FUNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

SDC in statistical education - the Polish experience

Tomasz Klimanek[1,2], Tomasz Józefowski[1,2], Andrzej Młodak[1,3]

[1]Statistical Office in Poznań, Poland, [2]Poznań University of Economics and Business, Poznań, Poland, [3]Calisia University – Kalisz, Poland

Abstract

In addition to the development of the principles and tools of statistical disclosure control, it is also important to raise public awareness of why and how these methods are applied. In particular, the public should be educated about the use of SDC to protect privacy on the one hand and to maximize the amount of publicly available information, on the other. Educational efforts in this regard should obviously start with the NSI staff responsible for efficient data protection, and then this knowledge should be disseminated among data users and taught to people who handle statistics and data managers in various institutions and economic entities. In the paper we present some activities in this area that have been undertaken in Poland recently. They include specialized training workshops for members of the team responsible for SDC methods, including current and future experts in this field at Statistics Poland as well as SDC topics covered in the preparatory training for new employees of statistical offices in Poland. We also offer a brief overview of a monograph on SDC methods, which is about to be published by Poznań University of Economics and Business (PUEB), which is addressed to all those interested in ways of handling sensitive data or applying for access to microdata or tabular data from various sources. We will also present the syllabus of a new course entitled "Data confidentiality protection methods", which is going to be launched at PUEB. Other educational initiatives, such as a tutorial on data protection for students writing their diploma theses are mentioned.

Key words: *statistical disclosure control, statistical education, training, preparatory training*

1. Introduction

The growing demand for detailed statistical information in recent years has led to the intensive development of statistical disclosure control (SDC) methods, processes, and tools. The main purpose of SDC is to ensure safe and efficient dissemination of data collected by official statistics. It is therefore necessary to prepare employees of national statistical institutes to efficiently implement SDC processes. NSI staff should have a broad knowledge of various SDC methods and tools, be able to apply them in specific situations and obtain sufficiently safe and useful data to be released. In particular, they should have the ability to choose appropriate SDC methods to handle data from a given survey taking into account its methodology and current (and sometimes even expected future) users' needs.

Also newly employed staff could benefit from SDC training. In Poland – as in many other countries – new employees are obliged to participate in a special training, which is usually the first stage of their professional career in public administration and civil service. We believe this training should cover problems connected with statistical confidentiality and SDC, even if only to a limited extent.

Official statistics are not the only data source that can be used for scientific or analytical purposes. There are many data custodians (both public, such as central government agencies and units of local government, and private, e.g. various economic entities), which collect data they need to perform their activities and tasks. For instance, data from public opinion polls or market research concerning preferences of customers (e.g. audience size of different TV stations or the structure of purchases based on data from cash registers) are very valuable for social studies and are of interest to scientists and analysts. This means that the knowledge of efficient ways of protecting the privacy of their stakeholders or customers is also important for them. In addition, the protection of

data privacy is regulated by various laws, which also require commonly available guidelines. To exemplify how such guidelines could look like, we present a Polish handbook written by current or former NSI employees, which we believe could provide useful knowledge about SDC to various data holders planning to release their data to external users.

The last but not least important group of persons who should be aware of the benefits, drawbacks and expected effects of SDC methods are end users of statistical data. Many of them are students preparing their diploma theses (mainly BSc., MSc or PhD), scientists conducting their research or various business analysts (e.g. in the field of market research). When users are familiar with the principles regulating access to data, general types of SDC methods (e.g. confidentiality rules) and the impact of SDC on data quality, they can then correctly assess the real value of resulting data analyses. Of course, it is difficult to reach all potential recipients with the right message, but by providing relevant information to students one can hope that in time this knowledge will reach all interested parties. Moreover, students often conduct their own polls or targeted surveys and should know how to properly protect sensitive information they collect. This is why we also present a proposal of a course syllabus called "Methods of protecting data privacy" addressed to students of economic studies who cover quantitative methods at the Poznań University of Economics and Business in Poznań, Poland.

The paper is organized as follows. In Section we present the aims and scope of trainings for experts, which cover the main principles of statistical disclosure control and the main stages of the SDC process. Section 3 describes how basic aspects of SDC are presented during the initial training of new NSI employees. Section 4 deals with selected problems and examples of SD education for a wide range of potential users. Section 5 provides the most important conclusions od suggestions for the future.

2. Training of experts

SDC education should obviously start with the NSI staff responsible for efficient data protection. The goal is to provide the staff with a broad knowledge of SDC and how these methods apply to survey methodology in order to maximally protect the privacy of respondents while simultaneously ensuring the maximum utility of data for end users. Basic knowledge about SDC should also be disseminated among data users and taught to people who handle statistics, e.g. data managers in various institutions and economic entities.

