**Statistical Disclosure Control and Quality Reporting**

Arijana Amina Ramić, Federal Statistical Office of Germany, ArijanaAmina.Ramic@destatis.de

**Abstract**

*The role of quality reporting is to demonstrate that high quality standards have been applied and achieved through in the statistical production processes, and to guarantee correct interpretation and use of the produced statistics. Statistical Disclosure Control (SDC) refers to the measures taken to protect data in accordance with confidentiality requirements, ensuring at same time that the usefulness of the data outputs is preserved to the greatest extent possible. Disclosure control measures reduce data quality (i.e. by suppressing data or changing detail levels), can affect the accuracy of information released (i.e.by data perturbation) or produce bias (i.e. using value rounding or noise addition), and limit access to certain groups (such as researchers). The degree and method of disclosure control may vary for different types of outputs as well as different statistics producers. The peer reviews in the European Statistical System (ESS) show that the highest standards for protection of the statistical confidentiality are applied across the ESS. Some quality indicators, as specific and measurable elements of statistical practice used to characterise the quality of statistics when possible, are also proposed for the information about the statistical confidentiality and security. Nevertheless, for the time being there is no information in the quality reports in the ESS about any disclosure measures applied or their influences at the quality of statistics produced and published. This paper focuses on showing the importance of the presentation of the SDC methods applied and their impacts at the statistics produced, in the quality reports. It provides a small analysis of the benefits for different kinds of the statistics users and of some possible and relevant quality indicators informing about SDC. Some information on the SDC measures applied and/or planned for the Population Censuses could serve as example of good practices of reporting about SDC.*

**Keywords:** Quality Reports, Statistical Disclosure Control (SDC), Quality Indicators on SDC

**1. Introduction**

Statistics is a common reference for society and the quality of statistics depends a lot of the users’ needs. Quality is a multi-dimensional concept, which encompasses all aspects of how well statistics are fit for their purpose. In the European Statistical System (ESS), quality of statistics is managed in the framework of the European Statistics Code of Practice (CoP), which, based on 15 Principles, is setting up the standards for developing, producing and disseminating European statistics. [11] Quality reporting underpins quality assessment, the starting point for quality improvements. Quality reports are used as an instrument to provide the users with information on quality of statistical outputs, to explain the content of the statistics, help the user to understand it and enable its correct interpretation and use. There is a wide range of different possible quality reports - according to the scope of the report, the level of detail, the producer or user orientation, or the perspective of process or output. [11] The guidelines for quality reporting are producer-oriented with a special focus on the statistical process and the ESS system. User-oriented quality reporting is much less detailed, focuses on the output quality and is a sub-set of the detailed, producer-oriented quality report. [11]

Confidentiality is a property of data indicating the extent to which their unauthorised disclosure could be harmful to the interest of the source or other relevant parties. [11] Confidentiality concept includes confidentiality policy - legislative measures or other formal procedures which prevent unauthorised disclosure of data that identify a person or economic, entity either directly or indirectly - and confidentiality data treatment - rules applied for treating the data set to ensure statistical confidentiality and prevent unauthorised disclosure. Statistical confidentiality is a fundamental principle of official statistics enshrined in the CoP and the United Nations (UN) Fundamental Principles of Official Statistics, and covered as the 8th process quality component in the *ESS* *Handbook on Quality Reporting* [11]. The 5th CoP Principle states that the privacy of data providers, the confidentiality of the information they provide and its use only for statistical purposes must be guaranteed. [7] The indicators for adhering to this principle are:

• Statistical confidentiality is guaranteed in law,

• Statistical authority staff signs legal confidentiality commitments on appointment,

• Substantial penalties are prescribed for any wilful breaches of statistical confidentiality,

• Instruction and guidelines are provided on the protection of statistical confidentiality in the production and dissemination processes. These guidelines are spelled out in writing and made known to the public,

• Physical and technological provisions are in place to protect the security and integrity of statistical databases,

• Strict protocols apply to external users accessing statistical microdata for research purposes. [12]

The 6th UN Fundamental Principle of Official Statistics states also that individual data are to be strictly confidential and used exclusively for statistical purposes, while the first four principles provide core rules about microdata confidentiality. [18]

Statistical Disclosure Control (SDC) is a technique used in data-driven research to ensure protection of the confidentiality of the respondents of the research. The aim of SDC is twofold: to identify the risks involved in releasing data, and to modify “risky data” in such a way that for the resulting data the disclosure risk is negligible. [8]

**2. Importance of Reporting About SDC**

SDC is not commonly known to the users and even often to the producers of statistics. Reporting about SDC would raise awareness on needs for SDC, SDC methods and impacts. Further, it would bring users and producers to work together on protection of confidentiality, and share the burden of working on SDC. And finally, it would raise the trust in the producers, which could bring the data providers provide more accurate data.

The most common reason that can cause breach of confidentiality is simple carelessness, such as forgetting to remove direct identifiers or leaving the cells containing sensitive information. Same is by producers and users of provided statistics. Usually researchers/users are not interested in individuals and tend to think of SDC procedures as barriers, not understanding the motivations behind. Some organized trainings on SDC motives, procedures and methods for researchers showed that usually researchers were not interested in attending such trainings and mostly expected from the training to get know how to get some data and how to get over the barriers that SDC meant for them. Mostly researchers do not consider themselves as potential “intruders” and do not absorb a lot of technical details nor retain details of legal framework governing access. [14]

In general, disclosure control measures reduce data quality. A protection method changes the data in order to reduce the disclosure risk and make the data release possible. Much of disclosure risk research focuses on the control side of the disclosure issue, asking: "what do we need to do in order to make this data safe?" However, this question is only one side of the problem that a data provider faces in controlling for risk. [4] Elliot and Purdam found that disclosure control measures had a significant impact on the usability of the data (analytical completeness) and on the accuracy of the data (analytical validity). [4]

Measuring information loss in a generic way reflects how much harm is being inflicted to the data by a given SDC method. There is little information loss if the analytic structure of the masked dataset is very similar to the structure of the original dataset. [3] Non-perturbative methods do not alter the data. They often suppress table cells/data or collapse some categories of variables into a single one. The disclosure risk of data can be lowered by applying non-perturbative methods. However, they might reduce the information content of the data drastically, leading to unnecessarily high information loss. The data structure might also change. Non-perturbative methods can be carried out consistently and tables might remain additive, but it is very difficult to maintain consistency among huge sets of many tables. [1] Data perturbation methods can affect the accuracy of information released. Perturbative methods deliberately change the data slightly. The information loss caused by a perturbative method can often be kept at a lower level than that caused by a non-perturbative method. It implies that users might find the data more useful. Many perturbative methods do not change the data structure. A general requirement for a perturbative method is that its effects on the data do not harm data quality, or at least only slightly [1] Disclosure control can also result in access to data being limited to certain groups, such as researchers. [19]

*2.1. SDC and Statistics Producers*

Official statistics producers are mainly official statistical institutions and to lesser extent other institutions such as ministries, universities, etc. [18] Statistical offices are obliged by the statistical laws to follow the CoP principles and ensure confidentiality for all data they use and produce. The nowadays possibilities for accessing the data are demanding a lot of work on SDC. As a first step statistics producers identify the user requirements for the output and priorities. The mode of access to the data has to be prepared and enabled depending on the intended user group (researchers, policy makers, journalists, the general public, etc.). For each group it should be decided which data should be released to them, or what kind of access they should have to the data. The next sub-process is the identification of potentially disclosive or sensitive data, depending on the type of data. Subsequently, the appropriate solutions to avoid disclosure are applied. But in order to determine the appropriate method, a lot of research has to be done: about the protection of the units as well as individual/single data, about the loss of information whereby the relevance, accuracy, comparison, consistence and additivity of the data is checked, about the usage of the data which includes providing the transparency and describing the complexity of the methods applied, enabling replicability of the data and the inferential statistics, as well as the reapplication of the method at possible new requests on the same data, and also about the resources needed in terms of staff, time, IT tools, etc. Quite often the SDC methods fail to guarantee the required level of protection and the producers are ought to evaluate the results of the disclosure activities and examine the data to assess its “degree of protection”. Evaluation and output checking should be performed by SDC-experts and context-experts, since the context-experts know the thresholds and meanings of variables, while SDC-experts know more about protection methods and the risks connected. [9] Preparing data and applying methods of restricting data or restricting access to data impose different kind of costs – not only for staff and their special trainings, but also for various measurements, tools and facilities. Training in ethical issues related to research, including fair information practices, as well as principles and practices related to research, should be part of the professional training of all those involved in the design, collection, distribution, and use of data obtained under pledges of confidentiality. [15]

Statistics producers consistently strive to minimise the burden on respondents, cultivate good cooperation with data providers and cooperate closely with stakeholder groups including the scientific communities. [7] Although the SDC is obligatory through the statistical laws, the producers maintain the confidentiality and protection also in order to gain the trust and to increase the satisfaction of the data providers. Data providers are the most important factor for the production of the official statistics and the more they trust the producers, the more they will be willing to provide correct and accurate data.

*2.2. SDC and Statistics Users*

In practice there may be a conflict between user demands and SDC. Different user groups have different demands. [8] They all want certain variables with certain detail in the data. This is not always possible due to the SDC applied. Not all users are aware of the SDC, reasons and methods behind; therefore their requests can be too general, too broad or even not relevant for their research. If researchers/users were aware of SDC, they could make concrete requests, without requiring additional irrelevant information, and they would submit relevant metadata about their research, needed for focused production and control by the producers. In this case the output checking would be context specific and no redundant work would be performed. If the SDC was a joint responsibility of researchers/users and the producers, then the common intention to protect data would split the work between them and enable faster and easier availability of needed data.

All risk control methods degrade the data to some extent and therefore reduce the ability of data users to conduct the analyses they need. The effects of SDC fall into two categories: 1. Reduction of analytical completeness - an example is the use of geographical thresholds in microdata sets leading to smaller administration units being grouped together, preventing researchers within those units from effectively using the dataset; 2. Loss of analytical validity - occurs when a disclosure control method has changed a dataset to the point where a user reaches a different conclusion from the same initial analysis. [4] Users are not always aware of these constraints and may blame statistics producer for inconsistent and inaccurate data.

**3. SDC in Quality Reporting**

The main tool for assessing progress in implementation of the CoP is the ESS peer reviews. They are based on a standard information package including replies from the national statistical institutes (NSIs) to a self-assessment questionnaire and the results of a user satisfaction survey. Interviews with representatives of the main user groups, including the scientific community, business representatives, government users, media, helped the peer reviewers to gain an external view on the national statistical institutes (NSIs) and their functioning. The Self-assessment Questionnaire (SAQ) for NSIs is the starting point for the Peer Reviews. [10] In the SAQ the producers are asked about all six indicators of the COP Principle 5 on Statistical Confidentiality, using a couple of sub-questions for each indicator. [10] According to the latest peer reviews reports the greatest strengths of the ESS lied in the areas covered by principles 2 (Mandate for data collection), 5 (Statistical Confidentiality) and 1 (Professional Independence), mainly dealing with the legal framework as well as implementing policies and practices. The peer reviews confirmed that the highest standards to protect statistical confidentiality were applied across the ESS, rooted in statistical law and implemented by internal procedures, techniques and physical safeguards. [5]

Although a lot research is performed for the SDC, not much is published for the users because any additional information might affect the degree of protection. The European and national legislations related to statistical confidentiality applied for the data in question as well as the rules applied for treating the data set with regard to statistical confidentiality are documented/described in the published quality reports.

