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 SDC 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. SDC should be combined with other tools (administrative, legal and 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 mixtures of tools.

Statistical Disclosure Control, First Edition. Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer and Peter-Paul de Wolf.

- *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 (Trewin 2007). Microdata protection is the youngest sub-discipline and has experienced continuous evolution in the last years. If microdata are freely disseminated then SDL 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 2.4.2) 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 NSIs. 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

1.1.1 Disclosure

A disclosure occurs when a person or an 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: (1) identity disclosure and (2) attribute disclosure. Identity disclosure occurs with the association of a respondents' identity with a disseminated data record containing confidential information (see 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 (see 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.

1.1.2 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.

1.1.3 Tabular data

There are two types of tabular output:

1. *Magnitude tables*. 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.
2. *Frequency tables*. 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.

1.1.4 Microdata

A microdata set $\mathbf{V}$ can be viewed as a file with n records, where each record contains m *variables* (also called *attributes*) on an individual respondent, who can be a person or an organisation (e.g. a company). Microdata are the form from which all other data outputs are derived and they are the primary form that data are stored in. While in the past, NSIs simply derived outputs of other forms, more and more, microdata are becoming a key output by themselves.

Depending on their sensitivity, the variables in an original unprotected microdata set can be classified into four categories which are not necessarily disjoint:

1. *Identifiers*. These are variables that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc. Since

the objective of SDC is to prevent confidential information from being linked to specific respondents, SDC normally assumes that identifiers in **V** have been removed/encrypted in a pre-processing step.

2. *Quasi-identifiers or key variables*. Borrowing the definition from Dalenius (1986), Samarati (2001), a quasi-identifier is a set of variables in **V** that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in **V** refer. Unlike identifiers, quasi-identifiers cannot be removed from **V**. The reason is that any variable in **V** potentially belongs to a quasi-identifier (depending on the external data sources available to the user of **V**). Thus, one would need to remove all variables (!) to make sure that the data set no longer contains quasi-identifiers.
3. *Confidential outcome variables*. These are variables which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
4. *Non-confidential outcome variables*. Those variables which contain non-sensitive information on the respondent. Examples are town and country of residence, etc. Note that variables of this kind cannot be neglected when protecting a data set, because they can be part of a quasi-identifier. For instance, if 'Job' and 'Town of residence' can be considered non-confidential outcome variables, but their combination can be a quasi-identifier, because everyone knows who is the doctor in a small village.

Depending on their data type, the variables in a microdata set can be classified as:

- *Continuous*. A variable is considered continuous if it is numerical and arithmetical operations can be performed on it. Examples are income and age.
- *Categorical*. A variable is considered categorical when it takes values over a finite set and standard arithmetical operations do not make sense. Two main types of categorical variables can be distinguished:
  - *Ordinal*. An ordinal variable takes values in an ordered range of categories. Thus, the $\leq$, max and min operators are meaningful with ordinal data. The instruction level and the political preferences (left-right) are examples of ordinal variables.
  - *Nominal*. A nominal variable takes values in an unordered range of categories. The only possible operator is comparison for equality. The eye color and the address of an individual are examples of nominal variables.

1.1.5 Risk and utility

SDC seeks to optimise the trade-off between the disclosure risk and the utility of the protected released data. Yet, SDC is a discipline which was born from daily statistical practice and any theory trying to bestow a unified scientific standing to it should be flexible enough to deal with the risk-utility trade-off in a rather vast number of situations (different data structures, different contexts of previously released information

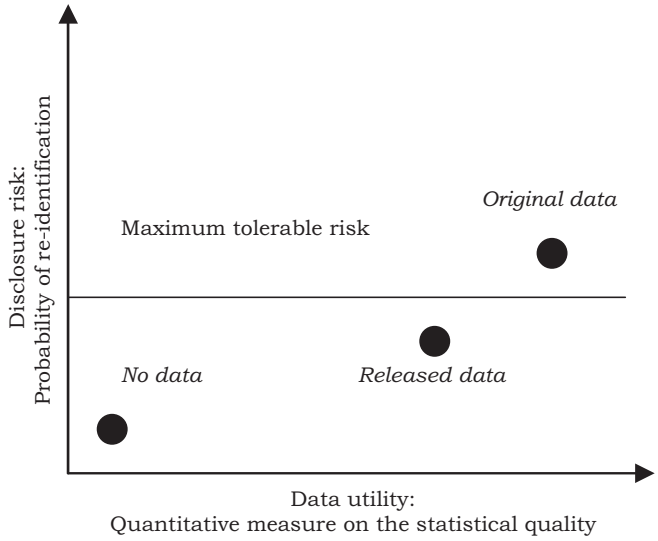

Figure 1.1 R-U confidentiality map.

and intruders' side knowledge, etc.). How to accurately measure risk and utility, how to distinguish legitimate users from intruders and how far to implement transparency are some of the open challenges SDC faces today, most of which are insightfully highlighted in Cox *et al*. (2011) and discussed in Domingo-Ferrer (2011).

We next review briefly some proposed models to deal with the risk-utility tension. One of them is merely descriptive (R-U maps depicting risk and utility), while the other two (*k*-anonymity and differential privacy) adopt a minimax approach: subject to a minimum guaranteed privacy, utility is to be maximised.

R-U maps

NSIs should aim to determine optimal SDC methods and solutions that minimise disclosure risk while maximising 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 then to find the balance in maintaining the utility of the data but reducing the risk below the maximum tolerable risk threshold.

k-anonymity

k-anonymity is a concept that was proposed by Samarati (2001); Samarati and Sweeney (1998); Sweeney (2002a, 2002b) as a different approach to face the conflict between information loss and disclosure risk.

A data set is said to satisfy k-anonymity for $k > 1$ if, for each combination of values of quasi-identifiers (e.g. name, address, age, gender, etc.), at least k records exist in the data set sharing that combination.

Note that, if a protected data set $\mathbf{V}'$ satisfies k-anonymity, an intruder trying to link $\mathbf{V}'$ with an external non-anonymous data source will find at least k records in $\mathbf{V}'$ that match any value of the quasi-identifier the intruder uses for record linkage. Thus re-identification, i.e. mapping a record in $\mathbf{V}'$ to a non-anonymous record in the external data source, is not possible; the best the intruder can hope for is to map groups of k records in $\mathbf{V}'$ to each non-anonymous external record.

If, for a given k, k-anonymity is assumed to be enough protection, one can concentrate on minimising information loss with the only constraint that k-anonymity should be satisfied. This is a clean way of solving the tension between data protection and data utility.

In Samarati (2001) and Sweeney (2002b) the approach suggested to reach k-anonymity is to combine generalisation and local suppression; in Domingo-Ferrer and Torra (2005), the use of microaggregation was proposed as an alternative; see Section 3.6 for a description of those methods.

k-anonymity has been criticised as being necessary but sometimes not sufficient to guarantee privacy. For example, if a group of k records sharing a quasi-identifier combination also shares the same value for a confidential variable (e.g. AIDS='YES'), it suffices for an intruder to know that John Smith's record is one of k records in the group to learn that John Smith suffers from AIDS. To remedy this, other privacy properties evolving from k-anonymity have been proposed; see Domingo-Ferrer and Torra (2008) for a survey of critiques and evolved models.

Differential privacy

In general, SDC methods take the data to be published as input and modify them with the aim of reducing the risk of disclosure by removing/changing combinations of variables likely to be re-identifiable. Differential privacy (Dwork 2006, 2011) tackles privacy preservation from a different perspective. On one side, differential privacy is not based on the precise understanding that certain combinations of variables might be re-identifiable. Instead, it seeks to guarantee, in a probabilistic sense, that after the addition of a new record to a database any information extracted from the database will remain close to what it was before. On the other side, differential privacy does not focus on a specific database. It seeks to protect all the possible databases that may arise from record addition.

The following formal definition of differential privacy can be found in Dwork (2006). A randomised function κ gives ε-differential privacy if, for all data sets D and D' differing in at most one row, and all $S \subset Range(\kappa)$

$$P[\kappa(D) \in S] \leq e^{\varepsilon} P[\kappa(D') \in S]. \tag{1.1}$$

Similarly to what happened with k-anonymity, the idea is that, among the randomisations offering differential privacy for a certain parameter value ε, one should choose the one causing minimal information loss.

Computationally, differential privacy is achieved by output perturbation; the responses are computed on the real data and the result is perturbed before release. In

most of the literature on differential privacy, the noise added for perturbation is taken to follow a Laplace distribution. However, it has been shown in Soria-Comas and Domingo-Ferrer (2011) that better noise distributions exist, in the sense that, given a parameter ε, they guarantee ε-differential privacy and have a smaller variance than the Laplace distribution (which implies less distortion for the data and hence less information loss).

However elegant, differential privacy has been criticised as not caring about data utility. In particular, if the original variables have bounded ranges, ε-differentially private data are likely to go off-range if ε is small (high protection); see Sarathy and Muralidhar (2011) and Soria-Comas and Domingo-Ferrer (2011) for details.

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:

1. Why is confidentiality protection needed?
2. What are the key characteristics and uses of the data?
3. What disclosure risks need to be protected against?
4. Disclosure control methods.
5. Implementation.

1.2.1 Why is confidentiality protection needed?

There are three main reasons why confidentiality protection is needed for statistical outputs:

1. 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.
2. 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.

3. 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.

1.2.2 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; and 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.

1.2.3 What disclosure risks need to be protected against?

Disclosure risk assessment 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 3.3.1. Risk-assessment methods for microdata are covered in Section 3.4 and different rules applied to assess the risk of magnitude and frequency tables are described in Chapters 4 and 5, respectively.

1.2.4 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 SDC 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. SDC methods used to reduce the risk of microdata, magnitude tables and frequency tables are covered in Chapters 3–5, respectively. Chapter 6 provides information on how disclosure risk can be managed by restricting access.

1.2.5 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 data set has been assessed for disclosure risk, and whether methods of protection have been applied. For quality purposes, users of a data set 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 SDC 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 chapter, 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 SDC has been included.