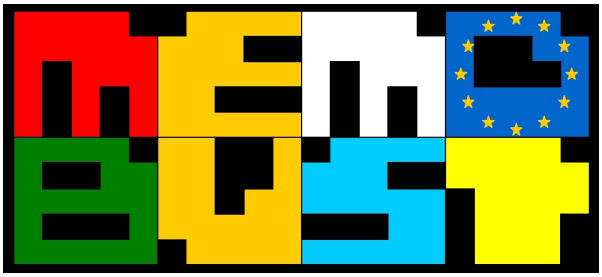

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Disclosure Control – Main Module

Contents

General section

1. Summary

Statistical disclosure control (SDC), or Statistical Disclosure Limitation (SDL) as it is also called, is an activity aimed at the protection of data that are to be released by an NSI. Protection means that individual entities (such as businesses) are not (readily) identified, and more particularly, confidential or sensitive information about such entities is not released to third parties. This to prevent misuse of data intended for statistical purposes. Instead of focusing on aggregates, the attention is directed at individual entities and their response, or the information that is available from them. This shift of attention may even be inadvertent, because certain aggregates happen to consist of one, or a few, entities of which one dominates the contribution.

The aim of SDC is twofold: to identify the risks involved in releasing data, and secondly to modify 'risky data' in such a way that for the resulting data the disclosure risk is negligible. The challenge in modifying the data is to do it in such a way that no (possibly sensitive) information about individual entities is disclosed, directly or indirectly, whereas the protected data are still of interest for statistical research and policy studies. The aim of SDC is not to hamper statistics, but to hamper non-statistical use of the data, such as 'unearthing' information on certain individuals. Statistics is not about individuals but about groups of individuals. So there is room to protect the privacy of individuals whilst serving the interests of society to provide it with statistical information, for research, policy making or general interest. Note that, in the context of this handbook, individuals usually mean individual businesses.

In case of business statistics, tables are the usual pieces of information that are released to users outside statistical institutes. Business populations are usually too skewed so that safe release of business data in microdata form is usually not possible for public use: large units cannot be protected, without rendering the microdata useless. In some countries it may be possible to allow researchers from bona fide institutes to have access to microdata, under strict conditions, and/or in safe settings. But the final results of this research are also in the form of aggregates, such as tables. So in practice, disclosure control of tables is more of an issue for business data than is the protection of microdata. For that reason the focus of attention in the present module is on the SDC of tables.

For tabular data the first task in protecting them is to define rules that separate safe from unsafe data. Once these rules have been specified they can be applied to the tables at hand. In case cells (in tables) have been found that are considered unsafe according to the rules applied, the next thing to do is to try to eliminate them by modifying the tables. For this a range of techniques is available. The problem is to apply them to the tables, in such a way that the resulting tables are safe (according to the rules that have to be considered) and the modification of the tables is minimal. For microdata a similar problem exists, but that will not be highlighted in the present module, for the aforementioned reasons.

For more detailed information about Statistical Disclosure Control issues, we refer to Hundepool et al. (2012), Hundepool and De Wolf (2011), Willenborg and De Waal (2001) and Willenborg and De Waal (1996).

2. General description

Microdata are data about individual entities, such as persons, households, companies, municipalities, etc. At NSIs business data are usually stored in microdata files. These files are used at NSIs as sources to produce aggregate data that can be released to external users. Public release of microdata for business data is typically not an option.

Tabular data are aggregate data, about groups of individual entities. It is convenient to divide the tabular data into two kinds: quantitative tables and frequency tables. Quantitative tables contain data for continuous variables, such as income, turnover, weight shipped, etc. Frequency tables contain numbers of units that have the properties of the respective cells in such a table, such as the number of business involved in a certain business activity in a specific part of a country (province, district, municipality, etc.). Frequency tables are not as often used in business statistics as magnitude tables. Like microdata, frequency tables are more favoured in social statistics than in business statistics. For this reason we focus on the protection of quantitative tables in the present handbook.

So, when publishing business data in the form of quantitative tables, the question is what to be aware of. How to prevent that information on certain businesses is revealed, maybe not exactly but with sufficient precision, by deduction and using certain prior knowledge. Moreover, how to appropriately modify such tables, such that the resulting tables will still be useful but will satisfy the safety rules as well.

