the quality of an adjusted table are global measures. A typical global measure would be statistics on the relative deviations between true and adjusted cell values. (Cox et al., 2004) hint at measuring the preservation of univariate properties like mean, variance, correlation, etc. in subsets of cells. (Cox et al., 2006) employ Goodness-of-fit statistics like the Kolmogorov Smirnov test to compare adjusted vs. original data.

In contrast to a global measure, a local measure will inform the data user on the reliability of each particular (adjusted) cell value. (Castro and Giessing, 2006b) discuss criteria that could be used by data providers to decide whether an adjustment is so small that there is no need for the cell to be flagged as "adjusted".

In order to be able to publish a single value in the tables whilst at the same time clarifying to the users that these values are not the original values, (Giessing, Hundepool and Castro, 2007) suggest to round the values resulting from CTA. The basis for rounding would be chosen in a way that the rounding interval includes both the true and the adjusted value. This strategy requires however some occasional post-processing to ensure that the protection of the sensitive cells is sufficient.

An implementation based on a simple heuristic to fix deviation senses for the adjustment of sensitive cells is available in the R-package *sdcTable* (Meindl, 2009). An algorithm for optimal CTA has been implemented on behalf of Eurostat by (Castro and Gonzalez, 2009).

## 4.6 Cell Key Method for Magnitude Tables

The '*Cell Key Method*' (CKM) was initially developed by the Australian Bureau of Statistics (Fraser/ Wooton (2016); Thompson et al. (2013)) and hence is sometimes referred to as '*ABS method*'. It is a post tabular perturbative method that maintains consistency between all tables that use the same microdata and configuration. This method adds sufficient '*noise*' to each cell so if an intruder tried to gather information by differencing, they would not be able to obtain the real data. It is one of the SDC methods recommended by Eurostat for the population and housing censuses 2021. In order to keep consistency between tables every record of the underlying microdata is equipped with a randomly generated number, the so called '*Record Key*'. This step is performed just once such that afterwards the Record Key is a fixed component of the microdata that is to be used for every subsequent evaluation. Now whenever the microdata are aggregated to form a table cell, the corresponding Record Keys are aggregated as well, forming the eponymous '*Cell Key*'. Using said Cell Key and a predefined table that encodes a probability distribution, which is tailored to the data/purpose, the corresponding noise can be looked up and added to the original cell value, before dissemination.

For more general information about the Cell Key Method, see Section 5.4.

The Cell Key Method was initially developed to protect frequency tables. By adding a controlled integer noise $v$ to a table cell $n$ a perturbed value $\hat{n} = n + v$ is generated. By applying this procedure to each table cell uncertainty about the real cell value is created. In this respect, this method has similar effects as simple deterministic rounding, but is unbiased and provides a higher level of protection when compared to a variant with similar level of information loss. A short comparison can be found in Enderle et al. (2018).

When adapting this method to magnitude tables, it must be noted that the amount of the noise $v$ needed, depends on the magnitude of the values in the microdata. Whereas in the case of frequencies it is a matter of protecting low counts and in particular individual cases, in the case of magnitude tables the magnitude of the individual values can vary a lot, and it is not sufficient, for example, to add a noise term of magnitude 5 when it is a matter of protecting a value that is one million. On the other hand, adding a noise term of one thousand to an original value of one hundred might be a little exaggerated. Since sufficient protection is particularly important in cases of dominance of one (or few) respondent(s), for the Cell Key Method for magnitude tables the amount of noise should correlate with the size of the largest contribution(s) to a cell value. For simplicity, we assume in the following that only the largest contribution is considered.

So, if $x$ is the actual value of a cell, $x_{max}$ is the corresponding largest contribution and $v$ is a limited random noise variable with a discrete distribution, then the perturbed value $\hat{x}$ computes as $\hat{x} = x + \delta \cdot x_{max} \cdot v$, where $\delta$ is an additional control parameter. Since CKM derives its protective effect from uncertainty, all published values must of course be subject to some perturbation (with some probability). But since the amount of perturbation depends on the magnitude of the largest contribution, the deviations

obtained are relatively small. Especially for cells without dominance issues, by adding (or subtracting) a fraction of the largest contribution value doesn't affect the cell value too much. Whereas in table cells with a single strongly dominating contribution, the change in value is correspondingly more significant. But this is precisely what is desired.

The calculations for the Cell Key Method are normally carried out by appropriate software, such as $\tau$-ARGUS, so that the user does not have to perform them themselves. Nevertheless, in the following, we give a brief, rough overview of how CKM is applied to continuous data, for interested readers. More detailed explanations, along with ideas for how to obtain control parameters such as $\delta$ can be found in Gießing and Tent (2019).

Since magnitudes usually aren't integers but continuous, it is not possible to represent all values in a finite table. Hence an interpolation technique is used, while the p-table generally maintains the same form as depicted in Table 5.25, in Section 5.4 about CKM for frequency tables. Since we don't want for a positive value $x$ to become negative after perturbation we require $0 \leq \hat{x} = x + \delta \cdot x_{max} \cdot v$ which is equivalent to $v \geq (-x)/(\delta \cdot x_{max})$. So $v$ has to be chosen with respect to $x/(\delta \cdot x_{max})$ and hence implicitly depends on both the cell value $x$ and the largest contributor $x_{max}$. Note, that in the case of frequency tables, the distribution of the noise $v$ depended only on the count $n$ to be perturbed, therefore, this value was used to look up the noise in the perturbation table. Accordingly, for the continuous case, the value $x/(\delta \cdot x_{max})$ is now used in the lookup step.

We now use Table 5.25 again, to illustrate such a lookup step. If $x/(\delta \cdot x_{max})$ is one of the numbers 0, 1, 2, 3 or any value larger than 3, the lookup step works just as is the case of frequency tables. So if said value is 3, for example, and if the corresponding Cell Key is 0.8, again we have to look for that row in Table 5.25, for which '*orig. value*' is 3 and for which *'lower bound'* $< 0.8 \leq$ *'upper bound'*. Which again is met in the last table row, for which the noise is given as 1. Hence the perturbed value computes as $\hat{x} = x + \delta \cdot x_{max} \cdot 1$. Imagine now a true cell value $x = 300$, with largest contributor $x_{max} = 200$ and an additional control parameter of $\delta = 0.5$. The corresponding result is then $\hat{x} = 300 + 0.5 \cdot 200 \cdot 1 = 400$. Please keep in mind that in this calculation example all values (including the p-table) are chosen in such a way that the calculation can be carried out as easily as possible and that the amount of noise is not anyway typical for practical use cases.

In case $x/(\delta \cdot x_{max})$ is a non-integer value smaller than 3, this value cannot be found in Table 5.25. To solve this issue interpolation is used, i.e. if $2 < x/(\delta \cdot x_{max}) < 3$, there exists a unique pair of positive numbers $a$ and $b$ such that $x/(\delta \cdot x_{max}) = 2 \cdot a + 3 \cdot b$ and $a + b = 1$. The algorithm, as implemented in $\tau$-ARGUS, for example, then computes both the noise $v_3$ for the case $x/(\delta \cdot x_{max}) = 3$ and the noise $v_2$ for the case $x/(\delta \cdot x_{max}) = 2$. The correct noise then results as $v_2 \cdot a + v_3 \cdot b$. We already looked up the noise $v_3 = 1$, so we still need $v_2$. So once again we need to search Table 5.25. After we spotted the row for which '*orig. value*' equals 2 and for which *'lower bound'* $< 0.8 \leq$ *'upper bound'*, we obtain a noise value of $v_2 = 0$. So if we consider the case where x = 300, $x_{max} = 200$ and $\delta = 0.6$, then $x/(\delta \cdot x_{max}) = 2.5$ and this can be written as $2 \cdot 0.5 + 3 \cdot 0.5$, i.e. in our

instance $a = b = 0.5$. The perturbed value is therefore calculated as $\hat{x} = 300 + 0.6 \cdot 200 \cdot (300 + 0.6 \cdot 200 \cdot (v_2 \cdot 0.5 + v_3 \cdot 0.5) = 300 + 0.6 \cdot 200 \cdot (0 \cdot 0.5 + 1 \cdot 0.5) = 360$ .

## 4.7 Measurement of disclosure risk and information loss

### 4.7.1 Disclosure risk

The disclosure risk for tabular data (especially magnitude) includes the assessment of:

- primary risk,
- secondary risk.

The primary risk concerns the threat for direct identification of an unit resulting from too low frequency or existence of outliers according to the presented magnitude in the cell. The secondary risk assessment is necessary due to the fact that primary confidential cells in detailed tables may still not ensure sufficient protection against re-identification: together with single cells also sums for larger groups, i.e. the margins are computed. Then the protection of the primary sensitive cells can be undone, by some differencing. Therefore the risk of such an event should also be assessed.

The key measure for assessing primary disclosure risk in the case of magnitude tables is based on the $(n, k)$-dominance or $p\%$ rules (c.f. 4.2.1, table 4.1). In some formal regulations or recommendations it is assumed $n = 1$ and $k = 75$. Hovever, in 4.2.1, following the reach practical experience, it is recommended to use $n > 1$ in this context. Quite commonly the $(n, k)$-dominance is combined with $k$-anonymity: a cell is regarded as unsafe if it violates the $k$-anonymity or the $(n, k)$-dominance. Therefore the dislosure risk can be measured as the share of cells violating the finally assumed principle. A broader discussion on an application of these rules to assess the disclosure risk is performed in 4.2. See also Hundepool et al. (2012).

However, this assessment concerns risk at the level of individual table cells. When designing tables, a risk measure at table level might be convenient. For frequncy tables, Antal, Shlomo and Elliot (2014) formulate four fundamental properties for a disclosure risk measure at table level:

- small cell values should have higher disclosure risk than larger,
- uniformly distributed frequencies imply low disclosure risk ,
- the more zero cells in the census table, the higher the disclosure risk ,
- the risk measure should be bounded by 0 and 1.

In the currently investigated case of magnitude table one should add one more condition concerning the specificity of this type of data presentation. That is, the larger is the share of the largest contributors in a cell the higher is the disclosure risk.

Shlomo, Antal and Elliot (2015) proposed in this context measure based on the entropy. According to their approach, a high entropy indicates that the distribution across cells is uniform and a low entropy indicates mainly zeros in a row/column or table with a few non-zero cells. The fewer the number of non-zero cells, the more likely that attribute disclosure occurs. The entropy is given as

$$H = -\sum_{i=1}^{k} \frac{c_i}{n} \cdot \log \frac{c_i}{n}, \quad (4.6.1)$$

where $c_i$ is the number of units contained in $i$-th cell, $k$ is the number of cell in the tables and $n$ is the total number of units covered by the table. The measure (4.6.1) is next normalized to the form

$$1 - \frac{H}{\log n}. \quad (4.6.2)$$

This approach, however, doesn't take our fifth condition into account. So, one can extend (4.6.2) to the form combining both aspects, e.g. in the following way:

$$r = 1 - \frac{1}{2}\left(\frac{H}{\log n} + \frac{1}{n}\sum_{i=1}^{n} \frac{x_{i(\max)}}{\sum_{j \in C_i} x_j}\right),$$

where $x_{i(\max)}$ is the share of the largest contributors to the cell $C_i$ and $x_j$ is the contribution of the $j$-th unit to it. This measure takes value from $[0, 1]$ and is easily interpretable.

The disclosure risk assessment can be also adjusted to the specificity of currently used SDC method. For instance, Enderle, Giessing and Tent (2020) have proposed an original risk measure for continous data perturbed using the Cell Key approach. It assumes that an intruder can know the amount of noise and the maximum deviation parameter. Then he/she can determine the feasibility intervals for original internal and margin cells using the noisy values. The risk estimate is then a weighted sum of risk probabilities, e.g. risk probabilites computed by summing up certain probabilities defined by the p-table relating to "risky" constellations, weighted by the empirical probability of the respective constellation to occur in a given test dataset.

### 4.7.2 Information loss

The basis of assessment of information loss for tables are differences between values of cells determined using microdata after application of SDC methods (or perturbed/hidden during creation of tables) and relevant values obtained on the basis of the original data. Let $T_0$ denote a table generated using original microdata and $T_1$ - analogous table created on

the basis of perturbed relevant microdata (or in which original cell vaules were pretrurbed - post-tabular pertrurbation). Denote by $T_l(c)$ the value of cell $c$ of table $T_l$, $l = 0, 1$.

The following metrics are commonly recommended for determining measures of information loss due to application of SDC process in magnitude tables:

- absolute deviation: $|T_1(c) - T_0(c)|$,
- relative absolute deviation: $\frac{|T_1(c)-T_0(c)|}{T_0(c)}$,
- absolute deviation of square roots: $|\sqrt{T_1(c)} - \sqrt{T_0(c)}|$.

for any $c$.

As one can observe, the absolute deviation of square roots has a sense in practice only if the cell values are nonnegative. However, this is the most common situation. Otherwise, $\sqrt{T_l(c)}$ can be replaced with $-\sqrt{|T_l|}$, $l = 0, 1$.

Using the metrics given above one can define complex measures of information loss for a given aggregate $A$ (it can be the whole table or a specific subset of table cells, referring to, for example, the same spatial unit - such as population numbers by age groups for a LAU 1 unit). The first of these measures is the average absolute deviation - the mean of absolute differences between cell values in original and modified tables:

$$\mathrm{AD}(A) = \frac{\sum_{c \in A} |T_1(c) - T_0(c)|}{n_A},$$

where $n_A$ is the number of cells included in the aggregate $A$.

One can also propose to use in this context the sum of relative absolute deviations, i.e. the the sum of relative differences between cell values in both tables:

$$\mathrm{RAD}(A) = \sum_{c \in A} \frac{|T_1(c) - T_0(c)|}{T_0(c)}.$$

The last - but not least - measure which we would like to present here is the measure based on absolute differences between square roots from cell values in original nad perturbed tables using the formula of Hellinger's distance (proposed in 1909 by German mathematician Ernst Hellinger (1883-1950)):

$$\mathrm{HD}(A) = \sqrt{\frac{1}{2} \sum_{c \in A} \left(\sqrt{T_1(c)} - \sqrt{T_0(c)}\right)^2}.$$

However, there may be missing data in the tables. When perturbative SDC methods are applied, these gaps will result mainly from missing data occurring in the original microdata being a basis of construction of tables. In this case, during computation of cell values one

can omit these gaps or earlier make an imputation of them. The only inconvenience may appear when there are no data concerning the categories defined by the cell. Then one must either resign from a given structure of the table (e.g. by combining some of the original categories into another, more coarse one), or the measure of loss should be based on the measurement of loss at the level of microdata corresponding to this cell, performed in the manner described in chapter 3.

More difficult is the situation where non-perturbative post-tabular SDC methods are used. For assessing information loss from the perspective of the data user, one can make a comparison of tables before and after the SDC process. For this, one should simulate a table, in which for missing cells their values are imputed. On the other hand, for controlling the process of selection of secondary cell suppression , information loss will be expressed as the sum of costs incurred due to it. The problem is whether the weight of each cell in the table is the same, or whether cells with a higher value have a higher weight (cf. the discussion in 4.3.1). In practice, suppression of too many cells with high values can significantly decrease the utility of disseminated data. The issue of information loss can be variously expressed, depending on preferences and needs of a user. In this way, it is possible to influence the operation of the algorithm for selecting cells for secondary suppression. In 4.3.1 (see also Hundepool et al. (2012)) we indicate the most common criteria taken into account when the cost function for cell suppression is defined:

- the same weights for all cells - for minimalization of the number of secondary suppressed cells,
- number of units in aggregate represented by a cell - leads to looking for possibilities of suppression of only such cells which will represent jointly as small number of units as possible,
- cell value - an optimal solution is leaving in publication as many cells with higher values as possible.

For more details, see discussion on cell costs and cost functions in 4.3.1.

### 4.7.3 References

Antal, L., Shlomo, N., & Elliot, M. (2014). Measuring disclosure risk with entropy in population based frequency tables. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings* (pp. 62-78). Springer International Publishing.

Box, G. E., Cox, D. R. (1964), An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26:211–252.

Enderle, T., Giessing, S., & Tent, R. (2020). Calculation of risk probabilities for the cell key method. In: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings* (pp. 151-165). Springer International Publishing.

Hundepool. A., Domingo-Ferrer. J., Franconi, L., Giessing S., Nordholt, E.S., Spicer, K. and de Wolf, P-P. (2012), *Statistical Disclosure Control*, series: Wiley Series in Survey Methodology, John Wiley & Sons, Ltd., Chichester.

Shlomo, N., Antal, L., & Elliot, M. (2015). Measuring disclosure risk and data utility for flexible table generators. *Journal of Official Statistics*, 31(2), 305-324.

