Specific Grant Agreement (SGA)

Harmonised protection of census data in the ESS

Contract N° 11111.2016.005-2016.367
under
FPA N° 11112.2014.005-2014.533

Date

2/10/2017

Work Package 3

Development and testing of the recommendations; identification of best practices

Deliverable D3.3

Recommendations for best practices to protect the Census 2021 hypercubes

Authors

Sarah Giessing and Eric Schulte Nordholt

Sensitivity

Available to NSIs

1. Introduction

As part of the Specific Grant Agreement 'Harmonised protection of census data in the ESS' Statistical Disclosure Control methods have been developed and tested. In this document recommendations are given how to protect the so-called hypercubes in the next Census round of 2021.

First some background is given in section 2. Then some details are discussed in section 3. Finally, some conclusions and recommendations are given in section 4.

2. Background

Article 20 of the Regulation (EC) No 223/2009 on European statistics makes reference to Statistical Disclosure Control (SDC) in the following way:

Within their respective spheres of competence, the National Statistical Institutes (NSIs) and other national authorities and the Commission (Eurostat) shall take all necessary regulatory, administrative, technical and organisational measures to ensure the physical and logical protection of confidential data (Statistical Disclosure Control).

The NSIs, other national authorities and the Commission (Eurostat) shall take all necessary measures to ensure the alignment of principles and guidelines with regard to the physical and logical protection of confidential data. The Commission shall ensure such alignment by means of implementing acts, without supplementing this Regulation. Those implementing acts shall be adopted in accordance with the examination procedure referred to in Article 27(2).

In 2015 the Working Group on Statistical Confidentiality was consulted on the list of Statistical Disclosure Control topics that should be analysed in more detail by means of collaborative projects between national statistical authorities.

The harmonized approach to the protection of Census data in the ESS was ranked third priority by the Working Group.

Following the expression of interest by the Working Group on Statistical Confidentiality, the Task Force on future EU censuses (gathering representatives of national statistical authorities responsible for the Census data) was approached to discuss the possible scope and objectives of the project. The TF participants expressed their interest in the outcomes of the project and underlined the importance of the proper and harmonised treatment of confidential data. They made some valuable suggestions with regard to the project's outcomes.

Population and Housing Census data represent an essential source of vital statistical information ranging from the lowest small-area geographical divisions to national and international levels. For the 2011 Population and Housing Census European legislation defined in detail a very large set of harmonised high-quality aggregate data in the form of multi-dimensional hypercubes from the population and housing censuses conducted in the EU Member States.

One of the biggest challenges of the 2011 Census round was identification and protection of confidential cells. Due to the high level of detail and multi-dimensional nature of the data transmitted, and the links between various dissemination formats, the treatment of statistical confidentiality was particularly difficult.

This project addressed confidentiality issues that were reported by national statistical authorities while processing 2011 Census data. Moreover, the project addressed the additional difficulty of the 2021 Census related to two parallel non-nested geographical classifications: regional breakdowns (NUTS/LAU) and supplementary classification by grid (1 km^2) squares. These grid data are discussed in deliverable D3.4 and the other Census hypercubes are discussed in this document.

In this project first the country specific data protection regulations and methods were reviewed by means of a questionnaire (in deliverable D2.2 the results of the questionnaire were presented). To be relevant recommendations for treatment of confidential census cells have to take into account the specific restrictions in the countries, the structure of and relationships in the data and the sensitivity of the information. Based on all that information and given the fact that combining rows or columns in the hypercubes is not a feasible protection strategy and knowing that cell suppressions are unwanted, the following protection strategy was proposed. First targeted record swapping is used and then random noise is added (via the cell key method).

Record swapping is a pre-tabular SDC method, and as such, it is applied to microdata. Some pairs of records are selected in the microdata set. The paired individuals/households match on some variables in order to maintain the analytical properties and to minimize the bias of the perturbed microdata set as much as possible. Record swapping exchanges some of the non-equal variable-values between paired individuals/households. The exchanged variables are often geographical variables. Since this exchange introduces uncertainty to the microdata, an intruder's inference about a certain individual/household might not be correct. Record swapping can for example be random or targeted. In case of random record swapping the individuals/households to be swapped are selected with equal probability, while in case of targeted record swapping records of high disclosure risk are determined and a pair to each of these records is selected. It is important to note that record swapping is applied to the microdata. Therefore, at least one of the variables of each hypercube needs to be swapped in order to obtain a perturbed hypercube that is actually different from the original one.