The way the SDC process is organised varies across countries. In Poland, there is a dedicated team responsible for SDC methods, which prepares the principles and guidelines for how to conduct the SDC process in various surveys. Moreover, units responsible for particular surveys should have their own methodologists that are responsible for SDC. To ensure that all NSI units in Poland have the required knowledge of SDC, a series of trainings have been conducted. They mainly took place in 2019 and were divided into basic and advanced courses. The basic training covered the following topics:

- general concepts and their definitions:
    - types of data disclosure
    - types of users from the point of view of SDC
- regulations and principles used in the international practice of SDC:
    - regulations in Poland, the European Union and other countries around the world
    - codes of good practices
- basic types of disclosed information – tabular data and microdata:
    - types of disclosed information
    - types of microdata from the point of view of SDC
- disclosure risk and information loss:
    - disclosure scenarios
    - individual, global and hierarchical risk
    - measurement of risk
    - information loss due to the application of SDC
    - ways of measuring information loss
    - the trade-off between minimizing disclosure risk and minimizing information loss
- main methods of protecting sensitive information in frequency and magnitude tables:
    - frequency tables:
        - non-perturbative methods
        - perturbative methods

    - magnitude tables:
        - non-perturbative methods
        - perturbative methods
- SDC methods and techniques for microdata:
    - anonymization as a preliminary stage of protecting sensitive data
    - non-perturbative methods:
        - subsampling
        - recoding
        - local data suppression
    - perturbative methods:
        - noise addition
        - microaggregation
        - data or range swapping
        - rounding
        - secondary subsampling
        - PRAM
- software used for protecting statistical confidentiality:
    - τ–Argus
    - μ–Argus
    - tools of the R environment
- problems connected with the preparation of microdata for external users (public use files) – e.g. data from the LFS or EU-SILC:
    - principles of data disclosure to specific users and their rights and obligations
    - organization and safety of microdata disclosure
        - stationary access (in research data centres)
        - remote access
        - control of microdata use.

The advanced training included specific methodological and technical problems of SDC and relevant practical exercises, i.e.:

- application of dedicated R packages – e.g. `sdcTable` and `sdcMicro`,
- methods of generating synthetic microdata and new algorithms for protecting statistical confidentiality, e.g. the cell-key method,
- problems connected with the preparation of microdata for external users (PUFs), e.g. from the LFS or the EU-SILC,
- methods of protecting statistical confidentiality of census data.

The third training was addressed to members of the general expert team. The topics covered were similar to those in the first two trainings, but they were presented in a more condensed manner. In a questionnaire after completing the training the participants said they were planning to use the presented methods and tools in their daily work. When asked to identify the main problems related to SDC, the mentioned things like the need for more training in the use of R, the lack of competent staff and the lack of experience in the use of SDC.

In the following years (among other things because of the COVID-19 pandemic) SDC training was offered in the form of direct individual consultations or on-line sessions, which also proved to be effective. Currently, the Team for Methods of Statistical Disclosure Control is preparing special guidelines for NSI employees (Statistics Poland (2023)) to support the implementation of the SDC process.

3. SDC issues for new employees of official statistics

The development of competent NSI staff involves equipping new employees with the knowledge and skills they need to perform tasks in the course of civil service. Preparatory service training for people employed in public administration includes learning about basic mechanisms and rules that govern specific public agencies. In the case of NSI employees preparatory service training consists of two parts: obligatory and optional. The obligatory part is divided into basic and extended. The basic part is in fact common for all government institutions and concerns the general functioning of the public administration sector. The extended part covers specific aspects of

official statistics (the legal basis, organization of official statistics, organization of statistical surveys, IT, safety of information, crisis management, international cooperation, publication policy, etc.).

Since 2021, the methodological aspects of SDC have been covered as part of topic 3 (principles of organising statistical surveys): Confidentiality and utility of output data in official statistics[1]. The study materials succinctly describe various aspects of data protection and SDC. Given the scope of the preparatory service training SDC issues had to be limited to the most important ones and included:

- the legal basis of statistical confidentiality
- anonymisation and pseudonymisation
- aims and basic concepts of SDC
- disclosure, disclosure risk and information loss
- user as an analyst and user as an intruder
- the SDC process, its components and stages
- SDC for microdata
- SDC for tabular data
- SDC for statistical outputs.

The materials do not include any mathematical and formal descriptions and do not cover specific details of different SDC tools. The effectiveness of the training is evaluated by means of a quiz at the end of this section to check if the knowledge has been properly acquired. Long-term effects of this training can be only be assessed after a few years at the earliest.

4. Education of potential data users and other stakeholders

As stated at the beginning, tom make sure that SDC methods and tools are well understood and used effectively, the knowledge about SDC needs to be broadly popularised not only among the NSI staff but also among various data holders responsible for the protection of data confidentiality and among end users interested in conducting analyses based on statistical data. The knowledge of SDC is also essential for people who collect any kind of sensitive data as part of their own research.

