Disclosure control of Census data by swapping grid-cell identifiers

Steinn Kári Steinsson, Ómar Harðarson, Violeta Calian

Statistics Iceland

1. Introduction

The goal of the present paper is to describe and evaluate the method proposed by Statistics Iceland for protecting the grid-data of the Census-2021. This is both a case-study and the description of a novel way of using data-swapping method of disclosure control: at tabular level.

Although the most frequently used statistical disclosure control (SDC) method for tabular data is the cell key method [1] and the most typical method for protecting microdata is based on random or targeted record-swapping [2], we applied a new approach which consists of swapping, in the grid *tabular* data, the cell-identifiers of all cells from the total set of cells which need to be protected according to the applied SDC criteria.

For instance, the identification risk is addressed by swapping the cell-IDs (i.e. the geospatial location) of all grid-cells having total number of individuals (N) smaller than a critical value (e.g. N0=9). This is done randomly, between grid-cells ensuring that population totals for municipalities and areas of the country remain unchanged by the protection process. The information about the procedure of data protection is carefully communicated to users, albeit without the specific parameters. The Census statistics are not modified in any way by this swapping process.

The variables of the census grid-data (tabular) are defined according to the standard dataset structures submitted to Eurostat: gender, age (0-16, 17-64, 65+), cas (total number of working people), pob.L (country of birth with values: (1) Iceland, (2) Europe, other than Iceland, (3) outside of Europe), roy (place of usual residence, with values: (1) same as last year, (2) different residence but in the country, (3) residence in different country). The total number of cells is over 100 thousand and the percentage of cells with permuted identifiers is between 2-3%, depending on the critical value. The main risk variable is employment, the swapping is defined via cell-grid identifiers and the similarity profile is defined by the size of the cell population conditional on region.

The results are compared with the ones obtained by using the SDC methods in the standard way. In a previous pilot study [4] we have already tested the recommended methods for the new system of small output areas and data of the previous Icelandic Census, but not on the grid system until now.

Evaluation and testing of the proposed solution consist of the following steps: (i) identifying the cells with risk of attribute disclosure, identification, and differencing risk (ii) applying the SDC method based on cell-ID swapping, for multiple values of the critical parameters (iii) evaluating the residual risk and information loss on the output dataset (iv) comparing the results with the record swapping and cell-key methods of data protection. The implementation of the testing, evaluation and risk-utility measurement steps makes use of several packages from the ensemble of SDC-tools [3] and will be shared as open source code[1].

Details concerning the data (section 2), methodology and results (section 3), conclusion and discussions (section 4) are described in what follows.

2. Multi-resolution data

There has been a gradual evolution of Census-data publication in terms of spatial resolution. When developing the statistical program for the Census 2011 data, Statistics Iceland created 42 statistical output areas with an average population size of 7,500 persons. The purpose was to allow for the presentation of the data with areas (regions and municipal subdivisions) of equivalent population sizes, given the huge differences in the local administrative units (LAU) population sizes. The statistical output areas were created respecting main geological, socio-economic and historical boundaries while satisfying the goal of relative equivalence in the population sizes, i.e. having the smallest population no smaller than half and the largest less than double of the achieved average.

For the purposes of Census 2021, Statistics Iceland has created a higher resolution geography, by building a set of minor statistical output areas of approximately 1,500 to 2,000 average population size[2]. It ensures that Iceland will fulfill the small area European requirements for census of population and housing, in addition to the 1 km^2 grid data[3]. As with the major statistical output areas, the minor areas are constructed by making use of information about existing administrative, historical as well as physical boundaries. Standard SDC-methods have been tested for this system already.

All census data are geocoded point-based information, providing sufficient flexibility to publish statistics for the newly proposed Icelandic Statistical Geography Standard (ISGS) hierarchy of statistical output areas (SOAs) but also for any type of territorial classification, including grids, according to the recommendations of European and UN statistical systems regarding the spatial dimensions of census[4].

Two grid-systems have been tested for the Census 2021: the national and the EU-grid, which have small differences due to geospatial reference systems. The total number of observations is 359122. There are 130850 households, 3811 grid cells in the Icelandic grid (3922 in the EU grid), 3 regions, 2 current employment status levels, 3 place of birth levels, 3 age levels and two gender levels.

---

[1] https://github.com/violetacln/testingSDCtools

[2] https://hagstofan.s3.amazonaws.com/media/public/2020/4172be8f-8d25-435b-a62e-f8a6c0afe436.pdf

[3] https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2018-3255714_en

[4] https://ec.europa.eu/eurostat/web/gisco/gisco-activities/integrating-statistics-geospatial-information/geostat-initiative

3. Methodology and results

When regarded through the Census-grid(s), Iceland looks like a "virtual archipelago", i.e. a large set of disconnected populated cells, many of them with a very small number of inhabitants, separated by large unpopulated areas. This means that even aggregated data for such cell-systems pose a high disclosure risk when published.

The main condition we formulated for any valid data protection method is that it should preserve the relevant distributions over regional (and more, e.g. urban/rural) divisions. An additional condition has been imposed, namely producing "credible results". This has been translated into mathematical terms by the requirement to preserve the (approximate) location of certain outlier type of characteristics/groups which are public knowledge.

The research questions of the present study are the following:

Q1: does swapping of only cell-ids of "vulnerable" (targeted) cells provide enough data protection? Should we swap the cell-ids of more cells, e.g. of a percentage of randomly chosen ones?

Q2: is the cell-id-swapping procedure equivalent/similar to the household-id swapping, differing only by the aggregation level? How do we quantify the similarity-relation?