## 4.8 References

Box, G. E. P. and Cox, D. R. (1964) *An analysis of transformations.* Journal of Royal Statistical Society, Series B, vol. 26, pp. 211—246.

Castro, J.(2002), *Network flows heuristics for complementary cell suppression: an empirical evaluation and extensions*, in LNCS 2316, Inference Control in Statistical Databases, J. Domingo-Ferrer (Ed), (2002) 59–73

Castro, J. (2003 a), *'Minimum-Distance Controlled Perturbation Methods for Large-Scale Tabular Data Protection'*, accepted subject to revision to European Journal of Operational Research, 2003

Castro, J. (2003 b) *User's and programmer's manual of the network flows heuristics package for cell suppression in 2D tables* Technical Report DR 2003-07, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain,2003; See http://neon.vb.cbs.nl/casc/deliv/ 41D6_NF1H2D-Tau-Argus.pdf

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

Castro, J., Giessing S. (2006a). *Testing variants of minimum distance controlled tabular adjustment*, in Monographs of Official Statistics. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 333-343

Catro, J., Giessing S. (2006b). *Quality issues minimum distance controlled tabular adjustment*, paper presented at the European Conference on Quality in Survey Statistics (Q2006), 24.-26. April 2006 in Cardiff

Catro, J., Gonzalez J.A. (2009). *A Package for L1 Controlled Tabular Adjustment*, paper presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Bilbao, 2-4 December 2009)

Cox, L. (1981), '*Linear Sensitivity Measures in Statistical Disclosure Control',* Journal of Planning and Inference, 5, 153 - 164, 1981

Cox, L. (2001), '*Disclosure Risk for Tabular Economic Data'*, In: 'Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies' Doyle, Lane, Theeuwes, Zayatz (Eds), North-Holland

Cox, L., Dandekar, R.H., (2002), '*Synthetic Tabular Data – an Alternative to Complementary Cell Suppression*, unpublished manuscript

Cox, L. H., Kelly, J. P., and Patil, R. (2004), *Balancing quality and confidentiality for multivariate tabular data*, Lecture Notes in Computer Science, 3050, 87–98.

Cox, L., Orelien, J. G., Shah, B. V. (2006), *A Method for Preserving Statistica Distributions Subject to Controlled Tabular Adjustment*, In: Domingo-Ferrer, Franconi (Eds.): Privacy in Statistical Databases 2006, Lecture Notes in Computer Science , Vol. 4302, Springer, Heidelberg (2006), p.1-11

De Wolf, P.P. (2002), 'HiTaS*: A Heustic Approach to Cell Suppression in Hierarchical Tables',* In: 'Inference Control in Statistical Databases' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)

De Wolf, P.P. (2007), '*Cell suppression in a special class of linked tables*', paper presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Manchester, December 2007, available at

http://epp.eurostat.ec.europa.eu/portal/page?_pageid=3154,70730193,3154_70730647&_dad=porta

De Wolf, P.P., Giessing, S. (2008) *How to make the $\tau$-ARGUS Modular Method Applicable to Linked Tables.* In: Domingo-Ferrer, Josep; Saygin, Yücel (Eds.): Privacy in Statistical Databases 2008, Lecture Notes in Computer Science , Vol. 5262, Springer, Heidelberg (2008), p.227-238.

Enderle, T., Giessing, S., Tent, R. (2018), '*Designing Confidentiality on the Fly Methodology – Three Aspects'*, In: Domingo-Ferrer, J., Montes, F. (eds) Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science(), vol 11126. Springer, Cham. https://doi.org/10.1007/978-3-319-99771-1_3

EU (2004), *Community statistics relating to the trading of goods between Member States,* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32004R0638:EN:HTML

Fraser, Bruce/Wooton, Janice (2005), '*A proposed method for confidentialising tabular output to protect against differencing'* Work session on Statistical Data Confidentiality, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf

Geurts, J. (1992), *'Heuristics for Cell Suppression in Tables'*, working paper, Netherlands Central Bureau of Statistics

Giessing, S. and Repsilber, D. (2002), '*Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine',* in 'Inference Control in Statistical Databases' Domingo-Ferrer (Editor), Springer Lecture Notes in Computer Science Vol. 2316

Giessing, S. (2004), *Survey on methods for tabular protection in ARGUS,* Lecture Notes in Computer Science. Volume Privacy in statistical databases 3050, 113, J. Domingo-Ferrer and V. Torra, Springer, Berlin

Giessing, S., Dittrich, S., Gehrling, D., Krüger, A., Merz, F.J., Wirtz, H. (2006), *, Bericht der Arbeitsgruppe „Geheimhaltungskonzept des statistischen Verbundes, Pilotanwendung: Umsatzsteuerstatistik*" ', document for the meeting of the „Ausschuss für Organisation und Umsetzung", Mai 2006, in German

Giessing, S., Hundepool, A., Castro, J. (2007) *Rounding Methods for Protecting EU-aggregates*, proceedings of the Joint UNECE/Eurostat Worksession on Statistical Confidentiality in Manchester, December 2007

Fischetti, M, Salazar Gonzales, J.J. (2000), '*Models and Algorithms for Optimizing Cell Suppression Problem in Tabular Data with Linear Constraints'*, in Journal of the American Statistical Association, Vol. 95, pp 916

Hundepool, A, Wetering, A van de, Ramaswamy, R, Wolf, PP de, Giessing, S, Fischetti, M, Salazar, JJ, Castro, J, Lowthian, P, (2005), $\tau$-ARGUS 3.1 user manual, Statistics Netherlands, Voorburg NL, Feb. 2005. http://neon.vb.cbs.nl/casc.

Ichim, D., Franconi, L. (2006), *Calibration estimator for magnitude tabular data protection,* Proceedings of the conference Privacy in Statistical Databases 2006, December 2006, Rome, Italy.

Loeve, A. (2001), '*Notes on sensitivity measures and protection levels'*, Research paper, Statistics Netherlands, available at http://neon.vb.cbs.nl/casc/related/marges.pdf

Meindl, B. *Linking Complementary Cell Suppression and the Software R,* paper presented at the New Techniques and Technologies (NTTS) Conference, 18.-20. Feb. 2009 in Brussels

Salazar, J.J. (2003) "*Partial Cell Suppression: a New Methodology for Statistical Disclosure Control*", Statistics and Computing, 13, 13-21

Repsilber, R. D. (1994), '*Preservation of Confidentiality in Aggregated data'*, paper presented at the Second International Seminar on Statistical Confidentiality, Luxembourg, 1994

Repsilber, D. (2002), '*Sicherung persönlicher Angaben in Tabellendaten'* - in Statistische Analysen und Studien Nordrhein-Westfalen, Landesamt für Datenverarbeitung und Statistik NRW, Ausgabe 1/2002 (in German)

Thompson, Gwenda/Broadfoot, Stephen/Elazar, Daniel (2013), '*Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics'*, Paper presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Ottawa 2013. [Zugriff am 1. November 2019]. Verfüg- bar unter: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf

Van der Meijden, R., (2006) ,*Improving confidentiality with τ-ARGUS by focussing on clever usage of microdata'*, in Monographs of Official Statistics. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 363-370

Wolf, P.P. de (2002). *HiTaS: a heuristic approach to cell suppression in hierarchical tables.* Proceedings of the AMRADS meeting in Luxembourg.

# 5 Frequency tables

## 5.1 Introduction

This chapter discusses disclosure controls for frequency tables, that is tables of counts (or percentages) where each cell value represents the number of respondents in that cell.

Traditionally frequency tables have been the main method of dissemination for census and social data by NSIs. These tables contain counts of people or households with certain social characteristics. Frequency tables are also used for business data where characteristics are counted, such as the number of businesses. Because of their longer history there has been relatively more research on protecting frequency tables, as compared with newer output methods such as microdata.

Section 5.2 of the chapter outlines the common types of disclosure risk and how the consideration of these risks leads to the definition of unsafe cells in frequency tables. The process of aggregating individual records into groups to display in tables reduces the risk of disclosure compared with microdata, but usually some additional SDC protection is needed for unsafe cells in tables. Disclosure control methods are used to reduce the disclosure risk by disguising these unsafe cells.

Advantages and disadvantages of a range of different SDC methods are discussed in Section 5.3 on a general basis. Well established methods of SDC for frequency tables are the cell key method, introduced in Section 5.4, and rounding, discussed in Section 5.5 which explains alternative techniques such as conventional and random rounding, small cell adjustment, and the mathematically much more demanding controlled rounding. Section 5.5 also provides information on how the software package $\tau$-ARGUS can be used to apply this method to frequency tables. Section 5.6 introduces the targeted record swapping which is a pre-tabular method, e.g. applied on the micro data before generating the table, and is intended to be used as protection method for tables instead of micro data.

Section 5.7 adresses a special kind of disclosure risk that can be connected to the release of means based on original data, when the underlying frequencies are protected by a perturbative, non-additive SDC method.

The final section of the chapter, Section 5.8, describes different information loss measures that can be used to evaluate the impact that different disclosure control methods have on the utility of frequency tables.

## 5.2 Disclosure risks

Disclosure risks for frequency tables primarily relate to 'unsafe cells'; that is cells in a table which could lead to a statistical disclosure. There are several types of disclosure risk and the associated unsafe cells can vary in terms of their impact. A risk assessment should be undertaken to evaluate the expected outcomes of a disclosure. In order to be explicit about the disclosure risks to be managed one should also consider a range of potentially disclosive situations and use these to develop appropriate confidentiality rules to protect any unsafe cells.

The disclosure risk situations described in this section primarily apply to tables produced from registration processes, administrative sources or censuses, e.g. data sources with a complete coverage of the population or sub-population. Where frequency tables are derived from sample surveys, e.g. the counts in the table are weighted, some protection is provided by the sampling process. The sample a priori introduces uncertainty into the zero counts and other counts through sample error.

It should be noted that when determining unsafe cells one should take into account the variables that define the population within the table, as well as the variables defining the table. For example, a frequency table may display income by region for males. Although sex does not define a row or column it defines the eligible population for the table and therefore must be considered as an identifying variable when thinking about these disclosive situations.

Disclosure risks are categorised based on how information is revealed. The most common types of disclosure risk in frequency tables are described below.

*Identification* as a disclosure risk involves finding yourself or another individual or group within a table. Many NSIs will not consider that self-identification alone poses a disclosure risk. An individual that can recall their circumstances at the time of data collection will likely be able to deduct which cell in a published table their information contributes to. In other words, they will be able to identify themselves but only because they know what attributes were provided in the data collection, along with any other information about themselves which may assist in this detection.

However, identification or self-identification can lead to the discovery of rareness, or even uniqueness, in the population of the statistic, which is something an individual might not have known about themselves before. This is most likely to occur where a cell has a small value, e.g. a 1, or where it becomes in effect a population of 1 through subtraction or deduction using other available information. For certain types of information, rareness or uniqueness may encourage others to seek out the individual. The threat or reality of such a situation could cause harm or distress to the individual, or may lead them to claim that the statistics offer inadequate disclosure protection for them, and therefore others.

Identification or self-identification may occur from any cells with a count of 1, i.e. representing one statistical unit. Table 5.1 presents an example of a low-dimensional table in a particular area where identification may occur.

Table 5.1: Marital status by sex

| **Marital Status** | **Male** | **Female** | **Total** |
|---|---|---|---|
| Married | 38 | 17 | 55 |
| Divorced | 7 | 4 | 11 |
| Single | 3 | **1** | 4 |
| Total | 48 | 22 | 70 |

The existence of a 1 in the highlighted cell indicates that the female who is single is at risk of being identified from the table.

Identification itself poses a relatively low disclosure risk, but its tendency to lead to other types of disclosure, together with the perception issues it raises means several NSIs choose to protect against identification disclosure. Section 5.3 discusses protection methods which tend to focus on reducing the number of small cells in tables.

*Attribute disclosure* involves the uncovering of new information about a person through the use of published data. An individual attribute disclosure occurs when someone who has some information about an individual could, with the help of data from the table (or from a different table with a common attribute), discover details that were not previously known to them. This is most likely to occur where there is a cell containing a 1 in the margin of the table and the corresponding row or column is dominated by zeros. The individual is identified on the basis of some of the variables spanning the table and a new attribute is then revealed about the individual from other variables. Note that identification is a necessary precondition for individual attribute disclosure to occur, and should therefore be avoided.

This type of disclosure is a particular problem when many tables are released from one data set. If an intruder can identify an individual then additional tables provide more detail about that person. In continuation of the example shown in Table 5.1, the cell disclosing the single female as unique will ultimately turn into a marginal cell in a higher dimensional table such as Table 5.2 below and her number of hours worked is revealed.

Table 5.2: Marital status and sex by hours worked

| **Marital status/ Hours worked** | **Male** | | | **Female** | | | **Total** |
|---|---|---|---|---|---|---|---|
| | **More than 30** | **16-30** | **15 or less** | **More than 30** | **16-30** | **15 or less** | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Married | 30 | 6 | 2 | 14 | 3 | **0** | 55 |
| Divorced | 3 | 4 | 0 | 2 | 2 | **0** | 11 |
| Single | 2 | 0 | 1 | 0 | 0 | **1** | 4 |
| Total | 35 | 10 | 3 | 16 | 5 | **1** | 70 |

The table shows how attribute disclosure arises due to the zeros dominating the column of the single female, and it is learned that she is in the lowest hours-worked band.

The occurrence of a 2 in the table could also lead to identification if one individual contributed to the cell and therefore could identify the other individual in the cell.

An example of potential attribute disclosure from the 2001 UK Census data, involves 184 persons living in a particular area in the UK. Uniques (frequency counts of 1) were found for males aged 50-59, males aged 85+, and females aged 60-64. An additional table showed these individuals further disseminated by health variables, and it was learned that the single male aged 50-59 and the single female aged 60-64 had good or fairly good health and no limiting long-term illness, while the single male aged 85+ had poor health and a limiting long-term illness. Without disclosure control, anyone living in this particular area had the potential to learn these health attributes about the unique individuals. Full coverage sources – like the Census – are a particular concern for disclosure control, because they are compulsory, so there is an expectation to find all individuals in the output. Although there may be some missing data and coding errors etc., NSIs work to minimise these, and the data issues are unlikely to be randomly distributed in the output. Certain SDC techniques can be adjusted to target particular variables (or tables) with more or less inherent data error. For example, providing more cell suppression for variables which are known to be better quality and have fewer data issues.

Another disclosure risk involves learning a new attribute about an identifiable group, or learning a group does not have a particular attribute. This is termed *group attribute disclosure*, and it can occur when all respondents fall into a subset of categories for a particular variable, i.e. where a row or column contains mostly zeros and a small number of cells that are non-zero. This type of disclosure is a much neglected threat to the disclosure protection of frequency tables, and in contrast to individual attribute disclosure, it does not require individual identification. In order to protect against group attribute disclosure it is essential to introduce ambiguity in the zeros and ensure that all respondents do not fall into just one or a few categories.

Table 5.3 below shows respondents in a particular area broken down by hours worked and income.

Table 5.3: Marital status by hours worked

| | **Hours worked** | | |
|---|---|---|---|
| **Marital status** | **Full time** | **Part time** | **Total** |

| | | | |
|---|---|---|---|
| Married | **6** | **0** | **6** |
| Divorced | 5 | 1 | 6 |
| Single | 2 | 2 | 4 |
| Total | 13 | 3 | 16 |

From the table we can see that all married individuals work full time, therefore any individual in that area who is married will have their hours worked disclosed.

The table also highlights another type of group attribute disclosure referred to as 'within-group disclosure'. This occurs for the divorced group and results from all respondents falling into two response categories for a particular variable, where one of these response categories has a cell value of 1. In this case, the divorced person who works part time knows that all other divorced individuals work full time.

*Differencing* involves an intruder using multiple overlapping tables and subtraction to gather additional information about the differences between them. A disclosure by differencing occurs when this comparison of two or more tables enables a small cell (0, 1, or 2) to be calculated. Disclosures by differencing can result from three different scenarios which will be explained in turn:

**Disclosure by *geographical differencing*** may result when there are several published tables from the same dataset and they relate to similar geographical areas. If these tables are compared, they can reveal a new, previously unpublished table for the differenced area. For instance, 2001 Output Areas (OA) are similar in geographical size to 1991 Enumeration Districts (ED), and a new differenced table may be created for the remaining area. A fictitious example of this is presented below in Table 5.4, Table 5.5 and Table 5.6.

Table 5.4: Single Person households and hours worked in Area A (2001 OA definition)

| | **Single Person Household Male** | **Single Person Household Female** |
|---|---|---|
| More than 30 | 50 | 54 |
| 16-30 | 128 | 140 |
| 15 or less | 39 | 49 |

Table 5.5: Single Person households and hours worked in Area A (1991 ED definition)

| | **Single Person Household Male** | **Single Person Household Female** |
|---|---|---|
| More than 30 | 52 | 55 |
| 16-30 | 130 | 140 |

| | | |
|---|---|---|
| 15 or less | 39 | 49 |

Table 5.6: New differenced table (via geographical differencing)

| | **Single Person Household Male** | **Single Person Household Female** |
|---|---|---|
| More than 30 | 2 | 1 |
| 16-30 | 2 | 0 |
| 15 or less | 0 | 0 |

The above example demonstrates how simple subtraction of the geographical data in Table 5.4 from Table 5.5 can produce disclosive information for the area in Table 5.6.

**Disclosure by *linking*** can occur when published tables relating to the same base population are linked by common variables. These new linked tables were not published by the NSI and therefore may reveal the statistical disclosure control methods applied and/or unsafe cell counts.

A fictitious example of disclosure by linking is provided below in Table 5.7 to Table 5.10, which are linked by employment status and whether or not the respondents live in the area.

Table 5.7: Area of Residence or Workplace

| | **Number working in area** | **Number living in area** | **Living and working area** |
|---|---|---|---|
| Area A | 49 | 102 | 22 |

Table 5.8: Employment status in Area A

| | **Number of Persons** |
|---|---|
| Employed | 85 |
| Not employed | 17 |
| Total | 102 |

Table 5.9: Males working and living in Area A

| | Living and working in Area A | Living in Area A and working elsewhere | Working in Area A and living elsewhere |
|---|---|---|---|
| Males | 21 | 58 | 23 |

Table 5.10: New differenced table (via linking)

| | **Living and working in Area A** | **Living in Area A and working elsewhere** | **Working in Area A and living elsewhere** |
|---|---|---|---|
| Males | 21 | 58 | 23 |
| Females | 1 | 5 | 4 |
| Total | 22 | 63 | 27 |

Table 5.10 shows the new data which can be derived by combining and differencing the totals from the existing tables. The linked table discloses the female living and working in Area A as a unique.

Importantly, when linked tables are produced from the same dataset it is not sufficient to consider the protection for each table separately. If a cell requires protection in one table then it will require protection in all tables, otherwise the protection in the first table could be undone.

The last type of disclosure by differencing involves **differencing of *sub-population tables*.** Sub-populations are specific groups which data may be subset into before a table is produced (e.g. a table of fertility may use a sub-population of females). Differencing can occur when a published table definition corresponds to a sub-population of another published table, resulting in the production of a new, previously unpublished table. If the total population is known and the subpopulation of females is gathered from another table, the number of males can be deduced.

Tables based on categorical variables which have been recoded in different ways may also result in this kind of differencing. To reduce the disclosure risk resulting from having many different versions of variables, most NSIs have a set of standard classifications which they use to release data. An example using the number of hours worked is shown below in Tables Table 5.11 to Table 5.13.

Table 5.11: Hours worked by sex in Area A

| | **<20** | **20 - 39** | **40 - 59** | **60 - 69** | **70 or more** |
|---|---|---|---|---|---|
| Male | 6 | 9 | 5 | 8 | 4 |
| Female | 10 | 38 | 51 | 42 | 32 |

Table 5.12: Hours worked by sex in Area A

| | **<25** | **25 - 39** | **40 - 59** | **60 - 69** | **70 or more** |
|---|---|---|---|---|---|
| Male | 7 | 8 | 5 | 8 | 4 |
| Female | 10 | 38 | 51 | 42 | 32 |

Table 5.13: New differenced table (via sub-populations)

| | **<20** | **20 - 24** | **25 - 39** | **40 - 59** | **60 - 69** | **70 or more** |
|---|---|---|---|---|---|---|
| Male | 6 | **1** | 8 | 5 | 8 | 4 |
| Female | 10 | 0 | 38 | 51 | 42 | 32 |

The example indicates how a new table can be differenced from the original tables, in particular a new hours worked group (for 20-24 hours) which reveals that the male falling into this derived hours worked group is unique.

More information on disclosure by differencing can be obtained from Brown (2003) and Duke-Williams and Rees (1998).

In addition to providing actual disclosure control protection for sensitive information, NSIs need to be seen to be providing this protection. The public may have a different understanding of disclosure control risks and their perception is likely to be influenced by what they see in tables. If many small cells appear in frequency tables users may perceive that either no SDC, or insufficient SDC methods have been applied to the data. Section 5.3 discusses SDC methods, but generally some methods are more obvious in the output tables than others. To protect against negative perceptions, NSIs should be transparent about the SDC methods applied. Managing perceptions is important to maintain credibility and responsibility towards respondents. Negative perceptions may impact response rates for censuses and surveys if respondents perceive that there is little concern about protecting their confidentiality. More emphasis has been placed on this type of disclosure risk in recent years due to declining response rates and data quality considerations. It is important to provide clear explanations to the public about the protection afforded by the SDC method, as well as guidance on the impact of the SDC methods on the quality and utility of the outputs. Explanations should provide details of the methods used but avoid stating the exact parameters as this may allow intruders to unpick the protection.

Table 5.14: Summary of disclosure risks associated with frequency tables

| **Disclosure Risk** | **Description** |
|---|---|
| Identification | Identifying an individual in a table |
| Attribute disclosure (individual and group) | Finding out previously unknown information about an individual (or group) from a table |

| | |
|---|---|
| Disclosure by differencing | Uncovering new information by comparing more than one table |
| Perception of disclosure | The public's feeling of risk based on what is seen in released tables |

## 5.3 Methods

There are a variety of disclosure control methods which can be applied to tabular data to provide confidentiality protection. The choice of which method to use needs to balance the how the data is used, the operational feasibility of the method, and the disclosure control protection it offers. SDC methods can be divided into three categories which will be discussed in turn below: those that adjust the data before tables are designed (pre-tabular), those that determine the design of the table (table redesign) and those that modify the values in the table (post-tabular). Further information on SDC methods for frequency tables can also be found in Willenborg & Ton de Waal (2001) and Doyle et al (2001).

Pre-tabular disclosure control methods are applied to microdata before it is aggregated and output in frequency tables. These methods include: record swapping, over imputation, data switching PRAM, sampling and synthetic mircrodata (see Section 3.4 or Section 5.6, for details of the methods). A key advantage of pre-tabular methods is that the output tables are consistent and additive since all outputs are created from protected microdata. Pre-tabular methods by definition only need to be applied once to the microdata and after they are implemented for a microdata set (often in conjunction with threshold or sparsity rules) they can be used to allow flexible table generation. This is because pre-tabular methods provide some protection against disclosure by differencing and any uncovered slivers will have already had SDC protection applied.

Disadvantages of pre-tabular techniques are that one must have access to the original microdata. Also, a high level of perturbation may be required in order to disguise all unsafe cells. Pre-tabular methods have the potential to distort distributions in the data, but the actual impact of this will depend on which method is used and how it is applied. It may be possible to target pre-tabular methods towards particular areas or sensitive variables. Generally pre-tabular methods are not as transparent to users of the frequency tables and there is no clear guidance that can be given in order to make adjustments in their statistical analysis for this type of perturbation.

Table redesign is recommended as a simple method that can minimise the number of unsafe cells in a table and preserve original counts. It can be applied alongside post-tabular or pre-tabular disclosure control methods, as well as being applied on its own. As an additional method of protection it has been used in many NSI's including the UK and New Zealand. As table redesign alone provides relatively less disclosure control protection

than other methods, it is often used to protect sample data, which already contains some protection from the sampling process.

Table redesign methods used to reduce the risk of disclosure include;

- aggregating to a higher level geography or to a larger population subgroup
- applying table thresholds
- collapsing or grouping categories of variables (reducing the level of detail)
- applying a minimum average cell size to released tables.

The advantages of table redesign methods are that original counts in the data are not damaged and the tables are additive with consistent totals. In addition, the method is simple to implement and easy to explain to users. However, the detail in the table will be greatly reduced, and if many tables do not pass the release criteria it may lead to user discontent.

Statistical disclosure control methods that modify cell values within tabular outputs are referred to as post-tabular methods. Such methods are generally clear and transparent to users, and are easier to understand and account for in analyses, than pre-tabular methods. However, post-tabular methods suffer the problem that each table must be individually protected, and it is necessary to ensure that the new protected table cannot be compared against any other existing outputs in such a way which may undo the protection that has been applied. In addition post-tabular methods can be cumbersome to apply to large tables. The main post-tabular methods include cell suppression, the cell key method, and rounding.

Table 5.15: Summary of Tabular Disclosure Control Methods

| | **Pre-Tabular** | **Table Redesign** | **Post-Tabular** |
|---|---|---|---|
| | Methods applied before tables are created | Methods applied as tables are created | Methods applied after tables are created |
| Tables and totals will be additive and consistent | Yes | Yes | No |
| Methods are visible to users and can be accounted for in analysis | No | Yes | Yes |
| Methods need to be applied to each table individually | No | Yes | Yes |
| Flexible table generation is possible | Yes | No | No (for Cell suppression) |

The main perturbative post-tabular methods of disclosure control are discussed in the two subsequent sections.

Cell suppression is a non-perturbative method of disclosure control, (it is described in detail in Chapter 4), but the method essentially removes sensitive values and denotes them as missing. Protecting the unsafe cells is called primary suppression, and to ensure these cannot be derived by subtractions from published marginal totals, additional cells are selected for secondary suppression.

Cell suppression cannot be unpicked provided secondary cell suppression is adequate and the same cells in any linked tables are also suppressed. Other advantages are that the method is easy to implement on unlinked tables and it is highly visible to users. The original counts in the data that are not selected for suppression are left unadjusted.

However cell suppression has several disadvantages as a protection method for frequency tables, in particular information loss can be high if more than a few suppressions are required. Secondary suppression removes cell values which are not necessarily a disclosure risk, in order to protect other cells which are a risk. Disclosive zeros need to be suppressed and this method does not protect against disclosure by differencing. This can be a serious problem if more than one table is produced from the same data source (e.g. flexible table generation). When disseminating a large number of tables it is much harder to ensure the consistency of suppressed cells, and care must be taken to ensure that same cells in linked tables are always suppressed.

A simple instance of a Cell Perurbation method is Barnardisation. Barnardisation modifies each internal cell of every table by +1, 0 or -1, according to probabilities. Zeros are not adjusted. The method offers some protection against disclosure by differencing, however table totals are added up from the perturbed internal cells, resulting in inconsistent totals between tables. Typically, the probability p is quite small and therefore a high proportion of risky cells are not modified. The exact proportion of cells modified is not revealed to the user. This is generally a difficult method to implement for flexible output.

The Cell Key Method (CKM, described in Section 5.4) is a much more advanced cell perturbation method which was developed by the Australian Bureau of Statistics (hence it used to be known as ABS Cell Perturbation method) to protect the outputs from their 2006 Census. The method is designed to protect tables by altering potentially all cells by small amounts. The cells are adjusted in such a way that the same cell is perturbed in the same way even when it appears across different tables. This method adds sufficient 'noise' to each cell so if an intruder tried to gather information by differencing, they would not be able to obtain the real data. When integrated into the table generation process, the method provides protection for flexible tables and can be used to produce perturbations for multiple large high dimensional hierarchical tables. It is one of the methods recommended by Eurostat for protection of the Census 2022 output data.

The method is less transparent than other methods, such as, for example, conventional rounding.

Rounding (discussed in Section 5.5) involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell. There are several alternative rounding methods including: conventional rounding, random rounding and controlled rounding. Properties of the different alternative methods (as compared in the summary table Table 5.15 ) vary widely between those variants.

## 5.4 Cell Perturbation - the Cell Key Method

The Cell Key Method is a post tabular perturbative disclosure control method, that adds noise to the original table cell values. Since the individual table cells are equipped with a noise independently, please note that this implies that the resulting table is no longer additive. For example, after perturbation, a population table could show that 1000 males, 1100 females and 16 non-binary persons live in an area, while the total count is 2109 persons. This non-additivity is part of the protective mechanism of this method and at the same time offers the advantage that the deviation from the original value can be kept as small as possible. It is not recommended, generally, to form such aggregates subsequently from perturbed values, because this would also add the sum of all noise terms to the aggregate, which can make the deviation undesirably large.

The Cell Key Method is a more informed post-tabular method of disclosure control since it utilizes pre-tabular microdata information during the perturbation stage. This is to achieve that cells are adjusted in such a way that the same cell is perturbed in the same way even when it appears across different tables. The method is highly dependent on the lookup table used, but it is flexible in that lookup tables can be specifically designed to meet needs, and different lookup tables could potentially be used for different tables. Furthermore, the lookup table can be designed to reflect other post-tabular methods (e.g. small cell adjustments or random rounding). The method provides protection for flexible tables and can be used to produce perturbations for large high dimensional hierarchical tables. As noted above, since perturbation is applied to each table cell independently, additivity gets lost. This is similar to the case of rounding but due to the complexity of the method, those inconsistencies in the data are harder to communicate. Theoretically one might add a post-processing stage to restore additivity, using, for example, an iterative fitting algorithm which may attempt to balance and minimise absolute distances to the stage one table (although not necessarily producing an optimal solution). However, restoring additivity tends to increase the noise, and may cause different perturbation for the same cell when it appears across different tables. It is therefore not generally recommended.

Please note that there is not one ultimate way of how to define Record Key, Cell Key and the lookup table: The Australian Bureau of Statistics for example relies on integer values for their Record Keys whereas the Center of Excellence (CoE) on SDC presented an approach where the Record Keys are uniformly distributed between 0 and 1, which should allow for more flexibility regarding noise design. We will focus on the latter approach here, which is also implemented in the software $\tau$-ARGUS and the R-package '*cellKey*'. In

the variant suggested by the CoE on SDC all digits before the decimal point are removed from the Cell Key, which makes it another random number that is uniformly distributed between 0 and 1. The lookup table now can be interpreted as the tabular representation of a piecewise constant inverse distribution function. By looking up values that are uniformly distributed, we thus obtain realizations of a random variable with the corresponding coded distribution.

It is possible to create lookup tables, which are also known as perturbation tables or p-tables, that are tailored to your needs, by using the freely accesible R-package '*ptable*'. The package allows, among other things, to specify a maximum for the noise you want to add and the probability for the noise to be zero, which is equivalent to retaining the original value. You also have the option to generate the distribution, coded inside the perturbation table, in such a way, that certain values, such as ones or twos, do not occur in the perturbed output at all. The method for creating such tables, implemented in the *ptable* package, is based on a maximum entropy approach as described, for example, in Giessing (2016) and yields a distribution with zero mean. Therefore, the distribution of the data will not get biased by adding the noise. For more information about the *ptable* package, please see the vignette or the reference manual on cran.

The protection effect arises from the uncertainty of a data attacker about whether and, if so, how much a value has been changed. Therefore, all published figures must be perturbed with the Cell Key Method, even those that do not pose a disclosure risk per se. But before the Cell Key Method can be applied, one has to consider, which maximum deviation is still acceptable and how large the variance of the noise should be. But one should always keep in mind that a low maximum deviation also leads to less protection and hence one cannot focus on information loss alone. It is especially risky to publish the maximum deviation, since a data attacker can use this information to draw further conclusions.

Table 5.16: Fictional example of a p-table

| orig. value | pert. value | prob. of occurrence | noise | lower bound | upper bound |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0.5 | -1 | 0 | 0.5 |
| 1 | 2 | 0.5 | 1 | 0.5 | 1 |
| 2 | 2 | 0.8 | 0 | 0 | 0.8 |
| 2 | 3 | 0.2 | 1 | 0.8 | 1 |
| 3 | 2 | 0.3 | -1 | 0 | 0.3 |
| 3 | 3 | 0.4 | 0 | 0.3 | 0.7 |
| 3 | 4 | 0.3 | 1 | 0.7 | 1 |

To illustrate how the Cell Key Method is used, Table 5.16 shows a purely fictional manually created perturbation table with a maximum deviation of 1 and without ones in the results after perturbation. As you can see the values in the colum 'original value' range from 0 to

3. This is because a different distribution is stored in the p-table for each of these values. Otherwise, negative values could arise, for example. This means that within a p-table several probability distributions for the noise are stored, which are used depending on the original value. In the given example for an original value of 1 the noise '$v$' is defined as a uniform distribution on the set $\{-1, 1\}$, whereas for an original value of 2 with a probability of 80% the noise is 0 and with a probability of 20% it is 1. For every original value which is at least 3 the lowest lines in the p table will be used to define the noise '$v$', which encode a symmetric distribution on $\{-1, 0, 1\}$.

Table 5.17: Exemplary Microdata

| ID | Sex | Record Key |
|---|---|---|
| A | male | 0.9 |
| B | male | 0.3 |
| C | male | 0.6 |

Now if we have a set of microdata which contains three male respondents with Record Keys 0.5, 0.3 and 0.4 respectively, as shown in Table 5.17, then in a table cell that aggregates those three respondents the corresponding sum of Record Keys is 0.9+0.3+0.6=1.8. Since for the Cell Key the digits before the decimal point are irrelevant, we get a corresponding Cell Key of 0.8. Now to identify the noise that has to be added to the original count of 3, we have to concentrate on those lines of the p-table, for which the original value is 3 and identify that line for which *'lower bound'* $< 0.8 \leq$ *'upper bound'*. This is the last row of our exemplary table, in which the value 1 is given for the noise. Hence the perturbed count for this cell computes as $\widehat{n} = n + v = 3 + 1 = 4$. At this point, it should be pointed out that if, in addition to frequencies, magnitudes and mean values are also published, the mean values should rather not be shown as original values, since otherwise there is a risk that the corresponding original frequency values can be disclosed. See the discussion in section 5.6.

### 5.4.1 Software implementing the Cell Key Method

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 $\tau$-ARGUS. For information how to use the software, we refer to the vignettes of the respective R-packages (on CRAN), to the manual of $\tau$-ARGUS [**?**] and to the quick references for CKM in $\tau$-ARGUS [**?**].

## 5.5 Rounding

Rounding involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell. It adds a small, but acceptable, amount of distortion to the original data. Rounding is considered to be an effective method for protecting frequency tables, especially when there are many tables produced from one dataset. It provides protection to small frequencies and zero values (e.g. empty cells). The method is simple to implement, and for the user it is easy to understand as the data is visibly perturbed.

Care must be taken when combining rounded tables to create user-defined areas. Cells can be significantly altered by the rounding process and aggregation compounds these rounding differences. Furthermore, the level of association between variables is affected by rounding, and the variance of the cell counts is increased.

There are several alternative rounding methods including; conventional rounding, random rounding, controlled rounding, and semi-controlled rounding, which are outlined below. Each method is flexible in terms of the choice of the base for rounding, although common choices are 3 and 5. All rounded values (other than zeros) will then be integer multiples of 3 or 5, respectively.

When using conventional rounding, each cell is rounded to the nearest multiple of the base. The marginal totals and table totals are rounded independently from the internal cells. An example of conventional rounding is provided below; Table 5.18 shows counts of males and females in different areas, while Table 5.19 shows the same information rounded to a base of 5.

Table 5.18: Population counts by sex

| | Male | Female | Total |
|---|---|---|---|
| Area A | 1 | 0 | 1 |
| Area B | 3 | 3 | 6 |
| Area C | 12 | 20 | 32 |
| Total | 16 | 23 | 39 |

Table 5.19: Population counts by sex (conventional rounding)

| | Male | Female | Total |
|---|---|---|---|
| Area A | 0 | 0 | 0 |
| Area B | 5 | 5 | 5 |
| Area C | 10 | 20 | 35 |
| Total | 15 | 25 | 40 |

The example shows the Males unsafe cell in Area A in Table 5.18 is protected by the rounding process in Table 5.19.

The advantages of this method are that the table totals are rounded independently from the internal cells, and therefore consistent table totals will exist within the rounding base. Cells in different tables which represent the same records will always be the same. While this method does provide some confidentiality protection, it is considered less effective than controlled or random rounding. Tables are not additive (e.g. row 3 of Table 5.17 does not sum to 35) and the level of information is poor if there are many values of 1 and 2. The method is not suitable for flexible table generation as it can be easily 'unpicked' when differencing and linking tables. For these reasons conventional rounding is not recommended as a disclosure control method for frequency tables. Conventional rounding is sometimes used by NSIs for quality reasons (e.g. rounding data from small sample surveys to emphasize the uncertain nature of the data). The distinction between rounding performed for disclosure control reasons and rounding performed for quality reasons should always be made clear to users.

Random rounding shifts each cell to one of the two nearest base values in a random manner. Each cell value is rounded independently of other cells, and has a greater probability of being rounded to the nearest multiple of the rounding base. For example, with a base of 5, cell values of 6, 7, 8, or 9 could be rounded to either 5 or 10. Marginal totals are typically rounded separately from the internal cells of the table (i.e. they are not created by adding rounding cell counts) and this means tables are not necessarily additive. Various probability schemes are possible, but an important characteristic is that they should be unbiased. This means there should be no net tendency to round up or down and the average difference from the original counts should be zero.

If we are rounding to base 3 the residual of the cell value after dividing by 3 can be either 0, 1 or 2.

- If the residual is zero no change is made to the original cell value.
- If the residual is 1, then with a probability of 2/3 the cell value is rounded down to the lower multiple of 3 and with a probability of 1/3 the cell value is rounded up to the higher multiple of 3.
- If the residual is 2, the probabilities are 2/3 to round up and 1/3 to round down.

| Original Value | Rounded Value (probability) |
|---|---|
| 0 | 0 (1) |
| 1 | 0 (2/3) or 3 (1/3) |
| 2 | 3(2/3) or 0(1/3) |
| 3 | 3 (1) |
| 4 | 3(2/3) or 6(1/3) |
| 5 | 6(2/3) or 3(1/3) |

| Original Value | Rounded Value (probability) |
|---|---|
| 6 | 6 (1) |

As an example, Table 5.21 shows a possible solution for Table 5.18 using random rounding to base 5.

Table 5.21: Population counts by sex (with random rounding)

| | Male | Female | Total |
|---|---|---|---|
| Area A | 0 | 0 | 0 |
| Area B | 5 | 0 | 5 |
| Area C | 10 | 20 | 35 |
| Total | 15 | 20 | 40 |

The main advantages of random rounding are that it is relatively easy to implement, it is unbiased, and it is clear and transparent to users. Table totals are consistent within the rounding base because the totals are rounded independently from the internal cells. All values of 1 and 2 are removed from the table by rounding, which prevents cases of perceived disclosure as well as actual disclosure. The method may also provide some protection against disclosure by differencing as rounding should obscure most of the exact differences between tables.

However, random rounding has disadvantages including the increased information loss which results from the fact that all cells (even safe cells) are rounded. In some instances the protection can be 'unpicked' and in order to ensure adequate protection, the resulting rounded tables need to be audited. Although the method is unbiased, after applying random rounding there may be inconsistencies in data within tables (e.g. rows or columns which do not add up like row 3 of Table 5.21 does not sum to 35) and between tables (e.g. the same cell may be rounded to a different number in different tables).

Unlike other rounding methods, controlled rounding yields additive rounded tables. It is the statistical disclosure control method that is generally most effective for frequency tables. The method uses linear programming techniques to round cell values up or down by small amounts, and its strength over other methods is that additivity is maintained in the rounded table, (i.e. it ensures that the rounded values add up to the rounded totals and sub-totals shown in the table). This property not only permits the release of realistic tables which are as close as possible to the original table, but it also makes it impossible to reduce the protection by 'unpicking' the original values by exploiting the differences in the sums of the rounded values. Another useful feature is that controlled rounding can achieve specified levels of protection. In other words, the user can specify the degree of ambiguity added to the cells, for example, they may not want a rounded value within 10%

of the true value. Controlled rounding can be used to protect flexible tables although the time taken to implement the method may make it unsuitable for this purpose.

Table 5.22 shows a possible rounding solution for Table 5.18, using controlled rounding to base 5.

Table 5.22: Population counts by sex (controlled rounding)

| | Male | Female | Total |
|---|---|---|---|
| Area A | 5 | 0 | 5 |
| Area B | 0 | 5 | 5 |
| Area C | 10 | 20 | 30 |
| Total | 15 | 25 | 40 |

The disadvantages of controlled rounding are that it is a complicated method to implement, and it has difficulty coping with the size, scope and magnitude of the census tabular outputs. Controlled rounding is implemented in the software package $\tau$-ARGUS, see Section 5.5.1 below for detailed information. Nevertheless, it is hard to find control-rounded solutions for sets of linked tables, and in order to find a solution cells may be rounded beyond the nearest rounding base. In this case users will know less about exactly how the table was rounded and it is also likely to result in differing values for the same internal cells across different tables.

Semi-controlled rounding also uses linear programming to round table entries up or down but in this case it controls for the overall total in the table, or it controls for each separate output area total. Other marginal and sub totals will not necessarily be additive. This ensures that either the overall total of the table is preserved (or the output area totals are all preserved), and the utility of this method is increased compared with conventional and random rounding. Consistent totals are provided across linked tables, and therefore the method can be used to protect flexible tables, although the time it takes to implement may make it unsuitable. Disadvantages of semi-controlled rounding relate to the fact that tables are not fully additive, and finding an optimal solution can prove difficult.

Table 5.23: Summary of SDC rounding methods

| | Conventional Rounding | Controlled (and semi-controlled) Rounding | Random rounding |
|---|---|---|---|
| Internal cells add to table totals (additvity) | No | Yes | No |

| | Conventional Rounding | Controlled (and semi-controlled) Rounding | Random rounding |
|---|---|---|---|
| Method provides enough SDC protection (and cannot be unpicked) | No | Yes | In some situations this method can be unpicked |
| Method is quick and easy to implement | Yes | It can take time for this method to find a solution | Yes |

There are some more specialised rounding methods which have been used at various times by NSIs to protect census data, two of these methods are described below.

Small cell adjustment was used (in addition to random swapping (a pre-tabular method)) to protect 2001 Census tabular outputs for England, Wales and Northern Ireland. This method was also used by the ABS to protect their tabular outputs from the 2001 Census.

Applying small cell adjustments involves randomly adjusting small cells within tables upwards or downwards to a base using an unbiased prescribed probability scheme. During the process:

- small counts appearing in a table cells are adjusted
- totals and sub totals are calculated as the sum of the adjusted counts. This means all tables are internally additive.
- tables are independently adjusted so counts of the same population which appear in two different tables, may not necessarily have the same value.
- tables for higher geographical levels are independently adjusted, and therefore will not necessarily be the sum of the lower component geographical units.
- output is produced from one database which has been adjusted for estimated under-count so the tables produced from this one database provide a consistent picture of this one population.

Advantages of this method are that tables are additive, and the elimination of small cells in the table removes cases of perceived as well as actual identity disclosure. In addition, loss of information is lower for standard tables as all other cells remain the same, however information loss will be high for sparse tables. Other disadvantages include inconsistency of margins between linked tables since margins are calculated using perturbed internal cells, and this increases the risk of tables being unpicked. Furthermore, this method provides little protection against disclosure by differencing, and is not suitable for flexible table generation.

### 5.5.1 Software - How to use Controlled Rounding in $\tau$-ARGUS

$\tau$-ARGUS (Hunderpool et al, 2005) is a software package which provides tools to protect tables against the risk of statistical disclosure ($\tau$-ARGUS is also discussed in Chapter 4). Controlled rounding is easy to use in $\tau$-ARGUS and the controlled rounding procedure (CRP) was developed by JJ Salazar. This procedure is based on optimisation techniques similar to the procedure developed for cell suppression. The CRP yields additive rounded tables, where the rounded values add up to the rounded totals and sub-totals shown in the table. This means realistic tables are produced and it makes it impossible to reduce the protection by "unpicking" the original values by exploiting the differences in the sums of the rounded values. The CRP implemented in $\tau$-ARGUS also allows the specification of hierarchical structures within the table variables.

Controlled rounding gives sufficient protection to small frequencies and creates uncertainty about the zero values (*i.e.* empty cells). (This is not the case for suppression in terms of how it is now implemented in $\tau$-ARGUS).

In Zero-restricted Controlled Rounding cell counts are left unchanged if they are multiples of the rounding base or shifted to one of the adjacent multiples of the rounding base. The modified values are chosen so that the sum of the absolute differences between the original values and the rounded ones are minimized (under an additivity constraint). Therefore, some values will be rounded up or down to the most distant multiple of the base in order to help satisfy these constraints. In most cases a solution can be found, but in some cases it cannot and the zero-restriction constraint in CRP can be relaxed to allow the cell values to be rounded to a nonadjacent multiple of the base. This relaxation is controlled by allowing the procedure to take a maximum number of *steps*.

For example, consider rounding a cell value of 7 when the rounding base equals 5. In zero-restricted rounding, the solution can be either 5 or 10. If 1 step is allowed, the solution can be 0, 5, 10 or 15. In general, let $z$ be the integer to be rounded in base $b$, then this number can be written as

$z = ub + r,$

where $ub$ is the lower adjacent multiple of $b$ (hence $u$ is the floor value of $z/b$) and $r$ is the remainder. In the zero-restricted solution the rounded value, $a$, can take values:

$$a = \begin{cases} a = ub & \text{if} \quad r = 0 \\ a = \begin{cases} ub \\ (u+1)b \end{cases} & \text{if} \quad r \neq 0. \end{cases} \tag{5.1}$$

If $K$ steps are allowed, then $a$, can take values:

$$a = \begin{cases} \max_{j \in -K,\ldots,K}(0,(u+j)) \cdot b & \text{if} \quad r = 0 \\ \max_{j \in -K,\ldots,K+1}(0,(u+j)) \cdot b & \text{if} \quad r \neq 0 \end{cases} \tag{5.2}$$

#### 5.5.1.1 Optimal, first feasible and RAPID solutions

For a given table there can exist more than one controlled rounded solution, and any of these solutions is a *feasible* solution. The Controlled Rounding Program embedded in $\tau$-ARGUS determines the *optimal* solution by minimising the sum of the absolute distances of the rounded values, from the original ones. Denoting the cell values, including the totals and sub-totals, with $z_{\mathrm{i}}$ and the corresponding rounded values with $a_{\mathrm{i}}$, the function that is minimised is

$$\sum_{i=1}^{N} |z_i - a_i| \quad ,$$

where $N$ is the number of cells in a table (including the marginal ones). The optimisation procedure for controlled rounding is a rather complex one (*NP*-complete program), so finding the optimal solution may take a long time for large tables. In fact, the algorithm iteratively builds different rounded tables until it finds the optimal solution. In order to limit the time required to obtain a solution, the algorithm can be stopped when the first feasible solution is found. In many cases, this solution is quite close to the optimal one and it can be found in significantly less time.

The RAPID solution is produced by CRP as an approximated solution when a feasible one cannot be found. This solution is obtained by rounding the internal cells to the closest multiple of the base and then computing the marginal cells by addition. This means that the computed marginal values can be many jumps away from the original value. However, a RAPID solution is produced at each iteration of the search for an optimal solution, and it will improve (in terms of the loss function) over time. $\tau$-ARGUS allows the user to stop CRP after the first RAPID solution is produced, but this is likely to be very far away from the optimal one.

#### 5.5.1.2 Protection provided by controlled rounding

The protection provided by controlled rounding can be assessed by considering the uncertainty (about the true values achieved) when releasing rounded values; that is the existence interval that an intruder can compute for a rounded value. We assume that the values of the rounding base, $b$, and the number of steps allowed, $K$, are known by the user together with the output rounded table. Furthermore, we assume that it is known that the original values are positive frequencies (hence nonnegative integers).

**Zero-restricted rounding.**

Given a rounded value, $a$, an intruder can compute the following existence intervals for the true value, $z$:

$$z \in \begin{cases} [0, b-1] & \text{if} \quad a = 0 \\ [a-b+1, a+b-1] & \text{if} \quad a \neq 0 \end{cases} \tag{5.3}$$

For example, if the rounding base is $b = 5$ and the rounded value is $a = 0$, a user can determine that the original value is between 0 and 4. If the rounded value is not 0, then users can determine that the true value is between $\pm 4$ units from the published value.

**K-step rounding**

As mentioned above, it is assumed that the number of steps allowed is released together with the rounded table. Let $K^*$ be the number of steps allowed, then an intruder can compute the following existence intervals for the true value $z$:

$$z \in \begin{cases} [0, (K+1)b-1] & \text{if} \quad a < (K+1)b \\ [a-(K+1)b+1, a+(K+1)b-1] & \text{if} \quad a \geq (K+1)b \end{cases} \tag{5.4}$$

For example, assume that for controlled rounding with $b = 5$ and $K = 1$, $a = 15$, then a user can determine that $z \in [6, 24]$.

**Very large tables**

The procedure implemented in $\tau$-ARGUS is capable of rounding tables up to 150K cells on an average computer. However for larger tables a partitioning procedure is available, which allows much larger tables to be rounded. Tables with over six million cells have been successfully rounded this way.

## 5.6 Targeted Record Swapping

Targeted Records Swapping (TRS) is a pre-tabular perturbation method. It's intended use is to apply a swapping procedure to the micro data before generating a table. Although it is applied solely on micro data it is generally considered a protection method used for tabular data and not recommended for protecting micro data. TRS can be used for tables with and without spatial characteristics, with the prior case containing also grid data products or tables created by cross-tabulating with grid cells.
During the TRS the spatial character of the data can be taken into account to some degree.