In an effort to make such knowledge easily accessible, a group of NSI employees from the Statistical Office in Poznań have prepared a handbook entitled *Confidentiality vs. utility of statistical information. Dilemmas of statistical disclosure control* (Młodak et al., 2023). The monograph is an attempt to provide a comprehensive description of all aspects of SDC, 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). 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 are explained. 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.

In particular, the monograph describes SDC methods for microdata, tabular data and output checking, presenting characteristics of each type of data, their sources and fundamental principles of their protection, mathematical methods 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 measures of protecting confidentiality which 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 for estimating re-identification risk. A

[1] The topic "safety of information" concerns only the legal basis of personal data protection and the system of safety management, which is not directly related to the methodology of SDC.

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. Therefore, the authors of the monograph pay a lot of attention to the problems of measuring disclosure risk and information loss. Taking into account various disclosure scenarios, two main types of disclosure risk are discussed:

- 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).

The monograph provides numerous examples showing how to construct such measures and how to interpret them; the examples demonstrate the usefulness of various measures 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.

The publication of the monograph was co-financed from the state budget under the "Excellent Science" programme of the Polish Minister of Education and Science and will be available both in electronic and printed form.

As already noted, statistical data are necessary to obtain reliable and valuable analytical outputs. It is worth noting that many students, when writing their theses (BSc., MSc., PhD) or other studies not only rely on available data from various sources but conduct their own polls, experiments or other surveys. Data collected in this polls may also contain sensitive information, which should be properly protected. Graduates can be also use data provided by official statistics or other custodians as part of their work. It is therefore important they should be aware of SDC rules, methods and tools and their impact on the final output.

It is with students' needs in this respect that Poznań University of Economics and Business (PUEB) has included a special course about statistical disclosure control in the curriculum. Its aims are to familiarise students with key issues and concepts connected with the protection of data confidentiality, to present methods and techniques that help to maintain confidentiality of released data and to teach them how to use dedicated software when working with sensitive information.

After completing this course student should know the aims, principles and basic concepts of SDC, the main stages of the SDC process, specific SDC methods and techniques and should understand the trade-off between minimization of disclosure risk and information loss due to the application of SDC tools. They should also be able to identify potential threats relating to data confidentiality and choose optimal SDC methods and tools in specific situations.

The detailed syllabus of the course includes the following topics:

- protection of data confidentiality – aims and basic principles
- definition of basic concepts
- typology of output data
- legal solutions and formal or ethical principles used in international practice regarding the protection of data confidentiality
- measurement and assessment of the risk of disclosure of confidential information depending on the type of output data
- non-perturbative and perturbative methods and techniques in SDC
- information loss and its measurement
- an overview of selected software used in the SDC process.

This lecture/course is conducted by experts in the field. The group of students expected to participate in the course and the number of lecturers may expand in the future.

It is also important that diploma thesis supervisors should pay special attention to the quality of data obtained by students and how it is affected by SDC and should instil in them the need to always protect sensitive information included in the collected data.

5. Concluding remarks

One of key tasks of modern statistics is to educate as many people as about the aims, rules and methods of SDC. A variety of measures are necessary to fulfil this task on the part of those who are involved in preparing and disseminating statistical data and those who use them. Initiatives and ideas presented in this article have already been implemented and can potentially reach many people, especially of these measures are further developed.

However, the proposed solutions do not exhaust the range of possibilities in this respect. One can think of other ways promoting SDC among various data holders, such as handbooks and guidelines disseminated among employees, organising internal trainings and lectures, providing assistance or consultations, if necessary. Currently, activities in this regard (apart from the SDC User Group within the Centre of Excellence) are still far from sufficient. It is therefore necessary to consider how to organise such cooperation and ensure it can be sustainable in terms of specialised staff involved and financing. Of course, particular organisations operate in different circumstances. While detailed regulations and mechanisms concerning data privacy depend on the sector and type of organisation, there are some common realities, principles and solutions which can be beneficial to all of them.

It is obvious that SDC topics in statistical education should be presented to university students, especially those whose fields of study depend on the collection and use of statistical data. However, pupils in secondary schools also learn fundamentals of statistics and conduct simple data analyses. It therefore reasonable that they, too, should become familiar with the most essential aspects of SDC. One way in which this can be achieved is by organising competitions, such as the European Statistics Competition or national competitions in statistics held in various countries (e.g. the annual Statistical Olympiad held in Poland since 2016). It would be good if such competitions included questions or subjects related to data protection and SDC.

6. References

Młodak, A., Pietrzak, M., Klimanek, T., Józefowski, T. & Lańduch, P. (2023), *Confidentiality vs. utility of statistical information. Dilemmas of statistical disclosure control*, Poznań University of Economics and Business Press, Poznań, Poland (in Polish).

Statistics Poland (2023), *Guide for official statistics service units on the control of statistical data disclosure control*, Collective study of the Team for Methods of Statistical Disclosure Control, Warsaw (under preparation).