2.1 Tables versus microdata

Microdata contain information on individual entities such as business or enterprises in the business statistics area. The individual entities in this area are more complex than those in the social statistics area (persons, households). They are usually also different in terms of size, measured in a variety of ways (number of employees, turnover, profit, etc.). Because of the skewness of certain identifying characteristics of such entities in the business world, it is impossible to release microdata to external users, as the extremer individuals can be recognised immediately. Protecting business microdata using SDC techniques usually does not work, or would produce data that are safe but useless for statistical analyses. So whereas in social research protected microdata sometimes can be released, in business statistics this is not an option.

The publication of business data is therefore typically as aggregate data. This means that data not about individual businesses or enterprises are published, but about groups of such entities. For instance, one might want to publish about businesses providing financial services in the various regions (provinces, districts) of a country. This kind of information is usually published in the form of tabular data, or tables.

However one should not be fooled by the fact that these data are about aggregates, and therefore would need no protection. There is still the possibility that information about individual companies can be inferred from aggregate data, maybe not exactly, but with sufficiently high precision. This happens, for instance, if there is an entity that stands out in a group of entities, in the sense that it dominates this group's total (say total turnover). But in certain cases it is possible to publish this kind of tabular information, but after having modified the original table somewhat. How such tables can be modified in order to make them suitable for publication is the subject matter of statistical disclosure control of tabular data.

2.2 Tabular data

Whereas microdata contain information on individual entities, tables contain information of groups of entities. In other words they are aggregate data. Naively one would perhaps expect that aggregate data do not present any disclosure risks as they are about groups of individuals. But this is generally not true, if only because the size of a group (represented in a table by a cell) may correspond to one individual in the population. Or it may be the case that a group of entities is very heterogeneous with respect to a particular variable, such as turnover. It may very well be that a single individual dominates the individuals corresponding to a particular cell in a table. Publishing the contents of this cell (as part of a bigger table) would disclose the turnover of a particular company, say, in a particular year, with an error that is related to the contributions of the fellow companies represented in this cell. If, for instance, a company attributes 99% of a cell value (where the remaining 1% is attributed by, say, 5 other companies) that cell value is a very good estimate of the contribution of that largest company in that cell. Tabular data are very important for releasing business data. In particular this is true for tables of magnitude data. In the handbook the module "Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables" is devoted to the disclosure limitation of tabular data.

2.3 Trade-off: Probability of disclosure versus information loss

When selecting the method or methods to be used, two competing aspects must be taken into account:

- Probability of disclosure (also called 'disclosure risk'). This is the probability, assuming some kind of disclosure scenario, that there will be an identification of an individual entity.
- Information loss. This is used to express the loss of data utility when applying SDC techniques to a data set or a set of tables. Often this concept is used informally, but in some cases it is formalised in the form of a target function. The SDC problem is then formulated as an optimisation problem. For example the number of suppressed cells in a table may be seen as an example of an information loss measure.

In general, there is a trade-off between disclosure risk and information loss: reducing the disclosure risk will lead to increased information loss, and vice versa. The choice of an acceptable risk level has to be made after careful deliberation by an NSI. This usually depends on the disclosure scenario that is assumed. Because it is easier in practice, the choices will crystallise into a set of rules that can be applied easily in practice, by various groups in a statistical office. Without such a set of rules, protecting data prior to release would be tailor-made, difficult to check, and arbitrary (each department uses its own rules). It is preferable to have a common set of rules, to be used across an NSI.

2.4 User needs and SDC

In practice there may be a conflict between user demands and SDC. The users want certain variables with certain detail in the data, but this is not possible due to the SDC applied by an NSI. There also may be different user groups, with different demands. Policy makers, academic researchers, journalists and the general public may all have their specific requests. The task of the NSI is to manage these requests as well as possible, keeping a firm eye on the protection of the data.

2.5 Data access

Access to business microdata may (in some countries), for instance, only be granted at the premises of the NSI (on site facility), under strict conditions (contractual arrangement, safe settings, controlled access, output checking, etc.). A more recent trend (especially with social statistics microdata) is to allow researchers to have remote access to such data. This access mode has advantages for both researchers and the NSI: the researchers can work with the data at their institute, whereas the NSI can keep the microdata within its own walls, and it can control and log the access to the data. See also the theme module "Dissemination – Dissemination of Business Statistics" in the current handbook.

Access to business microdata is only to allow researchers to use the detailed information they need for their analyses. In the end, however, only aggregate information can be published. The microdata are used as an intermediate data source. Or, in an alternative interpretation, the external users are given similar access rights to these data as the employees at the statistical office working in the area involved.

For restricted access the microdata are lightly protected. Direct identifiers are removed, as well as information that is irrelevant for the research purpose. Some regional variables may be recoded into broader categories if this is possible given the research goals. But this is not a modification comparable to the production of safe microdata for external release (in the area of social statistics). Restricted access is only available to a select group of researchers. A major part of confidentiality issues is dealt with using legal protection, i.e., is agreed upon in contracts. Moreover, each research proposal is evaluated beforehand and only that information is made available that is necessary to conduct this specific research.

3. Design issues

The following aspects need to be designed and organised for tables:

1. The formulation of criteria as to what are safe and unsafe tables. For tabular data there are certain rules that are applied to identify cells in a table that are considered unsafe (due to the dominance of a small group of contributors for such cells). For more information on this see the module "Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables".

2. The measures to be taken to modify unsafe tables into safe data. For instance tabular data can be protected by a combination of table restructuring and cell suppression. Or by rounding, or adding noise. The choice of a method may depend on the user group for which the data are prepared. For instance, to the general public tables with suppressed cells are acceptable, whereas academics would perhaps prefer tables where noise is added for protection. The goal of the measures taken is to produce data that are safe, and with minimum information loss compared to the original data. This aspect, however, we do not consider as a design issue, but as an algorithmic problem. It may involve solving a formal optimisation problem (and sometimes a big one). See the module "Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables".

3. Mode of access to the data, depending on the intended user group (researchers, policy makers, journalists, the general public, etc.). For each group it should be decided what data should be released to them, or what kind of access they should have to the data. There are several

possibilities. Data can be released on a website or in a publication. Certain researchers may be granted access to the microdata, under strict conditions, and via safe settings or via remote access or remote execution. See also the theme module "Dissemination – Dissemination of Business Statistics".

4. Available software tools

τ-ARGUS is a package intended to protect tabular data by various techniques, such as table redesign, various versions of cell suppression, rounding and controlled tabular adjustment. For more information see Hundepool et al. (2011). This package requires a commercial LP-solver (either Xpress or Cplex) for certain techniques (like cell suppression and rounding). The τ-ARGUS package itself, however, is free of charge. See also http://neon.vb.cbs.nl/casc/index.htm

There are other packages for the protection of tabular data, such as sdcTable (R package, no user interface available) and G-Confid (see, e.g., Statistics Canada, 2011). For a general discussion of different software tools, see Giessing (2013).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Giessing, S. (2013), Software tools for assessing disclosure risk and producing lower risk tabular data. Data Without Boundaries Deliverable 11.1 – Part B, February 2013. (http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d11-1b_software-tools-disclosure-risk-assessment.pdf).

Hundepool, A. and De Wolf, P. P. (2011), *Statistical disclosure control*. Methods Series, Statistics Netherlands, The Hague. See: http://www.cbs.nl/en-GB/menu/methoden/gevalideerde-methoden/publicatie-analyse/statistical-disclosure-control.htm.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and De Wolf, P. P. (2012), *Statistical disclosure control*. Wiley-Blackwell.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., De Wolf, P. P., Giessing, S., Fischetti, M., Salazar, J. J., Castro, J., and Lowthian, P. (2011), *τ-ARGUS user manual 3.5*. Statistics Netherlands, Voorburg.

Statistics Canada (2011), *G-Confid User Manual*. Internal report.

Willenborg, L. and De Waal, T. (1996), *Statistical disclosure control in practice*. Lecture Notes in Statistics, vol. 111, Springer.

Willenborg, L. and De Waal, T. (2001), *Elements of statistical disclosure control*. Lecture Notes in Statistics, vol. 155, Springer Verlag.

Interconnections with other modules

8. Related themes described in other modules

1. User Needs – Specification of User Needs for Business Statistics
2. Statistical Disclosure Control – Statistical Disclosure Control Methods for Quantitative Tables
3. Dissemination – Dissemination of Business Statistics

9. Methods explicitly referred to in this module

1. Cell suppression
2. Table redesign
3. Rounding

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. 6.4 Apply disclosure control

12. Tools explicitly referred to in this module

1. τ-ARGUS
2. sdcTable
3. G-Confid

13. Process steps explicitly referred to in this module

1. Statistical disclosure control

Administrative section

14. Module code

Statistical Disclosure Control-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	04-09-2013	first version	Leon Willenborg, Peter-Paul de Wolf	CBS (The Netherlands)
0.2	24-01-2014	revised version after review	Leon Willenborg, Peter-Paul de Wolf	CBS (The Netherlands)
0.3	04-02-2014	minor revision after EB review	Peter-Paul de Wolf	CBS (The Netherlands)
0.4	05-02-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:30