Harmonization of the protection of social statistics at Statistics Finland

Miika Honkala

Statistics Finland, miika.honkala@stat.fi

Abstract: In 2017, Statistics Finland has started a project whose aim is to standardize statistical disclosure control (SDC) practices for social statistics. Statistics Finland already has guidelines for SDC, but the new protection rules will be considerably more accurate. This paper presents the current SDC guidelines and the project where the protection practices are harmonized. The project has studied what sensitivity rules, protection methods and tools are currently used at Statistics Finland. The project needs to make major decisions in the future: what protection methods and tools are used in social statistics.

1 Introduction

In 2016, Statistics Finland started a project called STIINA (Social statistics integrated information architecture). It is a great project that affects the production of all social statistics. Its central objective is to move all population data of Statistics Finland to a same data store. STIINA3 (Methods and new data contents), which began in January 2017, is the subproject of the STIINA. The aim of the STIINA3 is to design common statistical disclosure control (SDC) practices for publishing social statistics. The idea is that when all population data are in the same data store in the future, all users of the population data have same the same protection practices. Nowadays, Statistics Finland has protection guidelines for social statistics but the departments of Statistics Finland have a lot of different solutions for implementing the protection. Constructing the new protection rules, aim is that the rules are considerably more accurate compared to the current guidelines.

The STIINA3 project has studied the current protection methods and tools of Statistics Finland. Processing data is currently carried out using SAS program. Therefore, aim is to find software (or softwares) for protection which can be used with SAS program. Aim is that the protection of tables is more automatic in the future. Harmonization of the protection methods has also challenges because very different social statistics are released. There are statistics related to population in general, education and wages, for example. It is difficult to find the practices that are suitable for all social statistics.

Harmonization of protection is also an international question. In 2016, started project Harmonised protection of census data in the ESS (see the project website). The ESS countries will carry out census in 2021. The aim of the ESS project was to design common protection practices for census tables released in the ESS countries. The results of the ESS project can possibly be utilized in the STIINA3 project.

2 Harmonization of the protection of social statistics

2.1 Current guidelines for protection

Statistics Finland has guidelines for releasing tabular social statistics. Limitation methods, as changing the classification of the variables, suppression and rounding, are currently in use. All these methods offer different possibilities for protection. For example, when using suppression, a lot of different threshold rules are in use. This causes that the different departments of Statistics Finland have various practices in protection, although the practices are based on the same guidelines. Actually, every department releasing social statistics may have own SDC practices.

Varying practices in different social statistics cause some problems. First, almost every protection task has its own expert. In the worst case, if some expert leaves Statistics Finland, nobody can implement the protection tasks. It would be better if many employees could do the same protections. Second, the current system is quite unclear. It is difficult to know what methods and other practices are used in different social statistics. Many departments release quite similar social statistics (from a protection perspective) using different protection methods. Third, it is difficult to check that all social statistics are released according to the protection guidelines. Sometimes the disclosure of the statistical units can unfortunately be possible. There are sometimes released tables containing cells with only one person, and then somebody can have a change to identify this kind of persons. One person in a cell is not allowed according to the current guidelines but it happens, unfortunately. Checking is hopefully easier in future when the new protection practices offer less possibilities to the protection, and when the protection is implemented using same software (or softwares).

2.2 Harmonization of the protection and its challenges

There are several reasons for harmonizing the protection of social statistics. One reason is closely related to the STIINA project. In the future, all population data are in the same data store. Therefore, it is sensible that all employees of Statistics Finland using same population data have similar protection practices when releasing social statistics. Another important aim is that when the protection practices are similar, several employees can carry out same protection tasks. Then there is no problem if someone changes a job. Protection is desired to be more automatic, so that it will take less time. Therefore, the aim is to find common software for protection. The new protection practices will also clarify protection practices because people in all departments know the common rules and so they know what practices are used in the other departments. When the new protection rules offer fewer opportunities for protection, and all social statistics are protected using the same software, it is also easier to control that social statistics are released based on the common practices.

Social statistics of Statistics Finland are currently protected using SAS program or Tau-Argus, but the statistics are produced using SAS. Therefore, aim is to find some SAS software (or softwares) for protection. In the future, all social statistics could possibly be protected using the same SAS software. It will make following the protection practices easier because all tables or unit-level data sets are checked using the same procedure. If all social statistics are released using the same protection tool, the maintenance of the tool is easier.

Harmonization of the protection has many challenges. Statistics Finland have social statistics related to roughly 70 different themes, for example population structure, families, migration, election, education, crime and wages. Some statistics are released in frequency tables and others in magnitude tables. Frequency tables and magnitude tables need different protection practices. It is not clear if a same protection method is suitable for all social statistics or not. Possibly one protection method is not enough, but two or even three methods may be required.

Choice of protection methods is also challenging. What method or methods are allowed? On the other hand, what protection methods are not allowed? It is difficult to define that some methods are not allowed with social statistics. Limiting methods, as suppressing and changing the classification of the variables, are currently used at Statistics Finland, but it could also be interesting to use perturbation methods in social statistics. With perturbation methods, the tables provide more information because all information is visible. The tables do not include holes, for example. However, the common SDC practices cannot offer many different possibilities. The common protection practices may probably contain either limiting methods of perturbation methods, but maybe not both.

There are several possible reasons why limitation methods are in use with social statistics but perturbative methods not. Limitation methods are usually old and maybe easier to understand. People are not necessary familiar with perturbation methods and the methods can be experienced as challenging. Tables released using limitation methods may also seem more reliable because all values in the tables are correct. Perturbation methods cause instead that the tables contain some wrong values. Many researchers do not accept such methods. With perturbation methods satisfying table additivity and consistency of linked tables may also be challenging (Hunderpool et al. 2012). Additivity and consistency in tables are important properties. If some perturbation methods were taken into use, the methods should willingly satisfy these two properties, or at least one of them.

The users of the social statistics have to be taken into account when harmonizing the protection. Statistical outputs have to provide as much information as possible. Harmonization of the protection must not prevent it. At the same time, the common protection practices have to be simple so that the rules do not provide too much possibilities. This is a challenging equation.

2.3 Similar ESS project

In September 2016 started project Harmonised protection of Census data in the ESS. ESS countries will carry out Census in 2021. The task of the project was to give guidance on the protection of the tables for the Census. This project and the STIINA3 project have similar objectives: to standardize protection practices. Therefore, the STIINA3 project can possibly utilize the results of the ESS project.

The ESS project has focused on two perturbation methods: record swapping (more about record swapping, see Shlomo 2007) and cell key method. Record swapping is pre-tabular method (Duncan et al. 2011) where a set of records are sampled from the microdata file. Aim is usually to find match for these records in some other geographic area. Record swapping has also been used earlier for protecting Census tables (Shlomo and Young 2008, Frend et al. 2012). Cell key method is quite new method. The results of the ESS project are very useful for the STIINA3 project because this kind of perturbation methods are not currently in use at Statistics Finland. The STIINA3 project can utilize these experiences about perturbative methods. The experiences can be taken into account when designing the common SDC practices in the STIINA3 project. In the ESS project, program codes for perturbation methods have also been developed and tested. Finland has participated in testing. These program codes, particularly SAS codes, are useful for Statistics Finland.

3 Current SAS softwares for protection

3.1 SAS EG macro

This macro, that has no official name, has been developed in the department of Population and living conditions at Statistics Finland. The macro is suitable for suppressing of tables. Suppressing is the only method the macro can perform. The macro is easy to use. A user has to enter only a few values for the macro, including among other things the name of the dataset and the threshold value. The macro inform primary and secondary cells that the user must suppress. The protection usually takes a couple of minutes.

The macro has some problems. First problem is related to the suppression of zero-cells (cells that have value 0). The macro codes include an option where the user can decide if the macro suppress zero-cells or not. However, the macro suppress always zero-cells. The users often want that the zero-cells are not suppressed, so this is an inconvenient property. Second problem is that the macro chooses the secondary cells in two ways. The secondary cells are marked with two values (2 or 4). It is difficult to know what the different values mean because this has not been documented. The macro codes do not tell it, too. Therefore, the macro needs to be developed so that the user knows how the macro works.

The macro has also some restrictions. The entering table cannot include more than three classifying variables, so that the macro works. Another restriction is that if the classified variable is hierarchical (contains subtotals), the macro cannot find out the cells to be protected. Unfortunately, these restrictions cause that the macro is not suitable for protection of all social statistics.

3.2 SAS2Argus

SAS2Argus ("Sas to Argus") is a SAS macro which uses Tau-Argus in protection. It is in use in some business statistics at Statistics Finland. The macro has been modified so that it works well with business statistics. The macro is not yet in use in any social statistics.

The macro works as a "bridge" between SAS and Tau-Argus (Kraftling 2011). The macro needs Tau-Argus in protection. Tau-Argus must be installed on the computer, otherwise the macro does not work. A user does not need to be able to use Tau-Argus. The macro calls Tau-Argus and transfer a data to a csv file. The macro also makes a metafile which contains information about the variables of the data. Tau-Argus needs that metafile. The macro makes this kind of files which can be done automatically, and protect the tables. The user has to make other files: for example hierarchy file, if needed.

The macro can utilize very many properties of Tau-Argus, but not all of them. A microdata file and tabular data can be used. Different sensitivity rules can be used: threshold rule, dominance rule or percentage rule. The user can protect tables using suppression, rounding or CTA (Controlled Tabular Adjustment) method. The macro has adequate properties for protection. It can possibly be used in all social statistics in the future. The weakness of the macro is that it is difficult to use. The user must be careful in order to get successful protection. Learning the all necessary functions of the macro takes time. These things must be taken into account when considering the general protection tool. The software that is used in all social statistics has to be easy to use, so that all employees releasing statistics can use it.

4 Conclusion

The STIINA3 and harmonization of the protection of social statistics are at the beginning at Statistics Finland. In spring 2017, the STIINA3 project had a definition project where the current SDC practices of social statistics were studied. An implementation project for the STIINA3 will begin in autumn 2017. It will take a few years. Decisions on protection methods or tools were not made in the definition project. The implemention project must make these difficult decisions.

Harmonization of the protection of social statistics is a difficult task. There are released social statistics related to dozens of different topics. Some statistics are released in frequency tables and the others in magnitude tables. It is difficult to find a protection method which is suitable for all social statistics. It is also difficult to find a protection tool that is easy to use, and has adequate properties for protection.

Limitation and perturbation methods also cause a problem. Both have pros and cons. Can both of these methods be used, if the aim is to standardize protection methods? Limitation methods are currently used at Statistics Finland, so it would be natural to continue to use them. On the other hand, tables protected using perturbation methods offer more information compared to limited tables. When harmonizing the protection practices, the users of the statistics need to be taken into account. Information loss has to be as minor as possible.

The end result of the STIINA3 project is hopefully that all social statistics are protected using common, clear practices. It means maybe few accepted protection methods and common SAS software that is easy to use. With new practices and tools, the protection procedure is fast and as automatic as possible.

References

Duncan, G.T., Elliot, M. and Salazar-Gonzales, J.J. (2011). *Statistical Confidentiality Principles and Practice* Statistics for Social and Behavioral Sciences. Springler, New York.

Frend, J., Abrahams, C., Forbes, A., Groom, P., Spicer, K., Tudor, C. and Youens, P. (2012). Statistical Disclosure Control in the 2011 UK Census: Swapping Certainty for Safety. In *ESSnet Workshop on Statistical Disclosure Control (SDC) of Census Data, Luxembourg*.

Harmonised protection of Census data, website,

https://ec.europa.eu/eurostat/cros/content/harmonised-protection-census-data_en (referenced 27.6.2017)

Hunderpool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P.-P. (2012). *Statistical Disclosure Control.* Wiley, Chichester, UK.

Kraftling, A. (2011). SAS2Argus user manual, http://neon.vb.cbs.nl/casc/ESSNet2/SAS2Argus%20User%20manual%20(Ver%201.0).pdf

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review*, 75, 199-217.

Shlomo, N. and Young, C. (2008). Invariant Post-tabular Protection of Census Frequency Counts. In *Privacy in Statistical Databases 2008*. Springler, Berlin.