Q3: would the newly proposed method provide microdata protection as well?

Q4: (i) in case Statistics Iceland decides to publish Census results based on more than one grid-system, could we evaluate the disclosure risk increase? (ii) Would two independent cell-id-swapping procedures ensure satisfactory data protection?

Q5: are there group-identification issues need to be resolved?

In order to answer these questions, we used a logical procedure which contains several stages:

S1. Start with microdata used for of the census-grid (D0). Create aggregated version, denoted R0.

S2. Apply to D0 the standard record swapping (household swapping) procedure as implemented in the R-package sdcMicro [3] as a first data protection method. Denote D1 the resulted microdata. Create the aggregated version, denoted R1.

Note: one may produce a variant of this step, by applying the record-swapping by using grid-cell units instead of household-units. Could call this result R2.

S3. Apply the newly proposed method, i.e. permutation (conditional on similarity profile as explained above) of the grid-cell-ids, directly to the tabular version (R0) of the original data. Denote the result by R3. Note that this is basically a perturbation method.

S4. Apply the cell-key method, as implemented in the R-package with same name[5] and denote the result R4. At this step we added a test concerning the perturbation noise, namely that it does not follow any particular spatial pattern, i.e. that the hypothesis of no association between the small area and perturbation count variables cannot be rejected[6].

S5. Calculate risk and utility measures for R0-R4 obtained for varying thresholds involved in defining these steps. This comparison should be made carefully, for both available Census-grids. One should consider for instance issues such as the fact that [5] "risk measures based on frequency counts (k-anonymity, individual risk, global risk and household risk) cannot be used after applying perturbative methods since their risk estimates are not valid. These methods are based on introducing uncertainty into the dataset and not on increasing the frequencies of keys in the data and will hence overestimate the risk".

S6. In order to investigate whether publishing both grids can lead to differencing disclosure issues, we created one more system of small areas which consists of the areas belonging to both grids and the areas belonging only to one or the other. If the cell-ids of the Icelandic-grid and EU-grid are permuted independently, the resulted system has a very low differencing disclosure risk.

The computations and comparisons we have completed so far show that one could answer into affirmative to questions Q1 (conditioning on region and threshold value of N0), Q2 (although the quantification is not well defined yet), Q4, Q5. The answer to Q3 is still inconclusive and one might need additional conditions to reach that goal, due to risk enhancing conservation of group identities. While evaluating the utility/information loss, we preferred entropy-based measures although employed all the standard ones implemented in the R-version of the SDC-tools.

4. Conclusions and discussion

We tested, in order to evaluate it and decide on its future use, a new procedure for disclosure control of aggregated data and tested it for the case of Census grid data. The advantage of our proposed method is that it could be automatically and straightforwardly applied to any tabular dataset and that it preserves consistency at chosen levels of aggregation.

This is still a work in progress but updates in terms of code and conclusions will be reported/linked on the open source repository[7] which already includes the information concerning our project on SDC for the small output area system and preliminary code for the new method evaluation. As verified in [4], we can confirm that the most critical stages in applying and evaluating an SDC method are: the identification of risk variables and the risk-utility analysis. The former is a rather subjective process which is based on legal, cultural and information types of conditions [1]. The latter is the object of an interesting statistical problem, i.e. evaluating the effect of multivariate transformations (as implicitly defined by all methods employed here) on multivariate data distributions. Measures for both risk and utility

[5] https://cran.r-project.org/web/packages/cellKey/index.html
[6] https://www.rdocumentation.org/packages/DescTools/versions/0.99.37/topics/GoodmanKruskalGamma
[7] https://github.com/violetacln/testingSDCtools

(standard, information based) should be used to define the parameters of the optimum regime of the employed SDC method.

As future work, we plan to investigate several new directions for protecting tabular data by employing synthetic data, although this is a usually a less popular line of research due to the „output“ character of aggregated data, such as: (i) using Bayesian estimates instead of raw counts and (ii) using deep-learning/cryptography inspired methods.
It is of high importance for a country like Iceland, where the risk of disclosure is encountered so frequently, to ensure data protection in a *systematic* manner and as *automatic* as possible.

References

[1] Hundepool, A., DomingoFerrer, J., Franconi, L., Giessing, S, Nordholt, E. S., Spicer, K., de Wolf, P.P., Shewhart, A., Wilk, S., Statistical Disclosure Control, Wiley, 2012.

[2] Shlomo, N., Statistical Disclosure Control Methods for Census Frequency Tables, International Statistical Review / Revue Internationale de Statistique, 75(2), 199-217, 2007.

[3] Templ M, Kowarik A, Meindl B (2015). “Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro.” *Journal of Statistical Software*, **67**(4), 1–36. doi:10.18637/jss.v067.i04.

[4] Methods of statistical disclosure control for aggregate data. With a case study on the new Icelandic geospatial system of statistical output areas, Violeta Calian. Statistical series: Working papers, 105(6), 2 September 2020, https://hagstofan.s3.amazonaws.com/media/public/2020/e9ea7160-5032-4580-9297-7b3b3cb634da.pdf

[5] Benschop, T., Machingauta, C., Welch, M. (2022) Statistical Disclosure Control: A Practice Guide, https://buildmedia.readthedocs.org/media/pdf/sdcpractice/latest/sdcpractice.pdf

[6] Shlomo, N., Antal, L. and Elliot, M. (2013) Measuring Disclosure Risk and Data Utility for Flexible Table Generators, Joint UNECE/Eurostat work session on statistical data confidentiality, https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf