Statistical Office in Poznań

SDC in statistical education – the Polish experience

Tomasz Klimanek
Tomasz Józefowski
Andrzej Młodak

Statistical Office in Poznań, Poland

Poznań, September 2023

Introduction

- The growing demand for detailed statistical data requires the use and promotion of Statistical Disclosure Control (SDC).
- The staff of national statistical institutes (NSI) should have a broad knowledge of various SDC methods and tools, be able to apply them to obtain sufficiently safe and useful data to be released.
- Newly employed NSI staff could benefit from SDC training, which should include an overview of basis SDC problems.
- The knowledge of efficient ways of protecting the privacy of data stakeholders or customers is also important for many data custodians, who collect data to perform their activities and tasks.
- Users of statistical data (students preparing e.g. diploma theses, scientists, analysts, etc.) should also be aware of the benefits, drawbacks and expected effects of SDC methods.
- We will present various ways of disseminating knowledge about data privacy protection and SDC among these target groups that have already been applied or that will be introduced in the near future in Poland.

Outline of the presentation

1. Training of experts
2. SDC issues for new employees of official statistics
3. Education of potential data users and other stakeholders
4. Concluding remarks
5. References

Training of experts

Basic assumptions

- SDC education should start with NSI staff responsible for efficient data protection. The goal is to provide them with a broad knowledge of SDC and how these methods can be applied to survey data in order to maximally protect the privacy of respondents while simultaneously ensuring the maximum utility of data for end users.
- Basic knowledge of 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.
- In Poland, there is a dedicated team responsible for SDC methods, which prepares the principles and guidelines on how to conduct the SDC process in various surveys. Moreover, units responsible for particular surveys should have their own methodologists 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.

Training of experts

Basic training for NSI staff

- General concepts and definitions
- Regulations and principles used in the international practice of SDC
- Basic types of disclosed information – tabular data and microdata
- 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 (non-perturbative and perturbative)
- SDC methods and techniques for microdata (non-perturbative and perturbative)
- Software used for protecting statistical confidentiality (τ-Argus, μ-Argus, R tools, etc.)
- Problems connected with the preparation of microdata for external users (public use files) – e.g. data from the LFS or EU-SILC.

Training of experts

Advanced and other trainings within NSI

- The advanced training included specific methodological and technical problems of SDC and 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 featuring similar issues but in a more condensed form was addressed to members of the expert team, who were planning to use the presented methods and tools in their daily work but faced problems such as insufficient experience in the use of R, the lack of competent and experienced staff in the use of SDC.
- Recently SDC training has been offered in the form of direct individual consultations or on-line sessions. The Team for Methods of Statistical Disclosure Control is preparing special guidelines on SDC for NSI employees (Statistics Poland (2023)).

SDC issues for new employees of official statistics

SDC in preparatory service

- Preparatory service training for people employed in public administration includes learning about basic mechanisms and rules that govern specific public agencies and consists of two parts: obligatory (divided into basic and extended) and optional
- Since 2021, the methodological aspects of SDC have been covered as part of topic 3 (principles of organising statistical surveys). The section *Confidentiality and utility of output data in official statistics* includes:
  - 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, tabular data and statistical outputs.
- The effectiveness of the training is evaluated by means of a quiz to check if the knowledge has been properly acquired.

Education of potential data users and other stakeholders

A monograph aimed at popularizing SDC

- The knowledge of SDC needs to be broadly disseminated not only among NSI staff but also among various data holders responsible for the protection of data confidentiality, users conducting analyses based on statistical data as well as 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 (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 and types of released data (including metadata, paradata and other data).
- The disclosure process, types of data users, typologies of statistical outputs with respect to the protection of sensitive information and the trade-off between disclosure risk and data utility are explained.

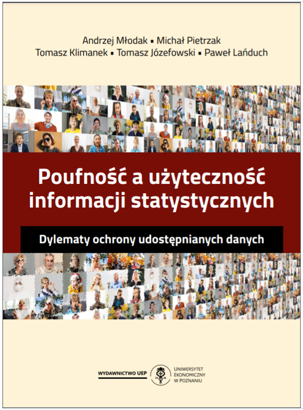
Education of potential data users and other stakeholders

A monograph aimed at popularizing SDC

- 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 attribute.
- The main purpose of SDC is to achieve an optimal trade-off between minimisation of disclosure risk and information loss. Therefore, the authors of the monograph pay a lot of attention to the problems of measuring these quantities (in the case of disclosure risk – in the context of internal and external risk)
- 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.

Education of potential data users and other stakeholders

A monograph aimed at popularizing SDC

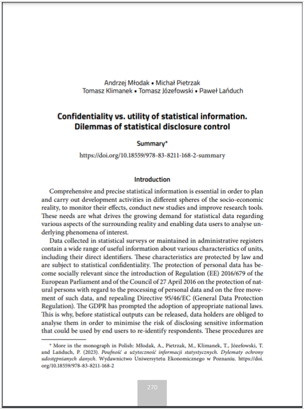


Andrzej Młodak ▪ Michał Pietrzak
Tomasz Klimanek ▪ Tomasz Józefowski ▪ Paweł Lańduch

Poufność a użyteczność informacji statystycznych

Dylematy ochrony udostępnianych danych

WYDAWNICTWO UEP
UNIWERSYTET EKONOMICZNY W POZNANIU



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

270

Fig. 1: The monograph's cover page (in Polish) & the first page of the summary (in English)

Education of potential data users and other stakeholders

Support for students

- 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 these polls may also contain sensitive information, which should be properly protected.
- Graduates can also use data provided by official statistics or other custodians as part of their work.
- Poznań University of Economics and Business (PUEB) has included a special course about statistical disclosure control in the curriculum (to start in 2024/2025). 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
  - to teach them how to use dedicated software when working with sensitive information.

Education of potential data users and other stakeholders

Support for students

- 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 training coverage may expand in the future.
- Diploma thesis supervisors should pay special attention to the quality of data obtained by students and how it is affected by SDC and instil in them the need to protect sensitive information included in the data.

Concluding remarks

Concluding remarks

- One of key tasks of modern statistics is to educate as many people as possible about the aims, rules and variety of SDC methods. Initiatives and ideas presented so far have already been implemented and can potentially reach many people.
- One can also think of other ways of 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.
- Students' education in SDC is very important. However, pupils in secondary schools also learn fundamentals of statistics and conduct simple data analyses and hence they should also become basically familiar with it. 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 problems related to data protection and SDC.

References

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).

Statistical Office
in Poznań

Thank you very much
for your attention!

poznan.stat.gov.pl/en