Statistical disclosure control considerations of publishing data on grid squares and territorial units in the application STAGE

Junoš Lukan*, Andreja Smukavec**

* Statistical Office of the Republic of Slovenia, junos.lukan@gov.si

** Statistical Office of the Republic of Slovenia, andreja.smukavec@gov.si

Abstract. The Statistical Office of the Republic of Slovenia (SURS) disseminate some of the geocoded data on a map using the application STAGE. Some of the demographic variables are published on a square grid with the size of the grid cell ranging from 10 km to 100 m and at the level of administrative units, such as statistical regions, municipalities and settlements. In theory, SURS could publish more data in this way, since much of the data are register-based. The interconnectedness of these two geographical classifications presents a challenge, however, for the existing statistical disclosure control methods. While combining different geographical classifications for the dissemination might increase data utility for spatial planning, understanding how disclosure risk is increased due to differencing is crucial. In this paper, a brief overview of the statistical disclosure control methods applicable to geospatial data is given. Some of the practices are tested on a set of demographic variables from the Slovenian Census 2011. Specifically, a procedure developed by the Office for National Statistics (ONS, UK) for the Census 2011, which combines a record swapping method and small cell adjustment, is considered. It is compared with the cell suppression method. As dissemination of some new social indicators on different territorial levels is foreseen, some results of these tests are also presented.

1 Introduction

The use of geospatial (statistical) information is rapidly increasing. There is a growing recognition amongst both the governments and the private sector that understanding of location and place is a vital component of effective decision-making (United Nations initiative on Global Geospatial Information Management, 2013). The ESS 2020 vision states (European statistical system, 2014, p. 3) that "there is a growing need to develop statistics with increasing geographical detail to support national and regional policy making" (e.g. development, implementation, monitoring and evaluation of national and EU policies). The question is how the portfolio of products and services of official statistics can be designed to reflect these information

requirements in the best possible way. To meet the user needs, Statistical Office of the Republic of Slovenia (SURS) wrapped various products and services in a portfolio called STAGE[1].

STAGE is a WebGIS application for cartographic visualisation and dissemination of geospatial statistical data, i.e. statistical data integrated to with spatial information. It consists of a web mapping application and a download service linked to the spatial database. STAGE is designed as an INSPIRE[2] compliant application with respect to the network services and the metadata descriptions. It is based on the Google Maps platform. The application enables the user to present statistical data while selecting the spatial level, content and time reference. In STAGE, geospatial statistical data are presented in time series (when available) for administrative units (cohesion and statistical regions, municipalities and settlements) down to grid cells of $10\,\mathrm{km} \times 10\,\mathrm{km}$, $5\,\mathrm{km} \times 5\,\mathrm{km}$, $1\,\mathrm{km} \times 1\,\mathrm{km}$, $500\,\mathrm{m} \times 500\,\mathrm{m}$, and $100\,\mathrm{m} \times 100\,\mathrm{m}$ size. Slovenia is a register-oriented country and it would be possible to publish a lot of information on a very detailed level. A collision between users' needs and protection of individual information can easily become a very demanding task.

SURS is one of the partners in a year-long SGA[3] "Harmonised Protection of Census Data", which was launched in September 2016. Five other statistical offices are involved: Statistics Netherlands (the coordinator), Statistics Finland, the National Institute of Statistics and Economic Studies (INSEE, France), the Federal Statistical Office of Germany (Destatis) and the Hungarian Central Statistical Office. The project will provide recommendations for the protection of the 2021 Census tables, where additional difficulties will arise due to the publication on two parallel non-nested geographical classifications: regular regional breakdowns (by Nomenclature of Territorial Units for Statistics, NUTS, and Local Administrative Units, LAU) and supplementary classification by grid squares ($1\,\mathrm{km}^2$).

The project team selected two statistical disclosure control (SDC) methods for testing: record swapping and cell-key method (additive random noise). The combination of methods was used in the UK 2011 Population Census. The UK Office for National Statistics (ONS) has agreed to collaborate with the project team and kindly offered their SAS codes to be used for the testing in the project. The results of SURS' testing on grid squares will be presented in this article. As the perturbation methods are not recommended for publications on regular regional breakdowns (NUTS/LAU) in Slovenia, the number of possible SDC methods to use on NUTS/LAU level is very limited. For this article, we tested a simple method of

[1] Accessible at `http://gis.stat.si/en/`

[2] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), OJ L 108, 25. 4. 2007, pp. 1–14, accessible at `http://inspire.ec.europa.eu/`.

[3] Specific Grant Agreements (SGAs) are a specific form of work for specific projects under Framework Partnership Agreement for Statistical Disclosure Control.

cell suppression, implemented in τ-Argus[4], in addition to the two ONS's methods. Of course it is important to check the disclosure risk, which arises due to the publication of the same data on two parallel non-nested geographical classifications.

Additionally, SURS collaborates in a project "Merging Statistics and Geospatial Information in Member States", establishing an integrated geospatial database on income statistics. The desired end result is an increased variety of available statistical information on small geographical areas. The data mostly come from the income tax breakdown and include information on salaries, scholarships, pension benefits, social transfers in kind, etc. Some indicators that are candidates for dissemination include: mean and median income per capita, distribution of income, Gini index, at-risk-of-poverty rate, etc. Testing started in spring 2017 on the frequency tables, where the Slovenian population was split in income quartiles, and will continue until the end of 2017. Some preliminary results are presented in section 3.4 of this paper.

2 Methods of statistical disclosure control of geospatial data

Geospatial data, when shown on a grid or at administrative units, can be easily transformed to tabular data. Therefore, most of the statistical disclosure control methods are just as applicable. Yet, since the data are shown on a map which can be overlaid with satellite imagery, some additional precautions need to be taken.

The problem of statistical disclosure control of geospatial data is approached in various ways in literature. A considerable part of work is focused on ensuring differential privacy (Dwork, 2011; Dwork et al., 2011). This can be achieved by generating synthetic in place of original data, either partially or fully synthetic data (e.g. Abowd, Gehrke, & Vilhuber, 2009; Drechsler, Bender, & Rässler, 2008; Drechsler & Hu, 2015; Saskhaug, 2011).

At present, SURS controls disclosure of geospatial data by employing the cell suppression method (Duncan, Elliot, & Salazar-González, 2011). If the frequency in a cell is under a certain threshold, that cell value is not shown. Both, this disclosure risk definition (Antal, Shlomo, & Elliot, 2014) and this SDC method (Quick, Holan, & Wikle, 2015) have been criticised.

2.1 Methods tested

The Office for National Statistics (ONS) controlled disclosure of the 2011 Population Census using two methods of SDC (Longhurst et al., 2007; Hundepool et al., 2012). The first part is a record swapping procedure, comprising four steps (Frend et al., 2012). In the first step, high risk records are identified. These are flagged according to specified risk variables (e.g. place of birth or employment) and a risk threshold. At each geographical level, risk scores are calculated which correspond roughly to rarity of the attribute at that level as defined by risk variables. The households to

[4]Accessible at `https://github.com/sdcTools/tauargus`

which high risk individuals contribute are flagged as high risk as well.

In the second step of the process, households are sampled for swapping. The high risk households are sampled with a higher probability, but low risk households can be sampled for swapping with finite probability, too. This probability is proportional to the number of high risk records at a particular geographical level and reciprocal to the number of all households at this level.

Households are matched in the third step. This is done iteratively, going from higher to lower geographical levels. The households are paired according to variables (e.g. the age and sex structure and the size of a household) that specify the matching profiles, which can be more or less detailed.

In the final step, the geographical variables are swapped between households. This has an effect of swapping households when shown on a map. All other attributes are conserved in this process.

The second method used by the ONS (Longhurst et al., 2007) is a cell perturbation method, designated the cell-key method. It is an adapted cell perturbation method, originally developed by the Australian Bureau of Statistics for purposes of their 2006 Population Census (see Hundepool et al., 2012).

The aim of the cell-key method is to perturb small value cells in a consistent fashion, so that the cells consisting of the same individuals are perturbed by the same value. This is achieved by using a random, but a prespecified perturbation table. Records are first assigned random record keys. The cell key is obtained by summing the record keys and using the modulo operation. Finally, the cell value and the cell key are looked up in the perturbation table to determine the perturbation value. The tables obtained using this method are non-additive in general.

2.1.1 Settings and parameters

Two geographical hierarchies were considered for record swapping. The first one focused on a $(1\,\mathrm{km})^2$ grid and the data included: $(1\,\mathrm{km})^2$ grid cell $<$ municipality $<$ statistical region $<$ cohesion region. The grid cells were split in such a way, that each fell in one municipality only. This increased the number of 13 425 original $(1\,\mathrm{km})^2$ cells to 14 725 inhabited cells, some of which were split among different municipalities. In the second scenario, only the grid hierarchy was considered as follows: $(100\,\mathrm{m})^2 < (500\,\mathrm{m})^2 < (1\,\mathrm{km})^2 < (5\,\mathrm{km})^2$. There were 159 040, 35 378, 13 425, and 851 inhabited cells of these sizes, respectively.

Four variables were chosen to determine the risk of disclosure for individual record. Records were flagged as risky if there were less than 25, 5 or 1 in the highest, the second, and the third level of geography, respectively. The proportion of households allowed to be selected for swapping was 10 %.

For cell-key method, a perturbation table was used which consisted of noise with (theoretical) variance of 1 and perturbations of ± 3 at most. A setting was used, which prevented zeros from being perturbed, so they were fixed at zero.

3 Results and discussion

3.1 The effects of both methods

The main focus in using both methods of SDC, record swapping and the cell-key method, was first placed on the grid with cells of size $(1\,\mathrm{km})^2$, the reason for which was twofold. First, under the aforementioned SGA, the methods are being tested for SDC use for the 2021 Population Census, because the data are planned to be published on a $(1\,\mathrm{km})^2$ grid[5]. Secondly, specifically for Slovenia, this size of a grid cell was deemed sufficient to publish more detailed age and sex distributions.

Absolute differences between the original and SDC controlled data were first determined and some descriptive statistics were then calculated. Table 1 shows the results of the first SDC scenario, where both, record swapping and the cell-key method, were applied to the data on a $(1\,\mathrm{km})^2$ grid.

Table 1: *Absolute difference (AD) and its statistics when comparing total population counts between original to swapped and perturbed data (O-SP), original data to data that were only swapped (O-S), and swapped to swapped and perturbed data (S-SP) across all* $(1\,\mathrm{km})^2$ *cells*

Statistic/Comparison	O-SP	O-S	S-SP
$\overline{AD}$	1.53	0.94	0.59
σ^2_{AD}	186.0	187.0	0.47
$\max(AD)$	425.0	425.0	3.0
$p(AD = 0)\,[\%]$	51.5	99.3	51.8
$p(AD \leq 1)\,[\%]$	89.5	99.3	90.1
$p(AD \leq 2)\,[\%]$	98.5	99.3	90.2
$p(AD \leq 3)\,[\%]$	99.3	99.3	100.0
$p(AD \leq 6)\,[\%]$	99.3	99.3	100.0

As noted in the second column of Table 1, the mean difference between the original cell values and the ones after applying record swapping and the cell-key method (O-SP) is 1.53. About half of the grid cells were left unchanged and there were only 0.7 % cells where the total population count was changed by more than 3.

The variability of the absolute differences is high, however, as is the maximum difference of 425 records. Comparing the third and the final column of Table 1, it can be concluded the high differences are due to record swapping. Compared to the cell-key method, record swapping only changes a small fraction of all cells, but the

[5]As proposed in Annex I of Task Force's ESTAT/F2/CENS/2016/06 and 07.

resulting differences can be quite high. Even when considering relative distances between original and swapped data, the cell values were altered up to 11.6 times.

Analysing the individual high differences, it was revealed that they are due to single swaps of large households. Some individuals are grouped into households of tens or hundreds of persons, such as retirement communities, campuses, and prisons. After failed attempts to match them with a comparable household, they are swapped with a much smaller household, resulting in a large difference between the original and swapped population count. This could be resolved by preventing large households to be swapped, either by excluding them from the input data set, rolling back the swap after the procedure or flagging them for low swap probability. Such modifications of the code are out of scope of this paper, so in its remainder only the cell-key method is considered.

3.2 Cell-key method results

The effects of the cell-key method were studied on the second geography hierarchy, consisting of grids only. The perturbation table used was the same as the one specified in Subsection 2.1.1.

3.2.1 $(100\,\mathrm{m})^2$ grid

Only the total population count was considered for the $(100\,\mathrm{m})^2$ grid. The absolute differences were again calculated and their distribution is shown in Figure 1.

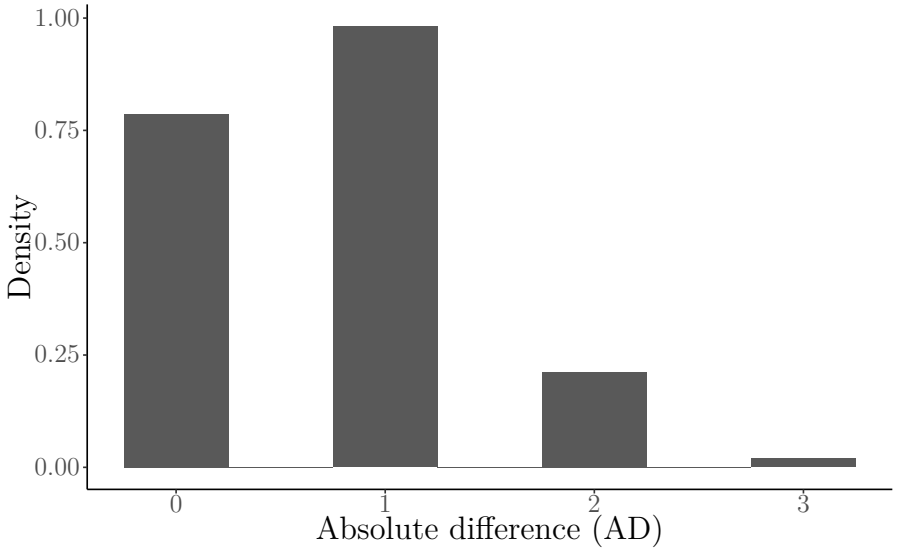

Figure 1: Distribution of absolute differences of total population count between original and perturbed data across all $(100\,\mathrm{m})^2$ cells

About half of the cells were changed for ± 1 as can be established from Figure 1. This distribution of absolute differences mirrors closely the distribution of perturba-

tions in the perturbation table and indeed, they do not differ statistically significantly ($\chi^2(df = 3) = 0.0007, p = 1$).

3.2.2 $(1\,\text{km})^2$ grid

The $(1\,\text{km})^2$ grid was analysed once again in more detail after only using the cell-key method. The results for the total population count followed closely the results for the $(100\,\text{m})^2$ grid. In addition to this item of statistics, the distributions of sex and age were considered.

It was found that the distributions of both, absolute and relative differences were similar for both categories of sex. In the perturbed data, there were 50.518 % of women, compared to the 50.514 % in the original data. These two proportions were not statistically significantly different ($z = -0.09, p = 0.927$).

Figure 2 shows how the grid-cell frequencies were perturbed for different five-year age categories.

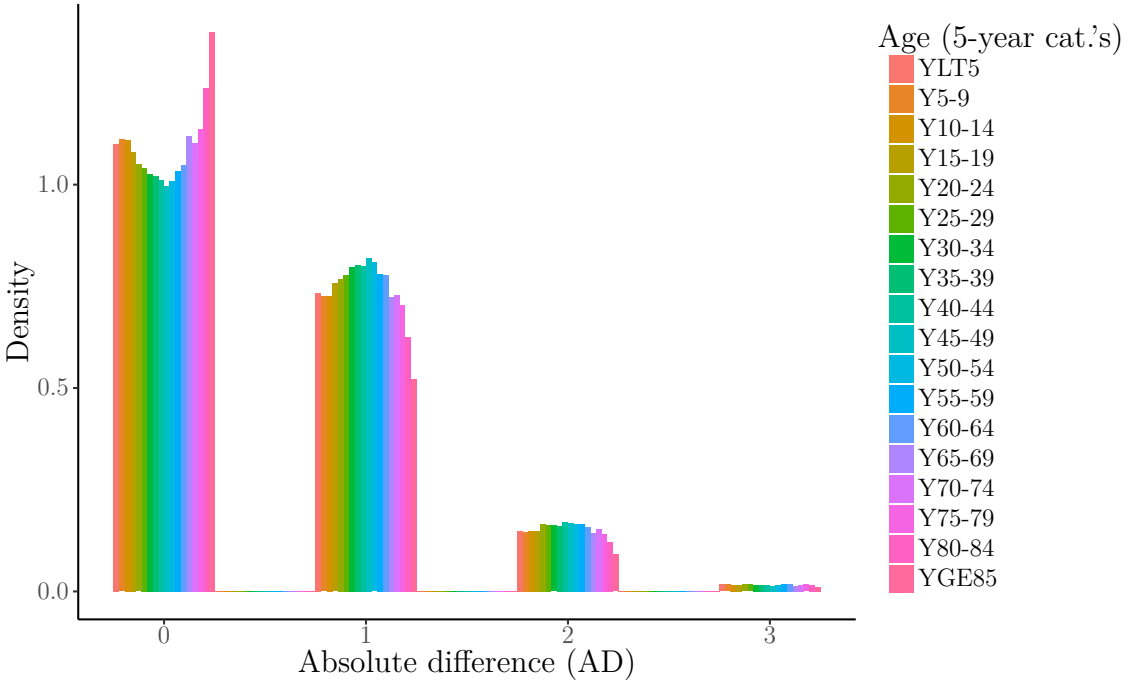

Figure 2: Distribution of absolute differences between the original and the perturbed (by the cell-key method) cell values, shown for different age categories

Comparing Figures 1 and 2, the distributions of absolute differences are different, regardless of the age group under consideration. This is because, for each individual age category, there were many cells with zero frequency. The zero cells were set to be left unperturbed by the cell-key method. Thus, there are more cells with a zero perturbation compared to other values of perturbations. The effect is more

prominent within the highest age groups, since there are the least records belonging to them, that is, there are more zero cells for those categories.

Despite these perturbations, the overall nation-wide age distribution was left more or less unchanged: the result of the chi-squared test was not statistically significant with $\chi^2(df = 17) = 1.44, p = 1$.

3.2.3 $(5\,\mathrm{km})^2$ grid

Since the total population count is on average larger for this larger grid, results pertaining to the total count, sex, and age distribution mirrored the results presented in the previous subsections. Additionally, the geospatial distribution was considered for this grid and is shown in Figure 3. No spatial patterns can be discerned, which was confirmed by a chi-squared test ($\chi^2(df = 850) = 311, p = 1$).

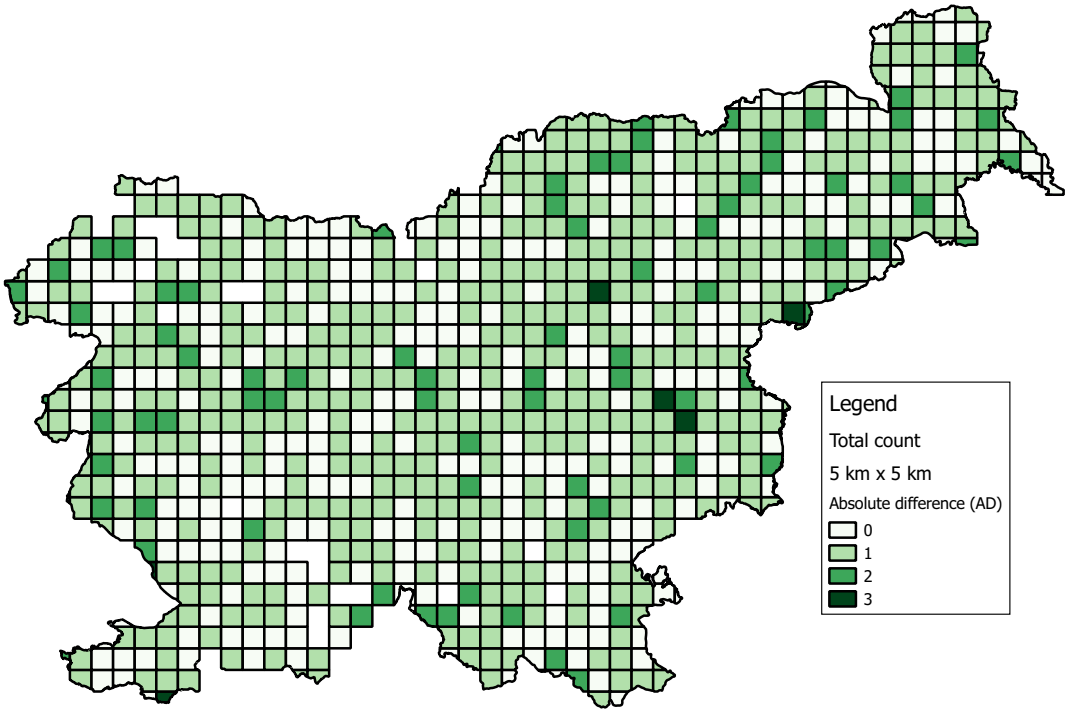

Figure 3: The spatial distribution of absolute differences between the original and perturbed population count

3.3 Comparison with cell suppression

The results of the methods described in the previous subsections were compared to another method of SDC, namely the simple cell suppression. For this purpose,

τ-Argus was used with the "Modular" method of suppression.

Including only the total count in $(1\,\mathrm{km})^2$ and $(5\,\mathrm{km})^2$ cells into a hierarchical table, resulted in some suppressions: over 5 % of the cells had to be suppressed (counting both primary and secondary suppressions) as the result of using minimum frequency count rule of 3.

The information loss is quickly increased in case of tables that are more detailed. A hierarchical table showing the age frequency counts on the largest grids was produced. Over half of the non-empty cells are suppressed in the final table. Moreover, owing to the asymmetric age distribution in the population, the higher age categories are suppressed more often than the others. This produces a skewed age distribution on a $(1\,\mathrm{km})^2$ grid (cf. subsection 3.2.2) if calculated from the protected data, which was confirmed by a chi-squared test ($\chi^2(df = 17) = 748, p < 0.001$).

3.4 Gross income on $(5\,\mathrm{km})^2$ grid

To test the possibility of using the cell-key method for income data, active population was split into four equal groups according to quartiles of the total gross income. The frequencies in each of the groups were perturbed by the cell-key method, where the maximum perturbation allowed was 3 and the noise variance was set to 2.

The overall national quartile income distribution was left unchanged. Also, the perturbed geospatial distribution of the quartile structure did not differ statistically significantly from the original one ($\chi^2(df = 3346) = 152, p = 1$). Indeed, even when comparing quartile distributions within each grid cell, no p-value resulting from a chi-squared test was significant. The minimum p-value was 0.083 and all were corrected to 1 when using the Holm-Bonferroni correction for multiple comparisons (Holm, 1979).

4 Conclusions and outlook

The effects of the three methods of statistical disclosure control were presented in this paper with focus on record swapping and the cell-key method.

Record swapping targets high risk records and their households as specified by risk variables and does not change the total population count in the majority of the cells on the grid. The changes that do occur, however, can be substantial. The cell-key method, on the other hand, perturbs most of the original values, but these perturbations are small (as set in the perturbation table) and rely heavily on the preset perturbation table. The perturbation table used in this paper did not take into account the rarity of a certain attribute, but all values were perturbed in equal fashion. This could be easily amended using a different perturbation table, but rarity is always only considered in terms of a specific tabulation and not with respect to microdata as is the case with record swapping.

The cell-key method has considerable advantages, especially in comparison to the

more often used cell suppression method (see subsection 3.3). Most importantly, it perturbs the data in a consistent fashion (the cells containing the same records are always perturbed in the same way) and retains some of the characteristics of the original data (see subsections 3.2.2, 3.2.3, and 3.4) while the usage of cell suppression quickly increases the information loss.

An additional argument why not to use cell suppression is the publication of the same data on two parallel non-nested geographical classifications. The implementation of the method with the goal to minimize the disclosure risk, caused not only by publication of additive tables on hierarchical administrative geographical levels, but also by publication on grid squares, would be a very demanding task.

It should be noted that the final data are non-additive when the cell-key method is used. The consequence of running the cell-key method separately for each geographical level is that the lower levels do not sum up to the values of the upper levels and the totals of different attributes (gender, age classes) are not necessarily equal. In section 3.4, the applicability of the cell-key method to income data was tested. In general, several differences need to be taken into account between income statistics and demographic variables for which the results were presented in the previous section. In this paper, only frequency tables were considered, whereas the data on income may include magnitude tabular data. Different sensitivity rules pertain to those (such as the p% rule and the dominance rule), which makes risk assessment more complex. The SDC methods themselves need to be amended, too. The cell-key method would need to be adjusted, for example, to perturb using multiplicative rather than additive noise (see Hundepool et al., 2012).

Acknowledgements

This paper is based on work carried out as part of the Eurostat project "Harmonized protection of CENSUS data in the ESS". The action has received EU funding under the grant agreement 11111.2016.005-2016.367. The paper reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

References

Abowd, J. M., Gehrke, J., & Vilhuber, L. (2009). Parameter exploration for synthetic data with privacy guarantees for OnTheMap. *Proceedings of the Joint UN-ECE/Eurostat Work Session on Statistical Data Confidentiality, 2–4 December 2009, Bilbao*. ECE/CES/GE.46/2009/WP.12

Antal, L., Shlomo, N., & Elliot, M. (2014). Measuring disclosure risk and information loss in population based frequency tables. In: J. Domingo-Ferrer (Ed.), *Privacy*

in Statistical Databases: Proceedings. PSD 2014. Lecture Notes in Computer Science, 8744. Cham: Springer. doi:10.1007/978-3-319-11257-2_6

Drechsler, J., Bender, S., & Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. *Transactions on Data Privacy 1 (2008)*, 1002–1050.

Drechsler, J. & Hu, J. (2015). Generating synthetic geocoding information for public release. *Proceedings of the UNECE Work Session on Statistical Data Confidentiality, 5–7 October 2015, Helsinki, Finland*.

Duncan, G. T., Elliot, M., & Salazar-González, J.-J. (2011). *Statistical confidentiality*. Statistics for social and behavioral sciences. New York: Springer.

Dwork, C. (2011). Differential Privacy. In: H. C. A. van Tilborg & S. Jajodia (Eds.), *Encyclopedia of cryptography and security*, pp. 338–340. New York: Springer US.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2011). Differential Privacy: A primer for the perplexed. *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 26–28 October 2011, Tarragona*. ECE/CES/GE.46/2011/WP.26

European Statistical System. (2014). *The ESS vision 2020*. Luxembourg: Eurostat.

Frend, J., Abrahams, C., Forbes, A., Groom, P., Spicer, K., Tudor, C., & Youens, P. (2012). Statistical disclosure control in the 2011 UK census: Swapping certainty for safety. *ESSnet on common tools and harmonised methodology for SDC in the ESS: The Workshop on Statistical Disclosure Control of Census data*.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. & De Wolf, P.-P. (2012). *Statistical disclosure control*. Wiley series in survey methodology. Malaysia: John Wiley & Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–75. Retrieved from: `http://www.jstor.org/stable/4615733`

Longhurst, J., Tromans, N., Young, C., & Miller, C. (2007). Statistical disclosure control for the 2011 UK census. *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 17–19 December 2007, Manchester, United Kingdom*. WP.28.

Quick, H., Holan, S. H., & Wikle, C. (2015, *preprint*). Generating partially synthetic geocoded public use data with decreased disclosure risk using differential smoothing. arXiv:1507.05529v1

Sakshaug, J. W. (2011). Synthetic data for small area estimation in the U.S. Federal Statistical System. *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 26–28 October 2011, Tarragona.* ECE/CES/GE.46/2011/WP.30

United Nations initiative on Global Geospatial Information Management. (2013). *Future trends in geospatial information management: The five to ten year vision* (Revised draft based on feedback provided following the Second Session of the UN-GGIM Committee of Experts on Global Geospatial Information Management, January 2013). New York: UN-GGIM. Retrieved June 21, 2017 from `http://ggim.un.org/docs/meetings/2ndHighLevelForum/UN-GGIM%20Future%20Trends%20Paper%20-%20Version%202.0.pdf`