### 5.6.1 The TRS noise mechanism

Expert level

Regardless of the table, be it count data or a magnitude table, the methodology of the TRS does not change. This is a direct consequence of the fact that the method is applied to the underlying micro data before generating any table.
Consider population units $i = 1, \dots, N$ where each unit $i$ has $p$ characteristics or variables $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbf{X} \in \mathbb{R}^{n \times p}$. Furthermore there exists a geographic hierarchy $\mathcal{G}^1 \succ \mathcal{G}^2 \succ \dots \succ \mathcal{G}^K$ where each $\mathcal{G}^k$ is the set of disjointly split areas $g^k_m$, $m = 1, \dots, M_k$ and each $g^k_m$ is further disjointly subdivided into smaller areas $g^{k+1}_m$,$m = 1, \dots, M_{k+1}$:

$$\mathcal{G}^k = \{g^k_m \mid g^k_i \cap g^k_j = \emptyset \text{ for } i \neq j\} \quad \forall k = 1, \dots, K$$

where

$$g^k_m = \dot{\bigcup}_{m=1}^{M_{k+1}} g^{k+1}_m \quad \forall k = 1, \dots, K-1 \quad .$$

The notation $a \dot{\cup} b$ refers to the disjoint union meaning that $a \cap b = \emptyset$.
With the above definition each unit $i$ in the population can be assigned to a single area $g^k_{m_i}$ for each geographic hierarchy level $\mathcal{G}^k$, $k = 1, \dots, K$. Consider as geographic hierarchy for example the NUTS regions, NUTS1 $\succ$ NUTS2 $\succ$ NUTS3, or grid cells, 1000m grid cells $\succ$ 500m grid cells $\succ$ 250m grid cells.
Given the geographic hierarchy levels $\mathcal{G}^k$, $k = 1, \dots, K$ calculate for each unit $i = 1, \dots, N$ risk values $r_{i,k}$. As an example one can choose $k$-anonymity as risk measure and a subset of $Q$ variables $\mathbf{x}_{q_1}, \dots, \mathbf{x}_{q_Q}$ to derive risk values $r_{i,k}$. They can be defined by calculating the number of units $j$ which have the same values for variables $\mathbf{x}_{q_1}, \dots, \mathbf{x}_{q_Q}$ as unit $i$ and taking the inverse.

$$c_{i,k} = \sum_{j=1}^{N} \mathbf{1}[x_{i,q_1} = x_{j,q_1}, x_{i,q_2} = x_{j,q_2}, \dots, x_{i,q_Q} = x_{j,q_Q}]$$

$$r_{i,k} = \frac{1}{c_{i,k}}$$

Having the risk values $r_{i,k}$ for each unit $i$ and each geographic hierarchy level calculated the TRS can be defined as follows:

1. Define initial, use-case specific, parameter.
   - A global swap rate $p$;
   - Define a risk value $r_{high}$ beyond which all units with $r_{i,k}$ are considered **high risk** for the geographic hierarchy level $k$;
   - A subset of $T$ variables $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_T}$ which are considered while swapping units;

2. Begin at the first hierarchy level $\mathcal{G}^1$ and select all units $j$ for which $r_{i,1} \geq r_{high}$.
3. For each $j$ select all units $l_1, \dots, l_L$, which do not belong to the same area $g^1_{m_j}$ and have the same values for variables $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_T}$ as unit $j$. In addition units $l_1, \dots, l_L$ cannot have been swapped already.

$$g^1_{m_j} \neq g^1_{m_l}$$

$$x_{j,t_1} = x_{l,t_1}, x_{j,t_2} = x_{l,t_2}, \dots, x_{j,t_T} = x_{l,t_T}$$

4. Sample for each $j$ one unit from the set $\{l \mid g^1_{m_j} \neq g^1_{m_l} \wedge x_{j,t_1} = x_{l,t_1} \wedge, \dots, \wedge x_{j,t_T} = x_{l,t_T}\}$ by normalising corresponding risk value $r_{l,1}$ and using them as sampling probabilities.
    - Previously swapped units should be excuded from this set.
5. Swap all variables, holding geographic information in $\mathbf{X}$, between unit $j$ and the sampled unit.
    - Some implementation of targeted record swapping consider only swapping specific variable values from $\mathbf{X}$ between $j$ and the sampled unit.
6. Iterate through the geographic hierarchies $k = 2, \dots, K$ and repeat in each of them steps 3. - 5.
7. At the final geographic hierarchy $k = K$ if the number of already swapped units is less than $p \times N$ additional units are swapped to reach $p \times N$ overall swaps.

If the population units refer to people living in dwellings and the aim is to swap only full dwellings with each other and not only individuals it can be useful to set

$$r_{i,k} = \max_{j \text{ living in same dwelling as } i} r_{j,k}$$

prior to applying the swapping procedure. In addition $\mathbf{x}_{1(q)}, \dots, \mathbf{x}_q$ should be defined such that they refer to variables holding dwelling information.
The above described procedure is implemented in the R package `sdcMicro` as well as in the software `muArgus`, alongside a multitude of parameters to fine tune the procedure.

### 5.6.2 Pros and cons of targeted record swapping

Indicated by the name of the method the TRS aims to swap micro data records prior to building a table and specifically targeting records during the swapping procedure which have a higher risk of disclosure with respect to the final tables. The protection of the TRS itself is considered to be the uncertainty that an identified unit $i$ has a considerable chance of actually being a swapped unit and that the information derived from this unit does not

contain the information of the original unit $i$. In general it is recommended to apply the TRS on the micro data set only once and afterwards build various tables from the same perturbed micro data. This creates a more drastic trade off between the number of records to swap and the utility of the final tables. The swapping procedure can indirectly take into account the structure of the final tables through the risk value derived from the subest of variables $\mathbf{x}_{q_1}, \dots, \mathbf{x}_{q_Q}$ and choice of the geographic hierarchy. However a large number of variables $\mathbf{x}_{q_1}, \dots, \mathbf{x}_{q_Q}$ and a high resolution in the geographic hierarchy can result in a high share of units with high risk values and consequently many potential swaps. A high swap rate, for instance beyond 10%, can quickly lead to high information loss, because the noise introduced through the swapping is not controlled for while drawing the swapped units. Thus it is not feasible to both address all possible disclosure risk scenarios while maintaining high utility in the final tables.

As with any method it is advised to thoroughly tune parameters to balance information loss and disclosure risk. Possible tuning parameters are:

- The geographic hierarchy and its depth of granularity.
- The construction of the risk values $r_{i,k}$ and $r_{high}$
- The choice of $\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_T}$
- The global swap rate $p$

Because the noise is applied to the microdata before building any tables the additivity between inner and marginal aggregates will always be preserved.

## 5.7 Publication of mean values

In this section, we will explain why original means should rather not be published when the Cell Key Method (or any other non-additive, perturbative SDC method, such as, e.g. Rounding) is applied to frequency counts. We will illustrate this with an example and show ways to publish the mean values in a safer variant. For this purpose we consider a certain population group of size $n$ and for every person $i \in 1, \dots, n$ of this population we denote their corresponding age with $x_i$. So the average age of this population group can be written as $(\sum_{i=1\dots n} x_i)/n$. We now consider the following example scenario, in which both (perturbed) frequencies and (original) mean values are published. Table 5.24 shows both the perturbed and the original frequency counts, as well as information about the age.

Table 5.24: An example table with (perturbed) ages and counts

| | Cell 1 | Cell 2 | Marginal |
|---|---|---|---|
| Original Count ($n$) | 8 | 12 | 20 |
| Perturbed Count ($\hat{n}$) | 9 | 14 | 19 |
| Original Sum of Ages ($x$) | 90 | 95 | 185 |

| | Cell 1 | Cell 2 | Marginal |
|---|---|---|---|
| Original Mean of Ages ($x/n$) | 11.25 | 7.9167 | 9.25 |

Attackers now have the perturbed frequencies as well as the original mean values at their disposal. In our attack scenario, we also assume that attackers know that the maximum deviation of the frequency count is 2. Attackers can therefore conclude that the original case numbers of the inner cells must be between 7 and 11 and between 12 and 16, respectively, and that the marginal must have originally had a value between 17 and 21. For an attacker, this results in the following possible combinations:

- The marginal is originally 19 and the inner cells are
  - 7 and 12
- The marginal is originally 20 and the inner cells are
  - 7 and 13
  - 8 and 12
- The marginal is originally 21 and the inner cells are
  - 7 and 14
  - 8 and 13
  - 9 and 12

The attackers can now multiply the original mean values of the table cells known to them with the thus calculated candidates for the associated frequencies. In this way, they obtain estimates for the original magnitude value for each cell. If now for each of those estimates, they sum up those values for the inner cells, they obtain another estimated value for the marginal value. This can now be used to identify the correct combination of frequency values, since for the correct ones the sum over the inner cell values is identical to the marginal value, as shown in Table 5.25.

Table 5.25: Sample calculation for an attacker

| Cell 1 | Cell 2 | Marginal | Est. Cell 1 | Est. Cell 2 | Sum of Estimates | Est. Marginal |
|---|---|---|---|---|---|---|
| 7 | 12 | 19 | 78.75 | 95 | 173.75 | 175.75 |
| 7 | 13 | 20 | 78.75 | 102.917 | 181.667 | 185 |
| 8 | 12 | 20 | 90 | 95 | 185 | 185 |
| 7 | 14 | 21 | 78.75 | 110.833 | 189.583 | 194.25 |
| 8 | 13 | 21 | 90 | 102.917 | 192.917 | 194.25 |
| 9 | 12 | 21 | 101.25 | 95 | 196.25 | 194.25 |

Additionally, through this calculation, the associated magnitude values are now known as well. If these are confidential, a further problem arises. The publication of original mean values is therefore not recommended. So, when using the Cell Key Method for frequency tables we recommend to use those perturbed counts also when generating mean values, i.e. if $n$ is an original count, $\hat{n}$ is the corresponding perturbed count and $m$ is the corresponding magnitude value, as it gets published, then, in order to avoid the disclosure risk described here, it is better to calculate the mean as $m/\hat{n}$.

## 5.8 Information loss

As described in Sections 5.3 to 5.5 there are a number of different disclosure control methods used to protect frequency tables. Each of these methods modifies the original data in the table in order to reduce the disclosure risk from small cells (0's, 1's and 2's). However, the process of reducing disclosure risk results in information loss. Some quantitative information loss measures have been developed by Shlomo and Young (2005 & 2006) to determine the impact various statistical disclosure control (SDC) methods have on the original tables.

Information loss measures can be split into two classes: measures for data suppliers, used to make informed decisions about optimal SDC methods depending on the characteristics of the tables; and measures for users in order to facilitate adjustments to be made when carrying out statistical analysis on protected tables. Here we focus on measures for data suppliers. Measuring utility and quality for SDC methods is subjective. It depends on the users, the purpose of the statistical analysis, and on the type and format of the data itself. Therefore it is useful to have a range of information loss measures for assessing the impact of the SDC methods.

The focus here is information loss measures for tables containing frequency counts; however, some of these measures can easily be adapted to microdata. Magnitude or weighted sample tables will have the additional element of the number of contributors to each cell of the table.

When evaluating information loss measures for tables protected using cell suppression, one needs to decide on an imputation method for replacing the suppressed cells similar to what one would expect a user to do prior to analysing the data (i.e. we need to measure the difference between the observed and actual values, and for suppressed cells the observed values will be based on user inference about the possible cell values). A naive user might use zeros in place of the suppressed cells whereas a more sophisticated user might replace suppressed cells by some form of averaging of the total information that was suppressed, or by calculating feasibility intervals.

A number of different information loss measures are described below, and more technical details can be found in Shlomo and Young (2005 & 2006).

- An exact Binomial Hypothesis Test can be used to check if the realization of a random stochastic perturbation scheme, such as random rounding, follows the expected probabilities (i.e. the parameters of the method). For other SDC methods, a non-parametric signed rank test can be used to check whether the location of the empirical distribution has changed after the application of the SDC method.
- Information loss measures that measure distortion of distributions are based on *distance metrics* between the original and perturbed cells. Some useful metrics are also presented in Gomatam and Karr (2003). A distance metric can be calculated for internal cells of a table. When combining several tables one may want to calculate an overall average across the tables as the information loss measure. These distance metrics can also be calculated for totals or sub-totals of the tables.
- SDC methods will have an impact on the variance of the average cell size for the rows, columns or the entire table. The variance of the average cell size is examined before and after the SDC method has been applied. Another important variance to examine is the "between"-variance when carrying out a one-way ANOVA test based on the table. In ANOVA, we examine the means of a specific target variable within groupings defined by independent categorical variables. The goodness of fit statistic $R^2$ for testing the null hypothesis that the means are equal across the groupings is based on the variance of the means between the groupings divided by the total variance. The information loss measure therefore examines the impact of the "between"-variance and whether the means of the groupings have become more homogenized or spread apart as a result of the SDC method.
- Another statistical analysis tool that is frequently carried out on tabular data are tests for independence between categorical variables that span the table. The test for independence for a two-way table is based on a Pearson Chi-squared statistic and the measure of association is the Cramer's V statistic. For multi-way tables, one can examine conditional dependencies and calculate expected cell frequencies based on the theory of log-linear models. The test statistic for the fit of the model is also based on a Pearson Chi-squared statistic. SDC methods applied to tables may change the results of statistical inferences. Therefore we examine the impact to the test statistics before and after the application of the SDC method.
- Another statistical tool for inference is the Spearman's rank correlation. This is a technique that tests the direction and strength of the relationship between two variables. The statistic is based on ranking both sets of data from the highest to the lowest. Therefore, one important assessment of the impact of the SDC method on statistical data is whether we are distorting the rankings of the cell counts.

In order to allow data suppliers to make informed decisions about optimal disclosure control methods, ONS has developed a user-friendly software application that calculates both disclosure risk measures based on small counts in tables and a wide range of information loss measures (as described above) for disclosure controlled statistical data, Shlomo and Young (2006). The software application also outputs R-U Confidentiality maps.

## 5.9 References

Brown, D., (2003) *Different approaches to disclosure control problems associated with geography, ONS*, United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians, Working Paper No. 14.

Doyle, P., Lane, J.I., Theeuwes, J.J.M. and Zayatz, (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* Elsevier Science BV.

Duke-Williams, O. and Rees, P., (1998) *Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure*, International Journal of Geographical Information Science 12, 579-605

Enderle, T., Giessing, S., Tent, R., (2020) *Calculation of Risk Probabilities for the Cell Key Method. In: Domingo-Ferrer, J., Muralidhar, K. (eds) Privacy in Statistical Databases*, Lecture Notes in Computer Science(), vol 12276. Springer, Cham. https://doi.org/10.1007/978-3-030-57521-2_11

Gomatam, S. and A. Karr (2003), *Distortion Measures for Categorical Data Swapping,* Technical Report Number 131, National Institute of Statistical Sciences.

Salazar, JJ, Staggermeier, A and Bycroft, C, (2005) *Controlled rounding implementation*, Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva

Shlomo, N. (2007) Statistical Disclosure Control Methods for Census Frequency Tables. International Statistical Review, Volume 75, Number 2, August 2007 , pp. 199-217(19) Blackwell Publishing.

Shlomo, N. and Young, C. (2005) *Information Loss Measures for Frequency Tables,* Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva.

Shlomo, N. and Young, C. (2006) *Statistical Disclosure Control Methods Through a Risk-Utility Framework,* PSD'2006 Privacy in Statistical Databases, Springer LNCS proceedings, to appear.

Willenborg, L., and Ton de Waal. (1996) *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics no 111 Springer-Verlag. New York.

# 6 Remote access issues

## 6.1 Introduction

Although very sophisticated methods have been developed to make safe microdata files, the needs of serious researchers for more detailed data cannot be met by these methods. It is simply impossible to release these very detailed microdata sets to users outside the control of the NSIs without breaching the necessary confidentiality protection. Nevertheless the NSIs recognize the serious and respectable requests by the research community for access to the very rich and valuable microdata sets of the NSIs. Therefore different initiatives have been taken by the NSIs to meet these needs.

The first step was the creation of Research Data Centres (RDCs), a special room in the NSI, where researchers can analyse the data sets, without the option to export any information without the consent of the NSI. In parallel to this initiative there are options for remote execution. Remote execution facilities are various kinds of systems where researchers can submit scripts for SAS, SPSS etc to the NSI. Remote access, where users can "log in" to a RDC from a remote desktop, has become commonly used.

As all these options allow the researcher to access unprotected sensitive data in some way, all possible precautions have to be taken. These options are certainly not available to the general public, but only to selected research institutes like universities and similar research institutes. Additionally, strict contracts have to be signed between the NSI and the researcher. Preferably also the research institute itself should sign the contract. This enables the NSI to take action against the institute itself as well as against the researcher, if something might go wrong. A common repercussion for the institute could be a ban for the whole institute to access these facilities. So it will be in the interest of the institute to ensure a correct behaviour of the researcher.