**4. Possible New Indicators on SDC for Quality Reports**

Quality indicators are statistical measures that give an indication of output quality. They differ from process variables, which give an indication of the quality of the process. However, some quality indicators can also give an indication of process quality. [9] In accordance to the indicators from the Principle 5 of the CoP [11] and the suggestions of the *Canadian Guidelines to Quality Reporting* [17], the United Nations Economic Commission for Europe suggests following quality indicators on SDC to be recorded:

• To which extent is business process using standard or well-known methods identification and protection of sensitive information

• To which extent is data protected from risk of disclosure of sensitive information

• To which extent is data actually protected; what is the residual risk of disclosure

• To which extent has usability of data been degraded; what is loss in precision or level of detail

• Are researchers who have access to micro data legally bound to uphold confidentiality and security protocols of the NSI

• Are research proposals submitted for approval by NSI analysts (analysts must approve the relevance of the analysis and the appropriateness of the methods)

• Are there policies in place that ensure outputs are vetted prior to their dissemination

• Are there confidentiality rules in place, such as a minimum number of units in a cell when doing cross-tabulations, and a maximum number of data requests per day with a maximum number of variables per request (to protect against penetration by an automated data mining process). [19]

All these indices are associated to reporting about "what to do to make the data safe". Only some of them could be of real interest for the users. There are still some more indicators that could be measured in the interest of data producers, but could also be of interest for the data users. Description of the process variables identified in the *Handbook on improving quality by analysis of process variables* [9] suggests the use of Number of cases where statistical disclosure control methods failed to protect the data; Indicators derived from feedback on suitability of disclosure-limited data products and Number of impingements by researchers using restricted data divided by total number of inspections of research data centres. Since these suggestions are not tried out in practice and there is no experience in measuring these indicators, further research is needed to assess them. [9]

Following indicators could be of interest for the users:

• Number of trainings on SDC provided

• Loss information/ Completeness rate before and after/ Distortion i.e. Bias of data before and after/ Suppression rate

• Percentage of records affected

• Percentage of overrepresentation of specific values

• Number of zeros/small counts vs. missing values case same presented

• List of sensitive/altered variables by microdata.

Provision of trainings on SDC for both data producers and data users should be a subject of measurement. An indicator on a number of provided i.e. received trainings could be connected to the statistical literacy and user satisfaction rates. Through this indicator the users could be informed about permanent possibility for trainings on SDC.

A loss function measures the extent of the difference between the original data and the data after disclosure control methods have been applied. For altered data (e.g. perturbation), the relative difference between the data before and after adjustment for confidentiality is measured. In the case of suppressed data, the suppression rate indicating the number of values suppressed compared to those released is often used. In general, for data suppressed in tables, the number of suppressed cells for which protection is inadequate – i.e. a too accurate approximation of the suppressed value can be obtained using information from other cells – must be identified. In the case of microdata, methods tend to measure the risk of disclosure using the re-identification method for a set of characteristic variables (called key variables) or by measuring matching attempts with an external file. Overall, the technique consists of identifying unique combinations of the population found in the released dataset. [19] These indicators cannot be always published but in specific cases could be.

Completeness rate and distortion i.e. bias of data before and after, as well as percentage of records affected, percentage of overrepresentation of specific values and number of zeros/small counts vs. missing values case same presented are relevant for the explanation of changes in quality of the research results.

Certainly additional studies should be conducted in order to measure disclosure risk if publishing more information about the impact of the applied SDC methods. Researchers could be interviewed in a case study for obtaining knowledge about their interest/need in additional information being at most relevant for their studies.

**5. Good Practices of Reporting on SDC**

Population and Housing Census data is an essential source of statistical information. For the 2011 Census different SDC methods were applied by the Member States. Inventory of the country specific data protection regulations and methods showed that around two third of the Member States considered only some variables sensitive, while some countries considered all census variables sensitive and a handful considered none of the census variables as sensitive. The variables that were most often considered sensitive were country of birth and country of citizenship. Cell suppression was the most popular method used. About half of the countries evaluated the disclosure risk and information loss. Many Member States were interested in testing a tool developed by others, so the cell suppression with Tau-Argus was the most cited method for 2021 Census. Identification of best practices shows that implementing cell suppression in a consistent way across countries is very difficult due to the major differences between the disclosure risk concepts and the rules used. Another problem arises with the management of differencing risk between hypercubes and grid level data. Global recoding was excluded due to the fixed hypercubes design. Perturbative methods thus seemed superior. The project team decided to use a combination of pre-tabular perturbation (targeted record swapping) and post-tabular perturbation (random noise). Both methods include parameters that are not fixed, so each country can decide upon them. [2]

Some known information regarding SDC of the Census in the United Kingdom is:

• Consistency and additivity across tabular output is a priority for users;

• Small counts (0's, 1's, and 2's) could be included in isseminated tables provided that uncertainty as to whether the small cell is a true value has been systematically created and creating that uncertainty does not significantly damage the data;

• Microdata are a sample of individual and household records drawn from the census data, which have been anonymised to protect confidentiality. A 3% and 5% sample of anonymised individuals were nondisclosive. A stricter access arrangement was put in place. [13]

Although certain calculations were performed to obtain the information loss, not much has been published. As the most complex and essential product of statistical authorities, Census data are a widely used for the research purposes. Therefore reporting more about SDC and its implications in census would be a great example for the more informative reporting for users.

**6. Conclusion**

There is a tension between providing easy access to research data and protecting confidentiality of those data. Managing this tension requires recourse to technical, administrative and legal solutions. [15] The primary motive for maintaining confidentiality comes from the legal and moral obligations of the statistics producers towards public. The second motive is underlying in the desire of the producers to gain cooperation of respondents and to obtain more accurate data. But democratic society demands wide access to high-quality research data, and good data require the public’s continued willingness to provide information, honestly and accurately. Such cooperation depends on the public’s confidence that its privacy and confidentiality will be respected and no harm caused while cooperating in research. The respondents who believe that their information will remain confidential are more likely to participate in the survey and accurately report their private information. [21] The more information is provided to the users (i.e. more information about SDC through trainings or reports such as quality reports for users) the more confidence they have into producers. SDC has an impact on usability and accuracy of the data. More studies of this area would enrich quality reports for users, providing more for them relevant indicators/information. Knowing more about SDC could help share this task between producers and users, simplify applied procedures, raise the quality of the researchers’ analysis results and improve cooperation between producers and users.

**References**

[1] Antal L. Enderle T. and Giessing S. (2017), Harmonised protection of census data in the ESS, Work Package 3.

[2] Buron M. (2017), Center of Excellence on Statistical Disclosure Control.

[3] Domingo-Ferrer J. and Torra V. (2017), Disclosure Control Methods and Information Loss for Microdata.

[4] Elliot M.J. and Purdam K. (2003), A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records, CASC Project Deliverable 5D2.

[5] Eurostat (2008), Report from the Commission to the European Parliament and the Council on Implementation of the Code of Practice.

[6] Eurostat (2009), ESS Standard for Quality Reports, 2009 Edition.

[7] Eurostat (2011), European Statistics Code of Practice.

[8] Eurostat (2011), Memobust Handbook, Theme: Statistical Disclosure Control – Main Module.

[9] Eurostat (2013), Handbook on improving quality by analysis of process variables.

[10] Eurostat (2013), European Statistical System Peer Reviews/ Self-assessment questionnaire on the implementation of the European Statistics Code of Practice.

[11] Eurostat (2014), ESS Handbook on Quality Reporting, 2014 Edition.

[12] Hundepool A. et al. (2010), Handbook on Statistical disclosure control, Version 1.2.

[13] Office for National Statistics (2009), Report of a Privacy Impact Assessment.

[14] Ritchie F., Green E., Newman J. and Parker T. (2017), Lessons learned in training ‘safe users’ of confidential data.

[15] Singer E. (2005), Access to Research Data: Reconciling Risks and Benefits, Journal of Law and Policy, Volume 14, Issue 1, Article 5.

[16] Shlomo N. (2007), Assessing the Impact of SDC Methods on Census Frequency Tables, in Work session on statistical data confidentiality, Eurostat.

[17] Statistics Canada (2009), Statistics Canada Quality Guidelines, Fifth Edition.

[18] United Nations (2015), United Nations Fundamental Principles of Official Statistics.

[19] UNECE (2017), Quality Indicators for the GSBPM - For Statistics derived from Surveys and Administrative Data Sources, Version 2.0.

[20] United States Census Bureau (2013), Checklist on Disclosure Potential of Data

[21] Yazdani A. (2015), Statistical Confidentiality and Disclosure Control, The Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC).