Andrzej Młodak ▪ Michał Pietrzak
Tomasz Klimanek ▪ Tomasz Józefowski ▪ Paweł Lańduch

Confidentiality vs. utility of statistical information. Dilemmas of statistical disclosure control

Summary*

https://doi.org/10.18559/978-83-8211-168-2-summary

Introduction

Comprehensive and precise statistical information is essential in order to plan and carry out development activities in different spheres of the socio-economic reality, to monitor their effects, conduct new studies and improve research tools. These needs are what drives the growing demand for statistical data regarding various aspects of the surrounding reality and enabling data users to analyse underlying phenomena of interest.

Data collected in statistical surveys or maintained in administrative registers contain a wide range of useful information about various characteristics of units, including their direct identifiers. These characteristics are protected by law and are subject to statistical confidentiality. The protection of personal data has become socially relevant since the introduction of Regulation (EE) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). The GDPR has prompted the adoption of appropriate national laws. This is why, before statistical outputs can be released, data holders are obliged to analyse them in order to minimise the risk of disclosing sensitive information that could be used by end users to re-identify respondents. These procedures are

* More in the monograph in Polish: Młodak, A., Pietrzak, M., Klimanek, T., Józefowski, T. and Lańduch, P. (2023). *Poufność a użyteczność informacji statystycznych. Dylematy ochrony udostępnianych danych*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. https://doi.org/10.18559/978-83-8211-168-2

known as statistical disclosure control (SDC). The first and simplest step undertaken during SDC is to remove key identifiers (such as a person's name, surname and personal identification number or a company registration number and its tax registration number).

Nowadays, however, such simple solutions, which have been used for many years, are far from sufficient. As the number of variables describing statistical units increases, the resulting number of possible combinations of categories of nominal or ordinal variables grows very quickly. Therefore, there is a considerable likelihood of the existence of unique combinations, which, in extreme cases, may contain single units that can be re-identified. This risk is particularly likely to occur in microdata (i.e., anonymised unit-level data) or multi-dimensional arrays of data, known as OLAP cubes; however, these are precisely the types of data which are increasingly sought after, especially by scientists and researchers. The risk of disclosure or re-identification can be ever greater if a survey or an administrative register contains information representing continuous variables measured on an interval or ratio scale. In such cases, it is possible to calculate how much particular groups contribute to total values (e.g., the share of employees in the public sector in the total number of employees). One should also consider the possibility that a potential end user might have access to other datasets, which could help them to identify individual units. Moreover, under current regulations, contact information of businesses (including sole proprietors) is publicly accessible, which means that their sensitive data need to be protected in a different way. Consequently, a number of advanced SDC methods have been developed (noise addition, post-randomisation, controlled tabular adjustment, etc.), which are based on mathematical statistics and IT solutions and have become an important methodological tool for official statistics.

SDC methods – description and discussion

Specific methods of statistical disclosure control have been developed for three main types of data:

- microdata,
- tabular data,
- output checking.

The monograph focuses on each of these three applications by describing specific characteristics of each type of data, their sources and fundamental principles of their protection, theoretical and IT tools used for this purpose as well as organisational and technological aspects of releasing statistical data, which are associated with the risk of unit re-identification or disclosure of sensitive information. In addition, it also describes ways of protecting confidentiality that can be applied to outputs of statistical analyses, such as descriptive statistics, estimates

of econometric models or charts. The authors have attempted to present the complexity of SDC and various rules of estimating re-identification risk. A comprehensive assessment of the effects of applying SDC should include an estimate of disclosure risk and the expected information loss due to the suppression or perturbation of sensitive information. The main purpose of SDC is to achieve an optimal trade-off between these two minimisation goals.

An optimal level of confidentiality regarding data collected by NSIs as well as those maintained by other holders can be ensured by employing appropriate SDC methods and effective measures of disclosure risk and information loss resulting from the application of SDC procedures. The choice of methods and their parameters depends on many different factors, such as laws and regulations, type and scope of data to be released in a given survey, organisation of data release or IT tools used for this purpose.

The monograph is an attempt to provide a comprehensive description of all of these aspects, including the goal and definition of statistical disclosure control, its formal and legal principles, particularly current regulations and international recommendations regarding data confidentiality, types of released data (including metadata, i.e., information about definitions of concepts that characterise units, reference periods of data collection and units for which data are collected, measurement methods, possible exceptions to rules of determining variables or missing data, explanations of known causes of deviations or gaps; paradata, i.e., information about the process by which the data were collected, which can be used to explain interpretations of the resulting characteristics of survey units as well as data not included in the survey itself but automatically collected during survey administration, such as the number of times a given respondent has visited the webpage with the survey form, the number of data changes made, the time taken to complete the form, etc. as well as other data). We explain the disclosure process, types of data users from the perspective of SDC, typologies of statistical outputs with respect to the protection of sensitive information and the trade-off between disclosure risk and data utility. One section of the monograph contains an overview of the most important SDC solutions and regulations implemented in different European countries and another one provides a description of the main kinds of microdata and the role of metadata and paradata in the SDC process.

The first of the two main factors that determine the effectiveness of the SDC process is the extent to which the risk of disclosure is minimised, i.e., the probability of identifying a particular statistical unit (its identity) or obtaining new, previously unknown information about it (its attributes) from data released by a national statistical institute or another data holder. Disclosure risk is measured by analysing potential disclosure scenarios in specific settings and considering the following commonly used concepts:

- uniqueness of combinations of quasi-identifiers (keys) that can be used to identify units that are at risk of disclosure – for categorical variables; these are typically *a priori* measures, which are calculated before the application of (non-perturbative or perturbative) masking methods at the implementation stage; according to the uniqueness approach, disclosure risk is measured by applying some basic rules: *k*-anonymity, *l*-diversity or *t*-closeness, which can be combined, in the case of sample surveys, with additional models accounting for the possibility of disclosing sensitive data in the original datasets as well as in the outputs (population estimates created by applying appropriate sampling weights),
- uniqueness of values in the neighbourhood of original values – for continuous variables; in this case *a posteriori* risk measures are used, i.e., those that are calculated by comparing microdata before and after the application of SDC methods.

Disclosure risk can be measured for each record separately and used to provide protection for selected records deemed to be at risk of disclosure; an alternative approach consists in using characteristics of such at-risk records to create a general definition of disclosure risk for the entire dataset. Therefore, the following levels of disclosure risk can be distinguished from the perspective of SDC:

- individual – for a single microdata record,
- global – for a whole microdata set,
- associated with a hierarchical structure of data – the impact of the hierarchical structure of the microdata on disclosure risk at a given level of the hierarchy is measured for a single record or a whole dataset.

Two types of risk are considered in disclosure scenarios:

- internal risk – when a unit can potentially be re-identified only on the basis of data made available to the user;
- external risk – when the user has access to other sources of data, which can be linked with the released dataset; this risk is much harder to measure because there is usually little or no information about other data sources available to the user, except for what can be inferred from the user's place of employment (e.g., if the user works at a labour office, they are likely to have access to the register of unemployed people, which can be linked with microdata from the LFS survey).

Methods of measuring disclosure risk for tabular data differ from those used for microdata or for output checking. In the latter case, disclosure risk is measured in the safe environment where accredited researchers are granted access to confidential microdata and IT tools and where results of their analyses are verified to ensure they are non-disclosive. To facilitate output checking, all output can be classified into a limited number of categories based on its functional form (tables, regression, etc.) and not on the scope of information contained in the data. Each

category is then labelled 'safe' or 'unsafe'. The fact that a particular output has been labelled 'unsafe' does not mean it will not be cleared for release. Similarly, outputs classified as safe will not necessarily be released outside the safe environment. Except certain well defined and limited cases, outputs classified should be released to researchers with no or minimal changes. Outputs are classified as unsafe in their present form if the likelihood of disclosure is assessed to be high, which means they cannot be released without substantial changes. In this case, the risk of disclosure can be assessed in two ways: either according to the rule-of-thumb model, which focuses on preventing confidentiality errors, or by choosing the principles-based model, where the goal is to minimise the risk of disclosure and to maximise data utility. It should be pointed out that values of risk measures should also be confidential and available only to those responsible for SDC.

The monograph presents the most important SDC methods for microdata and tabular data, which can be divided into two main groups:

- Non–perturbative masking – a family of SDC methods that employ various ways of suppressing sensitive information. As a result, the released dataset does not contain information about certain units or the amount of detail is reduced.
- Perturbative masking – a family of SDC methods consisting in distorting sensitive information in order to prevent its reconstruction by unauthorised users while minimising the loss of information.

These methods can be quite simple (e.g., anonymisation or local suppression) or more sophisticated, including methods based on advanced algorithms and mathematical tools (e.g., microaggregation, noise addition, rank swapping, targeted record swapping, controlled tabular adjustment or the cell key method). All of these theoretical tools are discussed in detail. In addition, there is a comparison of basic properties of the available methods for different types of data. The monograph also includes a presentation of SDC methods for data released in statistical publications, especially descriptive statistics, results of various analyses, charts and choropleth maps. A separate section is devoted to methods of and dilemmas associated with synthetic data generation. It should be emphasised that there is no universally accepted hierarchy of SDC methods in terms of the risk of disclosure and information loss – their usefulness depends on specific data and selected parameters.

Statistical disclosure control is inherently associated with the introduction of uncertainty regarding the values of released variables. This uncertainty is the result of suppressing or changing values in a microdata record or a table cell. Consequently, the application of SDC leads to a loss of information contained in the original data, which can negatively affect the quality of released information or calculations and estimates produced by data users. This is why, in addition to data, users should be informed about the expected information loss resulting from the application of SDC. As pointed out in the introduction, minimisation

of this loss is the second optimisation criterion of the SDC process, apart from the minimisation of the risk of unit re-identification and disclosure of sensitive information. This means that the utility of information in microdata, tabular data and analytic outputs released to users should be as close as possible to that of the original data. This requires effective methods of measuring information loss and ways of minimising it. The monograph explains the very concept of information loss and the most important categories of measures of information loss:

- Measures of changes in variable distributions, based on distance metrics between original and modified values. For example, for each geographical unit in the dataset, the distance between original and modified values is calculated and then the distances are averaged.
- Measures of impact of SDC on the variance of estimates, which account for the difference between variances for average values of certain data subsets or the entire dataset (in the cases of tabular data – columns, rows or the whole table) before and after the SDC process. Another approach is to conduct ANOVA for a selected dependent variable with respect to selected independent categorical variables. In this case, information loss is measured by comparing changes in the components of R^2 (by decomposing total variance into within-group and between-group variance) for the original dataset/table based on original data and the dataset/table modified as a result of applying SDC. The application of SDC can lead to heteroscedasticity, in other words, can cause between-group variance to decrease and within-group variance to increase or vice versa.
- Measures of impact of SDC on the strength of relationships, which involves comparing the direction and strength of relationships between certain phenomena with those observed in the original data. In this case, information loss can be measured by calculating correlation coefficients or conducting tests of independence between corresponding variables in selected breakdowns, in other words, for a specific contingency table. Other approaches can also be used.

The monograph provides numerous examples showing how to construct such measures, how to interpret them, which demonstrate their usefulness for different use cases of released data, including measures of estimation precision.

Nowadays, data collection, processing and analysis cannot be performed without the help of appropriate IT tools. This is also true in the case of statistical disclosure control. Although the development of this branch of statistics is only now starting to accelerate, a number of useful programming tools have already been created to facilitate the SDC process involving digital sets of numerical or symbolic data. The monograph presents an overview of the most commonly used IT tools for SDC. The section starts with the presentation of two, probably most well-known, open source programmes: τ-Argus and μ-Argus. Both were developed by

Statistics Netherlands (Centraal Bureau voor de Statistiek) in the course of a few European projects. They are Java-based programmes, available with and without a bundled JRE7 distribution. Other SDC tools have been implemented in the R software and include a number of dedicated packages, such as sdcTable, sdcMicro, recordSwapping, cellKey, etc. The section ends with a brief description of possibilities offered by other SDC tools.

The protection of data confidentiality is also associated with organisational problems related to releasing official statistics and the need to check them to prevent a disclosure of sensitive information. The monograph presents arguments in favour of releasing statistical outputs, including appropriately prepared unit-level data for scientific research purposes (Scientific Use Files), different ways in which access to such files can be granted as well as specific requirements associated with such forms of release. These include formal requirements that persons or institutions applying for access to microdata must satisfy, as well as requirements regarding the level of protection that the data administrator (typically a NSI, but this can be any data holder) should guarantee in order to prevent unauthorised persons from gaining access to sensitive information.

Structure of the monograph

The monograph consists of six chapters. The first one presents general concepts of SDC and relevant definitions, as well as regulations concerning the protection of sensitive information that are in effect in different countries, including Poland. The chapter contains a description of the main types of data that are released and the significance of metadata, paradata, and additional data from the perspective of SDC. The second chapter is devoted to the risk of disclosure and its measurement. It highlights differences between microdata and aggregated data in frequency and magnitude tables or outputs of statistical analyses. The third chapter contains a detailed description of various SDC methods and techniques that can be applied to the three categories of data mentioned above. It presents potential threats to data confidentiality associated with already published descriptive statistics, charts or outputs of statistical analyses and identifies ways in which such data can be used to reconstruct sensitive information. Aspects relating to information loss are discussed in the fourth chapter, which contains the definition of the problem and the main measures of information loss, including original measures developed by the authors. The chapter also analyses the impact of information loss on the quality of estimates produced from data treated with SDC techniques. The fifth chapter provides a detailed discussion of IT tools for SDC, with emphasis on τ-Argus, μ-Argus and two R packages: `sdcTable` and `sdcMicro`.

The sixth chapter focuses on the organisational aspects of providing access to data and the principles that should be followed in this regard. In particular, this

chapter describes different types of microdata that can be released, principles that should be followed in research data centres (RDC) and appropriate security measures, the process of access control and the scope of authorisation. The manograph ends with a conclusion summarising the main principles and recommendations regarding the application of SDC. There is also a glossary of key terms used in the monograph.

Given the general relevance of statistical disclosure control, the monograph is intended to serve as a practical reference guide for methodologists designing statistical surveys and persons responsible for survey quality and the confidentiality of collected information. For this reason, it contains numerous examples and practical discussions of particular aspects that should make the content more accessible.

Keywords: non-perturbative methods, perturbative methods, disclosure risk, information loss, microdata, tabular data, statistical outputs.

Translated by Grzegorz Grygiel