## 6.2 Research Data Centres (RDCs)

In order to meet the needs of the researcher community to analyse the rich datasets compiled by the NSIs, while safeguarding the confidentiality constraints, the first solution was to create special rooms in the premises of the NSIs (RDCs). The NSI makes available special computers for the researchers. On this computer the necessary software for the research will be installed by the NSI together with the necessary datasets. Ideally these

computers have no connection whatsoever to the internet and there is no email. Also drives for removable discs are not available and the use of memory sticks has to be blocked. The access to the internal production network of the NSI has to be blocked as well, preventing the possibilities of the researchers to access other sensitive information. Installing a printer is a risk as well as is the use of phones. Supervision of the RDC is always needed.

The datasets to be used in the RDC have to be anonymised (i.e. at least the name, address etc are removed). It is also advisable to restrict the variables available to the set that is needed for the specific research.

On these computers the researchers should nevertheless be able to fully analyse the data files and complete their analysis. When the research is finished the results have to be released to the researchers. Before this can be done, NSI staff has to check the research results. Unfortunately this is not a straightforward, easy task. This will be discussed in section 6.6.

The concept of RDCs is meeting many research needs and several NSIs have adopted this idea. RDCs have been implemented in the USA, Canada, The Netherlands, Italy, Germany, Denmark, Eurostat and several other countries.

The concept of RDCs has proved to be very successful. Many good research papers and theses have been completed for which these centres were indispensable. However there are some drawbacks. The most important one is that the researchers have to come physically to the premises of the NSIs. Even in a small country like the Netherlands, this is seen as a serious problem. Also the researcher cannot just try another option when he is back at his normal working place, because he has to travel to the NSI first. Also the fact that he cannot work in his normal working environment is considered a drawback.

## 6.3 Remote execution

As modern communication techniques have become available, the NSIs have investigated the possibilities to use these techniques. The first initiative is remote execution. In this concept the researchers will get a full description of all the metadata of the datasets available for research. However the dataset available will remain on the computers of the NSIs. The researchers will prepare scripts for analysing the datasets (with SPSS, SAS etc) and send them to the NSI (by email or via some internet page). The NSI will then check the script (e.g. for commands like List Cases, but also other unwanted actions like printing the residuals of a regression) before running it and after a second check send back the results to the researcher.

For the researcher this system has the advantage that he does no longer have to travel to the NSI. He can send a script whenever he wants. On the other hand he cannot directly run the script since this is done by the NSI. Hence correcting errors in a script can take much more time, depending on the turn around time of the NSI. This process could be

speeded up if the NSI will make available a fake dataset which corresponds to the original file in terms of structure but not in content. The main objective of this dataset is to avoid all unsuccessful submissions due to syntax errors etc.

For the researcher remote execution has several advantages (no need for travel) but also some drawbacks (slow turn around time). For the NSIs it is very time-consuming, as they have to check so many scripts and results. It is not uncommon in statistical analysis that several scripts are submitted and executed. But then the outcome proves to be not the optimal model and a new script is submitted. However the NSI does not know in advance which script is successful and has to check everything. This is very time-consuming if the NSI takes this seriously.

Examples of this kind of systems are the Luxembourg Income Study (Lissy) and the Australian RADL.

## 6.4 Remote access

Systems for remote access have become common over the years. The aim is to combine the flexibility for researchers to do all their analysis in a RDC while removing the constraints of travelling to the NSIs. Modern developments in the internet make it possible to set up a safe controlled connection, a VPN (virtual private network). A VPN is a technique to setup a secure connection between the server at the NSI and a computer of the researcher. It uses firewalls and encryption techniques. Also additional procedures to control the login procedure like software tokens or biometrics can be used to secure the connection. The most well know product behind this is Citrix, but other systems exist as well. Citrix has been developed to set up safe access to business networks over the internet without giving access to unauthorised persons. This will safeguard the confidentiality of the information on this network.

Some NSIs are using Citrix to set up a safe connection between the PC of the researcher and a protected server of the NSI. This approach was followed by Denmark, Sweden and the Netherlands. Slovenia is using Windows remote desktop services (similar to Citrix). Statistics Netherlands is currently using VMware Horizon Client (also similar to Citrix).

The main idea of a remote facility is that it should resemble the 'traditional' OnSite RDCs as much as possible, concerning confidentiality aspects.

The following aspects have to be taken into account:

1. Only authorized users should be able to make use of this facility,
2. Microdata should remain at the NSI,
3. Desired output of analyses should be checked on confidentiality,
4. Legal measures have to be taken when allowing access.

The key issue is that the microdata set remains in the controlled environment of the NSI, while the researcher can do the analysis in his institute. In fact it is an equivalent of the RDC. The Citrix connection will enable the researcher to run SPSS, SAS etc on the server of the NSI. The researcher will only see the session on his screen. This allows him to see the results on his analysis but also the microdata itself. This is completely equivalent to what he can see, if he would be at the RDC.

Citrix will only send the pictures of the screens to the PC of the researcher, but no data is send to him. Even copying the data from the screen to the hard disk is not possible. If the researcher is satisfied with some analysis and wants to use the results in his report, he should make a request to the NSI to release these results to him. The NSI has to check the output for disclosure risks and if this is OK the NSI will send the results to the researcher.

As the researchers will work with very sensitive data, all measures should be taken to ensure the confidentiality of the data. Therefore also legal measures have to be taken, binding not only the researcher himself but also the institute.

## 6.5 Licensing

Another access option for microdata releases available to NSIs is to release data under licence or access agreements. A spectrum of different data access arrangements can be provided. A variety of factors should be taken into account when granting approval for access – including the purpose of the access, the status of the user, the legal framework, the status of the data, the availability of facilities and the history of access. The levels of control over use and user applied within the licence should be balanced by the level of detail and/or perturbation in the microdata.

## 6.6 Confidentiality protection of the analysis results

### 6.6.1 Output checking

Output checking is the process of checking the disclosure risk of research results based on microdata files made available in RDCs. NSIs and other institutions can establish their own rules for output checking.

In 2009, a document 'Guidelines for the checking of output based on microdata research' was prepared within the European project ESSnet SDC. In 2015, this document was a basis for a document 'Guidelines for Output Checking' prepared within the DwB (Data without Boundaries) project. Both documents provide guidelines and practical advice for output checkers. Principles-based model and rule-of-thumb model are described; the former considers the entire context of the output, while the latter is based on strict rules.

The overall rule of thumb is defined and its application to different types of output is described. Organisational aspects of output checking are discussed.

### 6.6.2 Rules for designing programs for controlled teleprocessing using microdata of official statistics in Germany

#### 6.6.2.1 Introduction

As in many other countries, German official data are subject to statutory data protection regulations. Therefore, results produced on the basis of statistical data are checked for confidentiality and critical values are suppressed. This applies both to publications of the statistical offices and to the results of scientific projects supported by the research data centres (RDC) of German official statistics. Since the research data centres were set up in Germany in 2001, microdata requests from the scientific community have considerably increased. In this context, researchers do their own analyses of microdata coming from the wide range of statistics offered by the research data centres. In Germany, various ways of data access have been established for researchers. A way of access frequently used is controlled teleprocessing - or controlled remote data processing - , which runs on non-anonymised original data thus giving the researcher the opportunity to exploit the full information content of official microdata. For controlled remote data processing the user first of all gets a data structure file from the research data centre whose structure is that of the original file, but whose content is not. Based on that file, the data user develops a program code using the statistics software SAS, SPSS or STATA, checks it for errors and sends it to the research data centre via e-mail. The output of his program is then checked for confidentiality and returned also by e-mail. Checking the produced results for confidentiality is done manually.

Manual confidentiality checks are time-consuming and labour-intensive, as is shown in chapter 6.3. The responsible staff member must be familiar with both the analysed statistics and the specific research project. He must ensure primary and secondary confidentiality and keep track of the secondary confidentiality protection performed even with repeated outputs. In particular, he must check the transmitted program code for intentional and unintentional disclosure of the data. The experience of the research data centres of official statistics in Germany shows that the projects dealt with are very different in terms of volume and complexity of the transmitted programs. In some cases, there is much need for clarification and, depending on how the project proceeds, a variety of individual arrangements must be made with the data user. A more or less long start-up and acquaintance phase on both sides is necessary in every project. Combining that experience at the research data centres has led to the wish to reduce both the adjustment phase and the checking times by standardising the program design.

#### 6.6.2.2 Rules for program design

***Goals***

To simplify project handling, and especially the confidentiality checks, the research data centres of the Federation and the Länder have developed a catalogue of rules for program design when using controlled teleprocessing[9]. Its purpose is to allow:

- to shape the envisaged analysis steps in a traceable manner,
- to facilitate readability of the user-specific "programming handwritings",
- to apply uniform standards to projects performed at different locations of the research data centres of the Federation and the Länder,
- and, where necessary, to change the project staff at the RDC without causing delays in the project progress.

The benefits obtained from those rules for the research data centres involve a considerable advantage for users, too: The checking time for the output, for which users undoubtedly are urgently waiting for, will substantially be reduced.

***Program heading***

Specifically, the following requirements have been included in the rules:

The program developed by the researcher should include a program heading, indicating the project title, the data material used, and contact information of the data user. The project title and the information on data material used allow rapid identification of the program at the research data centre. The contact information is important because in case of faulty programs the user can immediately be contacted by telephone or e-mail in order to agree on modifications.

***Program explanation***

Also, a program explanation should be provided at the beginning of the program, describing its purpose. Where applicable, information is also required on how the current analyses fit into the project's state of affairs and progress, on the reference to previous and future analyses, or possible changes in the analysis strategy. The explanation should also contain a list of variables of the original dataset used and of the new variables created by the program code, including its labels. Such information facilitates orientation when checking the results. Repeated analyses involving minor changes will more easily be recognised and cross-table confidentiality, which requires comparison with previous outputs,

[9] The main persons involved in developing the rules for program design as part of the ad-hoc working group on "confidentiality and documentation" of the research data centres of the Federation and the Länder are: Heike Habla, Jörg Höhne, Ricarda Nauenburg, Ramona Pohl, Dr. Heinz Stralla and Alexander Vogel. The chapter was written by Andrea Harausz and Ricarda Nauenburg.

can be performed more rapidly. The names and contents of newly formed variables are more easily recognised by the checking person if they have been mentioned before.

***Output tables***

The output tables should meet specific format requirements. Every output element should have a title and a serial number. Tables must be self-explanatory, i.e. the table stub and rows must clearly show the table content. Tables should be manageable, that is, they must not exceed a reasonable size. Tables showing totals must have a column for the underlying number of cases to allow checking the frequency and dominance rules. These requirements are based, among other things, on the experience made with confidentiality checks of poorly structured and oversized tables. Output numbering has proved necessary to allow referring to a number when communicating with the data user, rather than having to describe in detail the content of an analysis (often quite similar to other ones) to identify the table.

***Path reference***

Path references to files used should be put at the beginning of the program. This provides an overview of what files are loaded by the program and what files are stored. This allows the research data centre to adjust the paths (which is frequently necessary) more rapidly than when having to search the entire program for load and store commands.

***Structure***

The program code should be clearly structured visually; loops should be indented and some space should be left between program blocks. Capital and small letters should be used consistently, program sections should be numbered, and long programs should be subdivided into modules. This, together with the following three items, will allow reducing the time required for getting acquainted with different programming styles and for understanding the program goals and output content.

***Comments***

Comments should be given on the program. All program blocks and individual commands needing explanation or analyses should be provided with intelligible and sufficiently detailed comments to facilitate understanding of the program steps. Especially the macros must be documented in detail. Variables must be indicated with their labels in the comment.

***Variable names***

Descriptive names should be given to newly formed variables; intelligible variable labels and, where applicable, value labels should be added.

***Presentation***

Designations of elements used in the code should not be changed (example: relations [everywhere either “>” or “GT”], missing values [0 or $-x$], etc.).

***Logging***

The log function of the statistics software (creating log files) must be activated by the user's program code or in any case must not be deactivated, depending on the statistics software. Creating the logs allows archiving the project, so that possible disclosures can be detected later.

***Changing the code***

Where minor changes are applied to the program, only the relevant program elements should be recalculated. The usual procedure for one's own workplace is to perform complete runs of a program with minor changes again and again until a satisfactory result has been obtained. For controlled teleprocessing, however, this means that nearly identical outputs have to be checked again and again.

#### 6.6.2.3 Summary

Different from the program rules required by automated systems, these rules do not completely block certain procedures and commands. Manual handling allows taking individual decisions – when calling up critical procedures, and depending on the data or the specific analysis – to what extent the analysis results thus obtained can be transmitted to the user or must be retained for confidentiality reasons. In this manner, users can make optimal use of the information content of the data.

Supplementary to the catalogue of rules, a model program (cf. Section 6.6.2.4) has been designed and is available for the statistics packages offered (SPSS, STATA and SAS). In addition, a glossary is provided, containing the terms used in the jargon of official statistics, in programming languages and statistical packages, scientific notation and everyday language and translating them into each other.

The program criteria developed are sent in the form of a letter to the researchers and are part of the data usage contract. The purpose of the letter is to ask users for understanding of the list of rules, to avoid misunderstandings from the beginning, and to rapidly and efficiently make users acquainted with the rules of controlled teleprocessing, which may be quite different from the work methods they are used to at their own workplace. When developing the criteria, the model programs and the glossary, the aim was to make them generally acceptable, thus making it easier for the very different scientific users to meet the requirements.

From the viewpoint of the research data centres, the individual items are maximum requirements for developing perfect program codes. If the criteria are not entirely met in some cases, it is up to the discretion of the RDC staff member to decide to what extent it is possible for him to check the results. As the research projects are handled individually and personally, interpreting the rules to the users' benefit is rather liberal. If, however, a project is getting very difficult to follow, the research data centres may refer to the rules that must be adhered to.

#### 6.6.2.4 Programming example in SPSS

```
*****************************************************************************.
*** open log file.
set journal on.
set journal="path\log_name of program code.jnl".
set printback=on.

*****************************************************************************.
*****************************************************************************.
***                 project title:  Women and Work in Germany
***                                 data:   Micro Census 1985
***
***      name of program code:  name of program code.sps
***          date of creation:  date
***                    author:  name
***                    e-mail:  e-mail address
***                     phone:  phone number
***
***       name of output file:  name of program code.spo
***
***      statistical software:  SPSS version 13.0
***
***                           outline:   analysis of multiple person numbers, data check
***
***                        variables:
***                                    ef1:    Land
***                                    ef2:    administrative district
***                                    ef3:    sample district number
***                                    ef4:    household number
***                                    ef5:    person number
***                                    ef6:    family number
***                                    ef23:   age
***                                    ef26:   population group
***                                    ef27:   population in private households
***                                    ef28:   population at family residence
***                                    ef38:   marital status.
***
***             created variables:
***                                 persnr: person number
***                                  famnr: family number
***                               hhnr: household number
***                             nrdiff: consistency of household number
```

```
***                              and family number
***                          piddiff: test on unique person number
***
***            weighting variable:
***                                gew1: weighting variable from ef253
***
***
*********************************************************************.
*********************************************************************.


*(1)*** original data: mc95org.
FILE HANDLE mz85org /Name='path\mc_1985.sav'.

**** save new variables: mc85working.
FILE HANDLE mc85working /Name='path\mc_1985_working_1.sav'.


*********************************************************************.
*(2)*** read data.
GET FILE=mc85org.

**** select private households (ef27 eq 1) at main residence (ef26 lt 3) and
**** at family residence ef28 eq 1.
SELECT IF (ef27 EQ 1) & (ef26 LT 3) & (ef28 EQ 1).


*********************************************************************.
*(3)*** generate household and family number from ef1-ef6.
**** pers ef1-ef5
**** hh ef1-ef4
**** fam ef1-ef4 + ef6.
SORT CASES BY ef1 ef2 ef3 ef4 ef5.
COMPUTE persnr=(ef1*10000000000)+(ef2*1000000000)+(ef3*100000)+(ef4*1000)+(ef5*10).
EXE.

SORT CASES BY ef1 ef2 ef3 ef4 .
COMPUTE hhnr=(ef1*10000000000)+(ef2*1000000000)+(ef3*100000)+(ef4*1000).
EXE.

SORT CASES BY ef1 ef2 ef3 ef4 ef6.
COMPUTE famnr=(ef1*10000000000)+(ef2*1000000000)+(ef3*100000)+(ef4*1000)+(ef6*1).
EXE.
```

```
**** labeling.
VARIABLE LABELS persnr 'person number'.
VARIABLE LABELS hhnr 'household number'.
VARIABLE LABELS famnr 'family number'.


*****************************************************************************.
*(4)*** weighting.
**** multiply the weighting variable by 0.1.
COMPUTE gew=(ef253 * 0.1).
EXE.

WEIGHT BY gew .
*****************************************************************************.
*(5)***compare household number (hhnr)  - family number (famnr) - person number (persnr).
COMPUTE nrdiff=0.

**** if family number (famnr) unequal household number (hhnr) then consistency

of household number and family number (nrdiff) = 1.

DO IF famnr NE hhnr.
    COMPUTE nrdiff=1.
END IF.

**** labeling.
VARIABLE LABELS nrdiff 'consistency of household number and family number'.
VALUE LABELS nrdiff 1 'household number and family number inconsistent'
                                0 'household number and family number consistent.

TITLE "output no. 1: constistency of household number and family number".
FREQ VAR=nrdiff.


*****************************************************************************.
*(6)*** test on unique person number (persnr).
SORT CASES BY persnr.
COMPUTE piddiff=0.
EXE.

**** if person number (persnr) is equal to previous person number, then test on unique pe
IF persnr EQ LAG(persnr) piddiff=1.
```

```
EXE.

**** labeling.
VARIABLE LABELS piddiff 'test on unique person number'.
VALUE LABELS piddiff    1'multiple person number'
                                0'unique person number'.

TITLE "output no. 2: multiple person numbers".
FREQ VAR=piddiff.


*************************************************************************.
*(7)*** user defined table (counts)
**** mean and valid cases (N) of age (ef23) by marital status (ef38).

CTABLES
  /VLABELS VARIABLES=ef38 ef23 DISPLAY=DEFAULT
  /TABLE ef23 [MEAN, VALIDN F40.0] BY ef38
  /CATEGORIES VARIABLES=ef38 ORDER=A KEY=VALUE EMPTY=EXCLUDE
  /TITLES
   TITLE= 'output no. 3: cross table age and marital status'.


*************************************************************************.
*(8)*** save new variables.
SAVE OUTFILE=mc85working.
```