Random noise, as a post-tabular method, is defined by noise probability distributions and by a mechanism to draw from the noise distributions. The variant used by the Australian Bureau of Statistics (ABS) builds on so called "cell keys" to ensure that the random noise added to a specific cell will always be exactly the same, even if the cell appears in e.g. different hypercubes. An implementation of random noise as outlined in the following may involve three "modules":

(1) Cell key module
(2) Module to determine noise based on cell key and noise distribution parameter matrix
(3) Module to restore additivity

The first module enforces consistency of the perturbation, the second determines the statistical properties of the perturbation. The third module achieves hypercube additivity, but might spoil consistency.

Both techniques are controlled by method specific parameters, see sec. 4 of deliverable D3.1 part I for more information. Deliverable D3.1 contains a detailed reasoning why these methods were selected, based on their advantages and disadvantages compared to other methods.

As always in Statistical Disclosure Control, parameters should be determined to yield minimum loss of information while decreasing disclosure risks to a level considered acceptable. The idea behind combining methods is that the combination with the other method might compensate e.g. for a typical type of disclosure risk of a method, allowing this way to choose for both methods parameters that yield a low level of information loss. With this combination of techniques a harmonised approach is proposed to protect confidential cells in the set of hypercubes to be provided. Obviously, countries might also choose to use only one of these methods.

3. Some details

After the decision in December 2016 to make use of the combination of record swapping and the cell key method the ONS kindly provided the census SDC software they had available. In the project team this software was adapted to meet European requirements. As the software has been written in SAS, not all countries are able to test this software on their own Census 2011 microdata.

Slovenia, Finland and France tested the chosen methods with their own data. Even though there were only two selected methods, record swapping and random noise (cell key method), there were lots of different variants of these methods available. For example, the choice of parameter values for record swapping and four different options for the perturbation table (ptable) and ways to deal with additivity (cf. section 4.2 in deliverable D3.1 part II) for the cell key method generate several different variants of these methods[1].

There were some software issues encountered during testing. This is why the project team saw it necessary to prepare a note to help the other countries outside the project to test the software. This note (see Annex A of deliverable D3.2) provides practical information on how to get started with the testing and which kinds of issues have already been encountered (and solved). Through the CROS portal the note was made available in addition to the software (see https://ec.europa.eu/eurostat/cros/content/testing-recommendations-codes-and-instructions_en).

Project members participating in the testing could choose which variants of the methods they wanted to test. All other European countries were invited to test the new methods with the SAS software provided on their data of the Census 2011. Invitation e-mails (together with a link to the CROS portal and with deliverable D3.1 as attachment) were sent to both the members of the Census Working Group (on 23 May 2017) and the members of the Working Group on Methodology (on 24 May 2017).

Many countries reacted positively and were glad that after a number of presentations about this project (see e.g. deliverables D4.1 – D4.4) the software was now made available to test.

[1] Deliverable D3.1 part II does not include the test results for a complex additivity module suggested for the cell key method. For a report on these tests see the paper 'Testing CTA as Additivity Module for Perturbed Census 2021 EU Hypercube Data' by Tobias Enderle and Sarah Giessing of the Federal Statistical Office of Germany (to be presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality in Skopje on 20-22 September 2017).

Some countries indicated they wanted to use one method only: either record swapping or the cell key method or some other method. That would be simpler, but could also lead to more information loss, if stricter parameter settings were used to compensate for not using the other method.

Unfortunately, many countries did not have the human resources to directly test this new software on their Census 2011 microdata. Also the fact that the software is till now only available in SAS made it impossible for quite a number of countries to join the testing. Not all countries have SAS as a standard tool on their network where the data are stored. Further, it should be mentioned that not all countries have enough SAS expertise to test the SAS routines on record swapping and the cell key method. Some countries have indicated that they were interested and downloaded the software, but would like to test later. They are of course welcome to do so, but their results can then not be taken into account in this project. Representatives of countries asked many questions. The general questions posed by many different countries were answered by Eric Schulte Nordholt, while the more technical questions regarding the software (posed by Greece, Malta and Portugal) were answered by Annu Cabrera, Tobias Enderle and Junoš Lukan.

4. Conclusions and recommendations

In this project we have been able to come up with SDC methodology to protect the hypercubes in a harmonised way. The testing (see deliverable D3.2) has indicated that different countries tend to choose different parameter values in applying record swapping and the cell key method. Test results of Slovenia, Finland and France can be found in the three annexes to this deliverable. The two methods (record swapping and the cell key method) can be applied independently from each other. It is too early to claim that this approach will lead to safe census hypercubes in all countries as not all countries have finalised their Census 2021 act and confidentiality rules. However, it looks like there is sufficient flexibility so that by varying parameter values all countries could make use of the confidentiality approach proposed.

With the software codes of this project, of the two methods proposed, it is more straightforward to implement the cell key method for EU census hypercubes. Within the scope of the testing phase of this project some countries participating in the testing have therefore concentrated on this method. An important – though trivial finding – is that with the cell key method, information loss is lowest when the method is applied directly to all cells of a hypercube (including higher level aggregates), while tolerating that lower level aggregates do not add up exactly to the respective margins (cf. Option 1 explained in 4.2 of part II of deliverable D3.1). Simple aggregation of noisy lowest level cells (cf. Option 2), on the other hand, has been found to lead incidentally to margins with very high absolute deviations. A study of a third, complex technique to restore hypercube additivity after application of the cell key method by balancing perturbations (i.e. Option 3) came to the recommendation that if additivity was to be restored at all, it should be restored only partially, i.e. providing for example a separate, perturbed additive hypercube per geographic area foreseen for a Census hypercube[2]. Option 1 (non-additive hypercubes) is a much simpler option, with lower information loss. However, non-additivity has some disadvantages: it requires special communication to

[2] See Enderle and Giessing (2017) in footnote 1

data users, and certain (very) small disclosure risks may arise, as discussed in (Giessing, 2016)[3]. This might be compensated by choosing a stricter parameter setting (which may increase information loss), or by combining with record swapping.

Implementing tests based on the project codes for record swapping is more challenging compared to testing the cell key method (with the "simple" additivity options 1 or 2). Its effect on the data seems to be – at least in the parameter settings used by the testers - stronger compared to the effect of the cell key based noise with option 1 (i.e. without restoring additivity). On the other hand, as record swapping is executed on the microdata, hypercubes protected by record swapping (only) will automatically be additive, and – depending on parameter selections – its effect on the data might even be less compared to that of noise after restoring additivity using option 2.

Also with record swapping there may be concerns of residual disclosure risk – at least when using less strict parameter settings. Therefore, even countries basically preferring record swapping might regard the combination with the cell key method as a way to avoid too strict parameter settings (like high swapping rates).

Given different confidentiality rules and sizes of European countries it is advisable to recommend not just a single method. Recommending a selection of two methods which may or may not be used in combination and can be controlled with parameters and options offers on the one hand a range of flexibility to the countries. It can avoid on the other hand that lots of completely different SDC approaches will be developed and used. Using (both or only one of the) two recommended methods, even with different parameter values, the output hypercubes will be similar enough for comparison of the statistics between countries. Such a similarity could hardly be achieved with completely different SDC approaches.

This project has provided some examples and advice for parameter choices for the methods (see also deliverable D3.2 and the annexes to this deliverable). As described in section 3 of this deliverable testers can choose from four different ptables of which one is recommended since it avoids the complex decisions which zero cells are perturbed into non-zero cells. Disturbing 0s into non-0s can only be used as a proper confidentiality approach if it concerns non-structural 0s. As information about which 0s are structural 0s is often not available, in many countries it is preferred to make only use of disturbing non-0s into 0s. In case of the cell key method, this does not require special ptables and unbiased outcomes can still be guaranteed, but on the other hand this disturbs information on populated grids. Although the project has provided examples and advice, more member states will certainly wish to experiment with the methods.

For this purpose, and for an introduction in the census practice of all European countries, it is preferable to make the software available in e.g. R. When translating the code provided by ONS from SAS to R, the code should also be simplified as much as possible so that more

[3] The inconsistency caused by lack of additivity theoretically allows data users who are aware of the perturbation to try to undo the protection by differencing attacks. Gießing (2016) evaluates different intruder scenarios (See 'Computational Issues in the Design of Transition Probabilities and Disclosure Risk Estimation for Additive Noise'. In: Domingo-Ferrer, J. and Pejić-Bach, M. (Eds.), Privacy in Statistical Databases, pp. 237-251, Springer International Publishing, LNCS, vol. 9867.)

countries will use it. Then the new approach can be incorporated in existing SDC software packages (ARGUS or the R-sdc packages) with a user interface. It is thus highly recommended to support the development of reusable open source software.

Annex A Test results Slovenia

Human resources used for testing: 1 SDC person

Time spent to set up the software and input data (i.e. time until the software runs correctly):

100 hours

Time spent for testing (adapting method parameters, measuring information loss, etc.):

80 hours

Overview of test results (cf. details in deliverable D3.2):

Method tested	Parameter setup	Findings & comments
Targeted record swapping	*10 % swapping rate, see sheet Record_swapping_parameters in test_results_infoloss_SLO.xlsx for other parameters.*	There are not a lot of changes in the swapped data. However, there are some large differences in the final tables. These are due to swaps of large households.
Cell key method (hypercube)	*Ptable_version_01, %let D = 3; %let suspend=; %let pzeroes=yes.*	
Cell key method (1km2 grid)	Same as for hypercube, but parameters for record swapping were different. See the appropriate sheet, again in deliverable D3.2.	The distribution of perturbations is as expected. There are a large number of perturbations, but these are mostly small (both, in absolute and relative sense).

Problems encountered in national environment, if any:

There are some large households in the data, such as retirement communities, prisons etc., which can have several hundred members. In the matching step, they cannot be accurately matched with any other household. Specifically, in step 5 (macro sdc_matching), number of persons in a household is top-coded at 8. This means that in the last step, all households with more than 8 persons are treated as equal. This can lead, however, to large (absolute and relative) discrepancies in the number of persons, whenever a household like this is swapped.

Each time that the record swapping was run, SAS needed to be restarted. Otherwise, the code did run, but the log file and the results that were output were identical to those from the previous run.

It was difficult to determine what all of the variables in the output of record swapping meant. Furthermore, the meaning of record swapping parameters was unclear. Both of

these were resolved in communication with Peter Youens from ONS, who helpfully provided additional clarification of the parameters and variables involved. It was evident other testers had problems similar to these from the emails they sent to Eric Schulte Nordholt, asking for help.

Concluding remarks:

Due to the problem mentioned above, the code for record swapping would need to be amended (such as to prevent the large households to be swapped, for instance).

The cell-key method seems to be appropriate. This mostly goes for showing the data on grids, where having non-additive tables is deemed to be acceptable. Non-additivity presents a larger problem for the data published at the level of local administrative units (NUTS levels).

Annex B Test results Finland

Human resources used for testing (profiles: IT and/or SDC and/or census?):

SDC/methodology staff did the testing. Help with the data was provided by a census specialist.

Time spent to set up the software and input data (i.e. time until the software runs correctly):

150 hours

Time spent for testing (adapting method parameters, measuring information loss, etc.):

130 hours

NB: The amount of time spent for setting up the software and input data and testing is just an estimate based on the total amount of time used for this SGA between February and August 2017. It is difficult to split the time between setting up and testing itself because these were often done simultaneously.

Overview of test results (cf. details in deliverable D3.2):

Method tested	Parameter setup	Findings & comments
Cell key method (hypercube and 1km2 grid)	ptable_version_01, aggregates=yes ptable_version_01, aggregates=no ptable_version_02, aggregates=yes	The results clearly showed the benefit of using the "aggregates=yes" option: the perturbation of total and sub-total cells was under control. The possibility to leave zero cells not perturbed is important. The ptable_version_02 was tested because it is not clear yet if small frequencies (1s, 2s) are desirable in the published hypercubes/tables even though this causes a bit more perturbation.

Problems encountered in national environment, if any:

No problems encountered because of national environment.

Concluding remarks:

The cell key method was by far easier to test than the record swapping. Statistics Finland did not have enough resources to test properly (with any usable results) the complex record swapping code even though time was spent working with it. It should be investigated if for the future use the record swapping code could be simplified and made more user friendly. At least some more documentation and instructions could be developed and maybe even a toy dataset to demonstrate the right structure of the input dataset.

Annex C Test results France

Human resources used for testing (profiles: IT and/or SDC and/or census?):

One SDC expert with help from a colleague working in geography methodology. Neither IT nor census profiles.

Time spent to set up the software and input data (i.e. time until the software runs correctly):

180 hours

Time spent for testing (adapting method parameters, measuring information loss, etc.):

60 hours

Note : As Finland explained it is rather difficult to split the time between setting up and testing. As the setting up also involved some debugging and transformation of the ONS swapping code, it is all the more difficult to precisely know how much time a country will need to use these codes.

Overview of test results (cf. details in deliverable D3.2):

Method tested	Parameter setup	Findings & comments
1km2 grid : Targeted record swapping + Cell key method	*Swapping rate : 10%* *p-table 1*	*We computed information loss measures using population count and 8 other counts (male, female, age under 15,…) between the 4 different sets of data obtained :* *1. original data compared to perturbed data* *2. original data compared to swapped data* *3. swapped data compared to swapped and perturbed data* *4. original data compared to swapped and perturbed data* *With the parameters chosen, swapping add a lot more noise than cell key method.*
Hypercube : Cell key method	*p-table 1,2,3,4 with aggregate=yes*	The code runs well and provide the expected results.

Problems encountered in national environment, if any:

The swapping code takes a long time to run. Due to memory size issues in SAS it was not possible to run the code on the whole dataset and the data had to be split between regions to reduce the number of households in the dataset.

The cell key method is by far faster and easier to use than the swapping code.

Concluding remarks:

The parameters choice is very important and it should enable each country to be satisfied with the level of perturbation. The variables used to target risky individuals and to swap households in the targeted record swapping code are also important parameters that will make it possible for any country to focus on the perturbation on their sensitive variables.