## 6.7 References

John Coder and Marc Cigrang (2003), *LISSY: A system for providing Restricted Access to Survey Microdata from Remote Sites*, Monographs in Official Statistics, Luxembourg

Anco Hundepool and Peter-Paul de Wolf(2005), *OnSite@Home: Remote Access at Statistics Netherlands*, Monographs of Official Statistics, Luxembourg

Lars-Johan Söderberg (2005), *MONA,- Microdata On liNe Access as Statistics Sweden*, Monographs of Official Statistics, Luxembourg

Lars Borchsenius (2005), *New developments in the Danish system for access to micro data*, Monographs of Official Statistics, Luxembourg

Dr. Sylvia Zühlke, Markus Zwick, Sebastian Scharnhorst and Thomas Wende (2005), *The research data centres of the Federal Statistical Office and the statistical offices of the Länder*, Forschungsdatenzentrum, Statistisches Bundesamt.

Brandt, M. et al. (2009). *Guidelines for the checking of output based on microdata research:* https://research.cbs.nl/casc/ESSnet/GuidelinesForOutputChecking_Dec2009.pdf

The DwB project, Work Package 11, extracted from the deliverable D11.8 (2015). *Guidelines for Output Checking:* https://cros.ec.europa.eu/system/files/2024-02/Output-checking-guidelines.pdf

# Glossary

In 2005 Mark Elliot (University of Manchester), Anco Hundepool (Statistics Netherlands), Eric Schulte Nordholt (Statistics Netherlands), Jean-Louis Tambay (Statistics Canada) and Thomas Wende (Destatis, Germany) took the initiative to compile a Glossary on Statistical Disclosure Control. A first version of the Glossary was presented at the UNECE worksession on SDC in Geneva in November 2005. This glossary can also be found via https://research.cbs.nl/casc/glossary.htm.

The underlined links in this glossary refer to other term in this glossary. This handbook contains also an index, but to avoid misleading cross references we have not indexed this glossary.

## A

**Analysis server:** A form of remote data laboratory designed to run analysis on data stored on a safe server. The user sees the results of their analysis but not the data.

**Anonymised data:** Data containing only anonymised records.

**Anonymised record:** A record from which direct identifiers have been removed.

**Approximate disclosure:** Approximate disclosure happens if a user is able to determine an estimate of a respondent value that is close to the real value. If the estimator is exactly the real value the disclosure is exact.

**Argus:** Two software packages for Statistical Disclosure Control are called Argus. $\mu$-ARGUS is a specialized software tool for the protection of microdata. The two main techniques used for this are global recoding and local suppression. In the case of global recoding several categories of a variable are collapsed into a single one. The effect of local suppression is that one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. Both global recoding and local suppression lead to a loss of information, because either less detailed information is provided or some information is not given at all. $\tau$-ARGUS is a specialized software tool for the protection of tabular data. $\tau$-ARGUS is used to produce safe tables. $\tau$-ARGUS uses the same two main techniques as $\mu$-ARGUS: global recoding and local suppression. For $\tau$-ARGUS the latter consists of suppression of cells in a table.

**Attribute disclosure:** Attribute disclosure is attribution independent of identification. This form of disclosure is of primary concern to NSIs involved in tabular data release and arises from the presence of empty cells either in a released table or linkable set of tables after any subtraction has taken place. Minimally, the presence of an empty cell within a table means that an intruder may infer from mere knowledge that a population unit is represented in the table and that the intruder does not possess the combination of attributes within the empty cell.

**Attribution:** Attribution is the association or disassociation of a particular attribute with a particular population unit.

## B

**Barnardisation:** A method of disclosure control for tables of counts that involves randomly adding or subtracting 1 from some cells in the table.

**Blurring:** Blurring replaces a reported value by an average. There are many possible ways to implement blurring. Groups of records for averaging may be formed by matching on other variables or by sorting on the variable of interest. The number of records in a group (whose data will be averaged) may be fixed or random. The average associated with a particular group may be assigned to all members of a group, or to the "middle" member (as in a moving average). It may be performed on more than one variable with different groupings for each variable.

**Bottom coding:** See top and bottom coding.

**Bounds:** The range of possible values of a cell in a table of frequency counts where the cell value has been perturbed or suppressed. Where only margins of tables are released it is possible to infer bounds for the unreleased joint distribution. One method for inferring the bounds across a table is known as the Shuttle algorithm.

## C

**Cell Key Method (CKM):** A post-tabular perturbative SDC method that adds noise to the original cell values. The tables are protected consistently, but they are no longer additive.

**Cell suppression:** In tabular data the cell suppression SDC method consists of primary and complementary (secondary) suppression. Primary suppression can be characterised as withholding the values of all risky cells from publication, which means that their value is not shown in the table but replaced by a symbol such as '×' to indicate the suppression. According to the definition of risky cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or presenting

a case of dominance have to be primary suppressed. To reach the desired protection for risky cells, it is necessary to suppress additional non- risky cells, which is called complementary (secondary) suppression. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the risky cells with the least amount of suppressed information.

**Complementary suppression:** Synonym of secondary suppression.

**Complete disclosure:** Synonym of exact disclosure.

**Concentration rule:** Synonym of (n,k) rule.

**Controlled rounding:** To solve the additivity problem, a procedure called controlled rounding was developed. It is a form of random rounding, but it is constrained to have the sum of the published entries in each row and column equal to the appropriate published marginal totals. Linear programming methods are used to identify a controlled rounding pattern for a table.

**Controlled Tabular Adjustment (CTA):** A method to protect tabular data based on the selective adjustment of cell values. Sensitive cell values are replaced by either of their closest safe values and small adjustments are made to other cells to restore the table additivity. Controlled tabular adjustment has been developed as an alternative to cell suppression.

**Conventional rounding:** A disclosure control method for tables of counts. When using conventional rounding, each count is rounded to the nearest multiple of a fixed base. For example, using a base of 5, counts ending in 1 or 2 are rounded down and replaced by counts ending in 0 and counts ending in 3 or 4 are rounded up and replaced by counts ending in 5. Counts ending between 6 and 9 are treated similarly. Counts with a last digit of 0 or 5 are kept unchanged. When rounding to base 10, a count ending in 5 may always be rounded up, or it may be rounded up or down based on a rounding convention.

## D

**Data intruder:** A data user who attempts to disclose information about a population unit through identification or attribution.

**Data intrusion detection:** The detection of a data intruder through their behaviour. This is most likely to occur through analysis of a pattern of requests submitted to a remote data laboratory. At present this is only a theoretical possibility, but it is likely to become more relevant as virtual safe settings become more prevalent.

**Data Intrusion Simulation (DIS):** A method of estimating the probability that a data intruder who has matched an arbitrary population unit against a sample unique in a target microdata file has done so correctly.

**Data protection:** Data protection refers to the set of privacy-motivated laws, policies and procedures that aim to minimise intrusion into respondents' privacy caused by the collection, storage and dissemination of personal data.

**Data swapping:** A disclosure control method for microdata that involves swapping the values of variables for records that match on a representative key. In the literature this technique is also sometimes referred to as "multidimensional transformation". It is a transformation technique that guarantees (under certain conditions) the maintenance of a set of statistics, such as means, variances and univariate distributions.

**Data utility:** A summary term describing the value of a given data release as an analytical resource. This comprises the data's analytical completeness and its analytical validity. Disclosure control methods usually have an adverse effect on data utility. Ideally, the goal of any disclosure control regime should be to maximise data utility whilst minimising disclosure risk. In practice disclosure control decisions are a trade-off between utility and disclosure risk.

**Deterministic rounding:** Synonym of conventional rounding.

**Direct identification:** Identification of a statistical unit from its formal identifiers.

**Disclosive cells:** Synonym of risky cells.

**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.

**Disclosure by fishing:** This is an attack method where an intruder identifies risky records within a target data set and then attempts to find population units corresponding to those records. It is the type of disclosure that can be assessed through a special uniques analysis.

**Disclosure by matching:** Disclosure by the linking of records within an identification dataset with those in an anonymised dataset.

**Disclosure by response knowledge:** This is disclosure resulting from the knowledge that a person was participating in a particular survey. If an intruder knows that a specific individual has participated in the survey, and that consequently his or her data are in the data set, identification and disclosure can be accomplished more easily.

**Disclosure by spontaneous recognition:** This means the recognition of an individual within the dataset. This may occur by accident or because a data intruder is searching for a particular individual. This is more likely to be successful if the individual has a rare combination of characteristics which is known to the intruder.

**Disclosure control methods:** There are two main approaches to control the disclosure of confidential data. The first is to reduce the information content of the data provided to the external user. For the release of tabular data this type of technique is called restriction based disclosure control method and for the release of microdata the expression

disclosure control by data reduction is used. The second is to change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased, but the information content is retained as much as possible. These are called perturbation based disclosure control methods.

**Disclosure from analytical outputs:** The use of output to make attributions about individual population units. This situation might arise to users that can interrogate data but do not have direct access to them such as in a remote data laboratory. One particular concern is the publication of residuals.

**Disclosure limitation methods:** Synonym of disclosure control methods.

**Disclosure risk:** A disclosure risk occurs if an unacceptably narrow estimation of a respondent's confidential information is possible or if exact disclosure is possible with a high level of confidence.

**Disclosure scenarios:** Depending on the intention of the intruder, his or her type of a priori knowledge and the microdata available, three different types of disclosure or disclosure scenarios are possible for microdata: disclosure by matching, disclosure by response knowledge and disclosure by spontaneous recognition.

**Dissemination:** Supply of data in any form whatever: publications, access to databases, microfiches, telephone communications, etc.

**Disturbing the data:** This process involves changing the data in some systematic fashion, with the result that the figures are insufficiently precise to disclose information about individual cases.

**Dominance rule:** Synonym of (n,k) rule.

## E

**Exact disclosure:** Exact disclosure occurs if a user is able to determine the exact attribute for an individual entity from released information.

## F

**Feasibility interval:** The interval containing possible values for a suppressed cell in a table, given the table structure and the values published.

**Formal identifier:** Any variable or set of variables which is structurally unique for every population unit, for example a population registration number. If the formal identifier is known to the intruder, identification of a target individual is directly possible for him or her, without the necessity to have additional knowledge before studying the microdata.

Some combinations of variables such as name and address are pragmatic formal identifiers, where non-unique instances are empirically possible, but with negligible probability.

## G

**Global recoding:** Problems of confidentiality can be tackled by changing the structure of data. Thus, rows or columns in tables can be combined into larger class intervals or new groupings of characteristics. This may be a simpler solution than the suppression of individual items, but it tends to reduce the descriptive and analytical value of the table. This protection technique may also be used to protect microdata.

## H

**HITAS:** A heuristic approach to cell suppression in hierarchical tables.

## I

**Identification:** Identification is the association of a particular record within a set of data with a particular population unit.

**Identification dataset:** A dataset that contains formal identifiers.

**Identification data:** Those personal data that allow direct identification of the data subject, and which are needed for the collection, checking and matching of the data, but are not subsequently used for drawing up statistical results.

**Identification key:** Synonym of key.

**Identification risk:** This risk is defined as the probability that an intruder identifies at least one respondent in the disseminated microdata. This identification may lead to the disclosure of (sensitive) information about the respondent. The risk of identification depends on the number and nature of quasi-identifiers in the microdata and in the a priori knowledge of the intruder.

**Identifying variable:** A variable that either is a formal identifier or forms part of a formal identifier.

**Indirect identification:** Inferring the identity of a population unit within a microdata release other than from direct identification.

**Inferential disclosure:** Inferential disclosure occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of home. As the purchase

price of a home is typically public information, a third party might use this information to infer the income of a data subject. In general, NSIs are not concerned with inferential disclosure for two reasons. First, a major purpose of statistical data is to enable users to infer and understand relationships between variables. If NSIs equated disclosure with inference, no data could be released. Second, inferences are designed to predict aggregate behaviour, not individual attributes, and thus often poor predictors of individual data values.

**Informed consent:** Basic ethical tenet of scientific research on human populations. Sociologists do not involve a human being as a subject in research without the informed consent of the subject or the subject's legally authorized representative, except as otherwise specified. Informed consent refers to a person's agreement to allow personal data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including awareness of any risks involved, of uses and users of the data, and of alternatives to providing the data.

**Intruder:** Synonym of data intruder.

## J

## K

**Key:** A set of key variables.

**Key variable:** A variable in common between two datasets, which may therefore be used for linking records between them. A key variable can either be a formal identifier or a quasi-identifier.

## L

**Licensing agreement:** A permit, issued under certain conditions, for researchers to use confidential data for specific purposes and for specific periods of time. This agreement consists of contractual and ethical obligations, as well as penalties for improper disclosure or use of identifiable information. These penalties can vary from withdrawal of the license and denial of access to additional data sets to the forfeiting of a deposit paid prior to the release of a microdata file. A licensing agreement is almost always combined with the signing of a contract. This contract includes a number of requirements: specification of the intended use of the data; instruction not to release the microdata file to another recipient; prior review and approval by the releasing agency for all user outputs to be published or disseminated; terms and location of access and enforceable penalties.

**Local recoding:** A disclosure control technique for microdata where two (or more) different versions of a variable are used dependent on some other variable. The different versions will have different levels of coding. This will depend on the distribution of the first variable conditional on the second. A typical example occurs where the distribution of a variable is heavily skewed in some geographical areas. In the areas where the distribution is skewed minor categories may be combined to produce a courser variable.

**Local suppression:** Protection technique that diminishes the risk of recognition of information about individuals or enterprises by suppressing individual scores on identifying variables.

**Lower bound:** The lowest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.

## M

**Macrodata:** Synonym of tabular data.

**Microaggregation:** Records are grouped based on a proximity measure of variables of interest, and the same small groups of records are used in calculating aggregates for those variables. The aggregates are released instead of the individual record values.

**Microdata:** A microdata set consists of a set of records containing information on individual respondents or on economic entities.

**Minimal unique:** A combination of variable values that are unique in the microdata set at hand and contain no proper subset with this property (so it is a minimal set with the uniqueness property).

## N

**NSI(s):** Abbreviation for National Statistical Institute(s).

**(n,k) rule:** A cell is regarded as confidential, if the n largest units contribute more than k % to the cell total, e.g. n=2 and k=85 means that a cell is defined as risky if the two largest units contribute more than 85% to the cell total. The n and k are given by the statistical authority. In some NSIs the values of n and k are confidential.

# O

**On-site facility:** A facility that has been established on the premises of several NSIs. It is a place where external researchers can be permitted access to potentially disclosive data under contractual agreements which cover the maintenance of confidentiality, and which place strict controls on the uses to which the data can be put. The on-site facility can be seen as a 'safe setting' in which confidential data can be analysed. The on-site facility itself would consist of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. Both the physical and the IT aspects of security would be considered here. The on-site facility also includes administrative and support facilities to external users, and ensures that the agreed conditions for access to the data were complied with.

**Ordinary rounding:** Synonym of conventional rounding.

**Oversuppression:** A situation that may occur during the application of the technique of cell suppression. This denotes the fact that more information has been suppressed than strictly necessary to maintain confidentiality.

# P

**Partial disclosure:** Synonym of approximate disclosure.

**Passive confidentiality:** For foreign trade statistics, EU countries generally apply the principle of "passive confidentiality", that is they take appropriate measures only at the request of importers or exporters who feel that their interests would be harmed by the dissemination of data.

**Personal data:** Any information relating to an identified or identifiable natural person ('data subject'). An identifiable person is one who can be identified, directly or indirectly. Where an individual is not identifiable, data are said to be anonymous.

**Perturbation based disclosure control methods:** Techniques for the release of data that change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased but the information content is retained as far as possible. Perturbation based methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. For example, an error can be inserted in the cell values after a table is created, which means that the error is introduced to the output of the data and will therefore be referred to as output perturbation. The error can also be inserted in the original data on the microdata level, which is the input of the tables one wants to create; the method will then be referred to as data perturbation - input perturbation being the better but uncommonly used expression. Possible perturbation methods are:
- rounding;

- perturbation, for example, by the addition of random noise or by the Post Randomisation Method;
- disclosure control methods for microdata applied to tabular data.

**Population unique:** A record within a dataset which is unique within the population on a given key.

**P-percent rule:** A (p,q) rule where q is 100%, meaning that from general knowledge any respondent can estimate the contribution of another respondent to within 100% (i.e., knows the value to be nonnegative and less than a certain value which can be up to twice the actual value).

**(p,q) rule:** It is assumed that out of publicly available information the contribution of one individual to the cell total can be estimated to within q per cent (q=error before publication); after the publication of the statistic the value can be estimated to within p percent (p=error after publication). In the (p,q) rule the ratio p/q represents the information gain through publication. If the information gain is unacceptable the cell is declared as confidential. The parameter values p and q are determined by the statistical authority and thus define the acceptable level of information gain. In some NSIs the values of p and q are confidential.

**Post Randomisation Method (PRAM):** Protection method for microdata in which the scores of a categorical variable are changed with certain probabilities into other scores. It is thus intentional misclassification with known misclassification probabilities.

**Primary confidentiality:** It concerns tabular cell data, whose dissemination would permit attribute disclosure. The two main reasons for declaring data to be primary confidential are:
- too few units in a cell;
- dominance of one or two units in a cell.

The limits of what constitutes "too few" or "dominance" vary among statistical domains.

**Primary protection:** Protection using disclosure control methods for all cells containing small counts or cases of dominance.

**Primary suppression:** This technique can be characterized as withholding all disclosive cells from publication, which means that their value is not shown in the table, but replaced by a symbol such as '×' to indicate the suppression. According to the definition of disclosive cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or representing cases of dominance have to be primary suppressed.

**Prior-posterior rule:** Synonym of the (p,q) rule.

**Privacy:** Privacy is a concept that applies to data subjects while confidentiality applies to data. The concept is defined as follows: "It is the status accorded to data which has been agreed upon between the person or organisation furnishing the data and the organisation receiving it and which describes the degree of protection which will be provided." There

is a definite relationship between confidentiality and privacy. Breach of confidentiality can result in disclosure of data which harms the individual. This is an attack on privacy because it is an intrusion into a person's self-determination on the way his or her personal data are used. Informational privacy encompasses an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviours, opinions and attitudes will be shared with or withheld from others.

**Probability based disclosures (approximate or exact):** Sometimes although a fact is not disclosed with certainty, the published data can be used to make a statement that has a high probability of being correct.

## Q

**Quasi-identifier:** Variable values or combinations of variable values within a dataset that are not structural uniques but might be empirically unique and therefore in principle uniquely identify a population unit.

## R

**Random perturbation:** This is a disclosure control method according to which a noise, in the form of a random value is added to the true value or, in the case of categorical variables, where another value is randomly substituted for the true value.

**Random rounding:** In order to reduce the amount of data loss that occurs with suppression, alternative methods have been investigated to protect sensitive cells in tables of frequencies. Perturbation methods such as random rounding and controlled rounding are examples of such alternatives. In random rounding cell values are rounded, but instead of using standard rounding conventions a random decision is made as to whether they will be rounded up or down. The rounding mechanism can be set up to produce unbiased rounded results.

**Rank swapping:** Rank swapping provides a way of using continuous variables to define pairs of records for swapping. Instead of insisting that variables match (agree exactly), they are defined to be close based on their proximity to each other on a list sorted on the continuous variable. Records which are close in rank on the sorted variable are designated as pairs for swapping. Frequently in rank swapping the variable used in the sort is the one that will be swapped.

**Record linkage process:** Process attempting to classify pairs of matches in a product space A×B from two files A and B into M, the set of true links, and U, the set of non-true links.

**Record swapping:** A special case of data swapping, where the geographical codes of records are swapped.

**Remote access:** On-line access to protected microdata.

**Remote data laboratory:** A virtual environment providing remote execution facilities.

**Remote execution:** Submitting scripts on-line for execution on disclosive microdata stored within an institute's protected network. If the results are regarded as safe data, they are sent to the submitter of the script. Otherwise, the submitter is informed that the request cannot be acquiesced. Remote execution may either work through submitting scripts for a particular statistical package such as SAS, SPSS or STATA which runs on the remote server or via a tailor made client system which sits on the user's desk top.

**Residual disclosure:** Disclosure that occurs by combining released information with previously released or publicly available information. For example, tables for nonoverlapping areas can be subtracted from a larger region, leaving confidential residual information for small areas.

**Restricted access:** Imposing conditions on access to the microdata. Users can either have access to the whole range of raw protected data and process individually the information they are interested in - which is the ideal situation for them - or their access to the protected data is restricted and they can only have a certain number of outputs (e.g. tables) or maybe only outputs of a certain structure. Restricted access is sometimes necessary to ensure that linkage between tables cannot happen.

**Restriction based disclosure control method:** Method for the release of tabular data, which consists in reducing access to the data provided to the external user. This method reduces the content of information provided to the user of the tabular data. This is implemented by not publishing all the figures derived from the collected data or by not publishing the information in as detailed a form as would be possible.

**Risky cells:** The cells of a table which are non-publishable due to the risk of statistical disclosure are referred to as risky cells. By definition there are three types of risky cells: small counts, dominance and complementary suppression cells.

**Risky data:** Data are considered to be disclosive when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit.

**Rounding:** Rounding belongs to the group of disclosure control methods based on output-perturbation. It is used to protect small counts in tabular data against disclosure. The basic idea behind this disclosure control method is to round each count up or down either deterministically or probabilistically to the nearest integer multiple of a rounding base. The additive nature of the table is generally destroyed by this process. Rounding can also serve as a recoding method for microdata.

**R-U confidentiality map:** A graphical representation of the trade off between disclosure risk and data utility.

# S

**Safe data:** Microdata or macrodata that have been protected by suitable Statistical Disclosure Control methods.

**Safe setting:** An environment such as a microdata lab whereby access to a disclosive dataset can be controlled.

**Safety interval:** The minimal feasibility interval that is required for the value of a cell that does not satisfy the primary suppression rule.

**Sample unique:** A record within a dataset which is unique within that dataset on a given key.

**Sampling:** In the context of disclosure control, this refers to releasing only a proportion of the original data records on a microdata file.

**Sampling fraction:** The proportion of the population contained within a data release. With simple random sampling, the sample fraction represents the proportion of population units that are selected in the sample. With more complex sampling methods, this is usually the ratio of the number of units in the sample to the number of units in the population from which the sample is selected.

**Scenario analysis:** A set of pseudo-criminological methods for analysing and classifying the plausible risk channels for a data intrusion. The methods are based around first delineating the means, motives and opportunity that an intruder may have for conducting the attack. The output of such an analysis is a specification of a set of keys likely to be held by data intruders.

**Secondary data intrusion:** After an attempt to match between identification and target datasets an intruder may discriminate between non-unique matches by further direct investigations using additional variables.

**Secondary disclosure risk:** It concerns data which is not primary disclosive, but whose dissemination, when combined with other data permits the identification of a microdata unit or the disclosure of a unit's attribute.

**Secondary suppression:** To reach the desired protection for risky cells, it is necessary to suppress additional non-risky cells, which is called secondary suppression or complementary suppression. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the disclosive cells at the highest level of information contained in the released statistics.

**Security:** An efficient disclosure control method provides protection against exact disclosure or unwanted narrow estimation of the attributes of an individual entity, in other words, a useful technique prevents exact or partial disclosure. The security level is accordingly high. In the case of disclosure control methods for the release of microdata this protection is ensured if the identification of a respondent is not possible, because the identification is the prerequisite for disclosure.

**Sensitive cell:** Cell for which knowledge of the value would permit an unduly accurate estimate of the contribution of an individual respondent. Sensitive cells are identified by the application of a dominance rule such as the (n,k) rule or the (p,q) rule to their microdata.

**Sensitive variables:** Variables contained in a data record apart from the key variables, that belong to the private domain of respondents who would not like them to be disclosed. There is no exact definition given for what a 'sensitive variable' is and therefore, the division into key and sensitive variables is somehow arbitrary. Some data are clearly sensitive such as the possession of a criminal record, one's medical condition or credit record, but there are other cases where the distinction depends on the circumstances, e.g. the income of a person might be regarded as a sensitive variable in some countries and as quasi-identifier in others, or in some societies the religion of an individual might count as a key and a sensitive variable at the same time. All variables that contain one or more sensitive categories, i.e. categories that contain sensitive information about an individual or enterprise, are called sensitive variables.

**Shuttle algorithm:** A method for finding lower and upper cell bounds by iterating through dependencies between cell counts. There exist many dependencies between individual counts and aggregations of counts in contingency tables. Where not all individual counts are known, but some aggregated counts are known, the dependencies can be used to make inferences about the missing counts. The Shuttle algorithm constructs a specific subset of the many possible dependencies and recursively iterates through them in order to find bounds on missing counts. As many dependencies will involve unknown counts, the dependencies need to be expressed in terms of inequalities involving lower and upper bounds, rather than simple equalities. The algorithm ends when a complete iteration fails to tighten the bounds on any cell counts.

**Special uniques analysis:** A method of analysing the per-record risk of microdata.

**Statistical confidentiality:** The protection of data that relate to single statistical units and are obtained directly for statistical purposes or indirectly from administrative or other sources against any breach of the right to confidentiality. It implies the prevention of unlawful disclosure.

**Statistical Data Protection (SDP):** Statistical Data Protection is a more general concept which takes into account all steps of production. SDP is multidisciplinary and draws on computer science (data security), statistics and operations research.

**Statistical disclosure:** Statistical disclosure is said to take place if the dissemination of a statistic enables the external user of the data to obtain a better estimate for a confidential piece of information than would be possible without it.

**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.

**Statistical Disclosure Limitation (SDL):** Synonym of Statistical Disclosure Control.

**Subadditivity:** One of the properties of the (n,k) rule or (p,q) rule that assists in the search for complementary cells. The property means that the sensitivity of a union of disjoint cells cannot be greater than the sum of the cells' individual sensitivities (triangle inequality). Subadditivity is an important property because it means that aggregates of cells that are not sensitive are not sensitive either and do not need to be tested.

**Subtraction:** The principle whereby an intruder may attack a table of population counts by removing known individuals from the table. If this leads to the presence of certain zeroes in the table then that table is vulnerable to attribute disclosure.

**Suppression:** One of the most commonly used ways of protecting sensitive cells in a table is via suppression. It is obvious that in a row or column with a suppressed sensitive cell, at least one additional cell must be suppressed, or the value in the sensitive cell could be calculated exactly by subtraction from the marginal total. For this reason, certain other cells must also be suppressed. These are referred to as secondary suppressions. While it is possible to select cells for secondary suppression manually, it is difficult to guarantee that the result provides adequate protection.

**SUDA:** A software system for conducting analyses on population uniques and special sample uniques. The special uniques analysis method implemented in SUDA for measuring and assessing disclosure risk is based on resampling methods and used by the ONS.

**Swapping (or switching):** Swapping (or switching) involves selecting a sample of the records, finding a match in the data base on a set of predetermined variables and swapping all or some of the other variables between the matched records.

**Synthetic data:** An approach to confidentiality where instead of disseminating real data, synthetic data that have been generated from one or more population models are released.

# T

**Table server:** A form of remote data laboratory designed to release safe tables.

**Tables of frequency (count) data:** These tables present the number of units of analysis in a cell. When data are from a sample, the cells may contain weighted counts, where weights are used to bring sample results to the population levels. Frequencies may also be represented as percentages.

**Tables of magnitude data:** Tables of magnitude data present the aggregate of a "quantity of interest" over all units of analysis in the cell. When data are from a sample, the cells may contain weighted aggregates, where quantities are multiplied by units' weights to bring sample results up to population levels. The data may be presented as averages by dividing the aggregates by the number of units in their cells.

**Tabular data:** Aggregate information on entities presented in tables.

**Target dataset:** An anonymised dataset in which an intruder attempts to identify particular population units.

**Targeted Record Swapping (TRS):** A pre-tabular perturbative SDC method that applies a swapping procedure to the microdata before generating a table. The tables are additive and protected consistently.

**Threshold (rule):** Usually, with the threshold rule, a cell in a table of frequencies is defined to be sensitive if the number of respondents is less than some specified number. Some agencies require at least five respondents in a cell, others require three. When thresholds are not respected, an agency may restructure tables and combine categories or use cell suppression or rounding, or provide other additional protection in order to satisfy the rule.

**Top and bottom coding:** It consists in setting top-codes or bottom-codes on quantitative variables. A top-code for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is replaced by the upper limit or is not published on the microdata file at all. Similarly, a bottom-code is a lower limit on all published values for a variable. Different limits may be used for different quantitative variables, or for different subpopulations.

**Top coding:** See top and bottom coding.

## U

**Union unique:** A sample unique that is also population unique. The proportion of sample uniques that are union uniques is one measure of file level disclosure risk.

**Uniqueness:** The term is used to characterise the situation where an individual can be distinguished from all other members in a population or sample in terms of information available on microdata records (or within a given key). The existence of uniqueness is determined by the size of the population or sample and the degree to which it is segmented

by geographic information and the number and detail of characteristics provided for each unit in the dataset (or within the key).

**Upper bound:** The highest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.

## V

**Virtual safe setting:** Synonym of remote data laboratory.

## W

**Waiver approach:** Instead of suppressing tabular data, some agencies ask respondents for permission to publish cells even though doing so may cause these respondents' sensitive information to be estimated accurately. This is referred to as the waiver approach. Waivers are signed records of the respondents' granting permission to publish such cells. This method is most useful with small surveys or sets of tables involving only a few cases of dominance, where only a few waivers are needed. Of course, respondents must believe that their data are not particularly sensitive before they will sign waivers.

## X

## Y

## Z

# Index

cell suppression, 116–118, 121, 127, 130, 134, 135, 138, 139, 141, 142, 145, 147, 149, 151, 153, 155–157, 159, 160, 166, 167, 170, 174, 180, 181, 190, 197

disclosure risk, 4, 5, 8, 9, 18–20, 23–25, 27–33, 35, 39–43, 45, 52, 54, 55, 59, 70, 73, 87, 89, 91, 96, 100, 104, 115, 116, 118, 119, 121, 126, 127, 130, 133, 140–144, 146, 152, 163, 164, 166, 167, 171–174, 177, 178, 181, 183, 195, 197, 198

information loss, 27

optimal approach, 154

PRAM, 54, 58, 73–77, 93, 106, 179

quasi-identifier, 80

rank swapping, 46, 48, 52, 53, 86, 88, 103
record linkage, 29, 30, 41–46

# Index

cell suppression, 116–118, 121, 127, 130, 134, 135, 138, 139, 141, 142, 145, 147, 149, 151, 153, 155–157, 159, 160, 166, 167, 170, 174, 180, 181, 190, 197

disclosure risk, 4, 5, 8, 9, 18–20, 23–25, 27–33, 35, 39–43, 45, 52, 54, 55, 59, 70, 73, 87, 89, 91, 96, 100, 104, 115, 116, 118, 119, 121, 126, 127, 130, 133, 140–144, 146, 152, 163, 164, 166, 167, 171–174, 177, 178, 181, 183, 195, 197, 198

information loss, 27

optimal approach, 154

PRAM, 54, 58, 73–77, 93, 106, 179

quasi-identifier, 80

rank swapping, 46, 48, 52, 53, 86, 88, 103
record linkage, 29, 30, 41–46