Hagtíðindi
Statistical Series

Greinargerðir
Working papers

2 September 2020

Methods of statistical disclosure control for aggregate data

With a case study on the new Icelandic geospatial system of statistical output areas

Aðferðir við hindrun rekjanleika í samanteknum gögnum

Athugun á miðlun upplýsinga um ný íslensk smásvæði

Abstract

The goal of the present paper is to explain and test the main statistical disclosure methods and their R-implementations for aggregate data protection. We apply these methods to a 2011 census dataset enriched with the geographical attributes of newly built small output areas of a high resolution Icelandic geography. Several measures of risk versus utility are examined in order to evaluate the performance of the methods.

Útdráttur

Markmiðið með þessari greinargerð er að útskýra og prófa helstu aðferðir við hindrun rekjanleika í samanteknum gögnum með tölfræðihugbúnaðinum R. Til að sýna fram á virkni aðferðanna er þeim beitt til að hindra rekjanleika í niðurstöðum íslenska manntalsins frá 2011. Niðurstöðurnar byggja á gögnum sem hafa verið auðguð með svæðisbundnum þáttum sem eru byggðir á nýjum smásvæðum fyrir Ísland. Aðferðirnar eru metnar út frá ýmis konar mælingum á nytsemi og áhættu við birtingu niðurstaðnanna.

Acknowledgement

This work is partly supported by the *Eurostat grant 831732 - 28-IS-Merging*.

Statistics Iceland

Introduction

This study describes the main methods for statistical disclosure control (SDC) for tabular data which will be implemented by Statistics Iceland for protecting published datasets with high number of dimensions. The purpose is to test the record swapping and cell-Key perturbation methods which we define in what follows and establish an efficient and reliable workflow for the dissemination of large tables. The recent development of a new system of small geographical regions built for the 2021 Icelandic census created the opportunity to test census hypercube data combined with spatial information of high resolution.

The paper is organized as follows: the main SDC concepts are defined in the first section and the new small output area statistical system which will be employed for the 2021 census is described in the second section. The third part explains the SDC methods and the purpose and types of risk-utility analysis. The fourth is dedicated to the case study of protecting census data with high resolution geospatial information. We conclude with comments and future work plans.

Statistical disclosure control for aggregate data

Disclosure risk and protection

Disclosure takes place when information about the identity or attributes of persons or organizations can be learned from disseminated data [1]. National statistical institutes systematically publish high quality statistical outputs which need to be protected from disclosure risks while keeping high standard for information content.

Most official statistical data sets are aggregate data type, i.e. tabular data, containing count and magnitude type of variables. Census data is a typical and celebrated example. Disclosure risk can manifest itself as small counts, as attribute disclosure or as disclosure by differencing, i.e. comparing two different tables which for example have different geographical levels of detail, such as grids and NUTS classification[1] system.

The methods for protecting the published data against disclosure risk can be applied to microdata (pre-tabular) or to aggregate data (post-tabular) and they can alter or not the data. The non-perturbative methods do not change the data and rely instead on suppressing cells defined as high risk (primary and secondary level of exposure) or on globally recoding variables, i.e. combining several categories of given variables in order to decrease the risk of disclosure but decreasing information content in the process. Perturbative methods, such as record swapping or applying random noise, do not change the data structure but alter slightly the data values while keeping the quality and information content intact.

Harmonized census data protection across Europe

Eurostat has supported a process of evaluation of statistical disclosure control methods and decided on harmonizing the methods across Europe [2], in view of having consistent Census national data sets which can be combined in a systematic way into European – level data. Eurostat recommends the use of targeted record swapping (pre-tabular method) and the random noise method (post-tabular), based on the extensive experience of Statistical Institutes such as Australian Bureau of Statistics (ABS) and on theoretical studies e.g. [3]. These methods do not change the structure of Census hypercubes but keep the information loss at minimal levels while being easy to adjust according to each country‘s needs.

[1] https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Nomenclature_of_territorial_units_for_statistics_(NUTS)

High resolution spatial statistics at Statistics Iceland

Main purpose of the geospatial project

Statistics Iceland created a higher resolution spatial statistics production for census 2021, by:
(i) building an improved Icelandic Statistical Geography Standard (ISGS), i.e. hierarchy of small area geographies designed for register based census data (the statistical output areas, SOAs)
(ii) formulating a geospatial strategy and methodology for referencing administrative records
(iii) using the Icelandic INSPIRE geospatial data together with census 2011 data as test data

Description of the project

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.

In the past months 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. 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[2]. As with the major statistical output areas, the minor areas are constructed by taking into account existing administrative, historical as well as physical boundaries.

All census data are geocoded point-based information, providing sufficient flexibility to publish statistics for the proposed ISGS hierarchy (of 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[3].

Methods for SDC of aggregate data

Record swapping

This is a SDC method which is applied to microdata with the purpose of protecting the aggregate, tabular datasets. According to this method, records of microdata describing attributes of individuals/households are paired so that some of the attributes match (such as household size) and some of the non-matching attributes (such as geographical location) are swapped between the pairs. A good review of the method and a risk utility assessment analysis is given in [4].

Statistics Iceland is testing the R-implementation as the package *recordSwapping* which is still an evolving open source project[4]. As explained by the authors, „The implementation of the procedure was done purely in C++ and is based on the SAS code on targeted record swapping from ONS (https://ec.europa.eu/eurostat/cros/content/2-record-swapping_en). There are however substantial differences between the SAS- and C++-Code. Some of these differences are the result of improving the run-time for the C++ implementation. The R-Package is just as a front end to easily call the procedures and for testing purposes.“ Swaps are made according to k-anonymity criterion of

[2]https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2018-3255714_en
[3]https://ec.europa.eu/eurostat/web/gisco/gisco-activities/integrating-statistics-geospatial-information/geostat-initiative
[4]https://rdrr.io/github/sdcTools/recordSwapping

SDC, at all levels of the data hierarchy. This criterion states that the information for any person contained in a dataset cannot be distinguished from at least k-1 individuals whose information is also included in the dataset.

Random noise method

This is a post-tabular method and consists of adding random amounts to count-cells, the amounts been defined by given noise probability distributions and drawing mechanisms. In order to ensure consistency between data (e.g. census) hypercubes, any given cell has same noise added irrespective of the hypercube it occurs in. It has originally been implemented at the Australian Bureau of Statistics as described in [5] and developed further as a Eurostat supported project, as shown in the documentation related to the R-implementation as the package *cellKey*[5] and its companion *ptable*[6], both included in the sdcTools suite dedicated to the harmonized data protection methods.
The implementation includes three modules: (i) the cell-key module, which enforces consistency of perturbation, (ii) the random drawing module, which determines the noise amount depending on the cell-key and on the noise distribution parameter matrix and (iii) the additivity restoring module. The record keys and perturbation defining tables are designed to fulfil a set of conditions such as fixed variance, mean zero, absolute values of perturbation amounts less than integer threshold values, i.e. the noise will not change the data in a significant way.

Risk-utility analysis

Two factors should be taken into account when choosing and tuning the SDC method for each dataset. One is the residual risk of disclosure for the protected data and the other is the loss of information caused by data protection, i.e. the utility factor.
The risk may be evaluated, as shown in [6], by various measures: (i) the inverse of the variance of the confidential counts, (ii) the percentage of cells left unchanged, (iii) the probability of an observed difference of 1 (or higher, in a more general setting) corresponding to a true difference of 1. An extended disclosure risk measure based on Information Theory was recently proposed and tested in [7], which has the properties of taking values between 0 and 1 and it depends on the conditional entropy of the original distribution given the confidential distribution as well as on the overall population size of the table and the number of zeros.
The theoretically [7] and implementation-wise (such as the R-package *cellKey*) accepted measures of utility/information loss are: the average percentage change, the average absolute difference, the mean variation and the Hellinger's Distance, the absolute distance between original and perturbed values (d1), the relative absolute distance between such values (d2), the absolute distance between square-roots of original and perturbed values (d3).

Decision on SDC method settings

The optimum regime of minimum information loss and maximum risk protection is not uniquely defined. Few studies have been dedicated to the evaluation of the impact of SDC methods on these competing effects. Census tabular outputs'protection methods have been revised in a risk-utility framework in [9], [10], while in [11] a theoretical comparative study was implemented in the context of attribute disclosure and cell perturbation, random record swapping and random rounding. The articles in [12] and [7] define global disclosure risk measures based on information theory as developed in [13] and [14] and build corresponding probabilistic models, in addition to using information theory based utility measures.

A competing criterion has been proposed in [8], based on so-called acceptance limits and formulated as: "A change is acceptable, if (A OR B) AND C, where: rule A says: "absolute difference is less than a"; rule B says: "relative

[5]https://github.com/sdcTools/cellKey
[6]https://github.com/sdcTools/ptable

absolute difference is less than r%"; rule C says: "square root distance is less than s". An information measure based on this criterion should involve the rate of hypercube cells with acceptable changes defined by A OR B) AND C, would depend on the parameters a,r,s" and it is still under development.

Case study

Data

The data used for this study consist of 2011-census records enriched with geospatial information regarding the recently built small output areas (SOAs) and the municipality level. The total number of observations is 315556. There are 121421 households, 196 small areas, 76 municipalities, 8 educational attainment levels, 8 current activity status levels, 170 place of birth levels, 102 age levels and two gender levels.

Results of testing the record swapping method

The test data contains a hierarchy of two geographical levels as described above, two socio- demographic variables (educational attainment and gender) in addition to household and household size variables. The risk variables are defined as gender and education level, the swapping is defined via household identifiers, the k_anonymity is set at 3, the similarity profile is defined by the size of household.

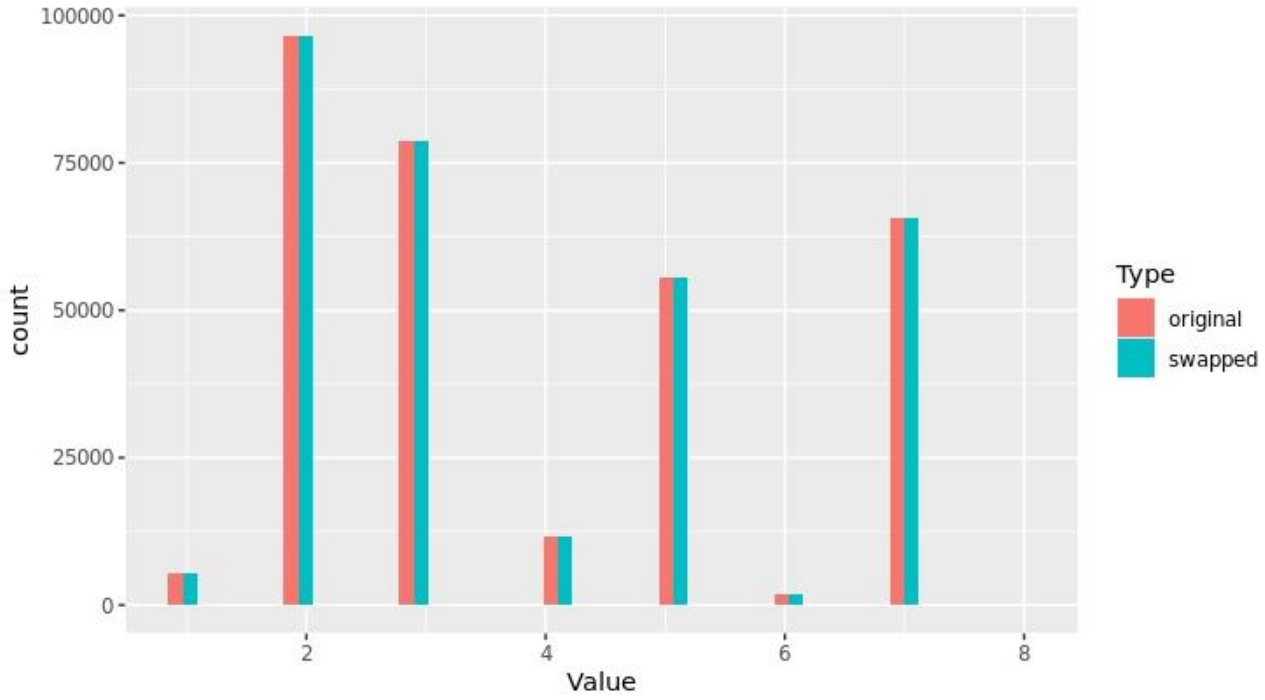
The resulted data frame obtained by using the *recordSwap()* function of the package *recordSwapping* was compared to the original. The number of resulted swapped households was 15778 for this example.

Figure 1. Example: R-table of the first few lines of the microdata with swapped geographical information

```
>md_swapped
        edu sex svf smsv household n
     1:   5   1   1   11         1 1
     2:   2   2   1   14         2 1
     3:   2   1   1    1         3 2
     4:   2   2   1    1         3 2
     5:   3   2   1    1         5 1
```

Similar tests were made for different demographical attributes but same geographical hierarchy. An illustrative example is shown in Figure 2, where the educational attainment categories do not show any difference in frequencies between the original and swapped data, confirming the consistency of the method.

Figure 2. Educational attainment categories do not change marginal counts after record swapping



Results of testing the cell key method

The test data contains the same attributes and geographical information as in the previous tests. The basic perturbation table settings were chosen as the default values offered by the R-package *cellKey,* i.e. D=2 (the perturbation parameter for maximum noise), V=1.05 (the perturbation parameter for variance), js=1 (the threshold value for blocking small frequencies), pstay=NA (choosing the maximum entropy solution and not presetting probability of frequencies to remain unperturbed).

The following results are obtained for this setup: 3053 cells are perturbed while 2266 (or 42%) are not changed. The distribution of the noise can be described by the number (cnt) and/or the percentage (pct) of cells perturbed with a value of -2, -1, 0, 1, 2 as shown in Figure 3.

Figure 3. The noise distribution

	noise	cnt	pct
1:	-2	320	0.06016168
2:	-1	1194	0.22447829
3:	0	2266	0.42601993
4:	1	1222	0.22974243
5:	2	317	0.05959767

By calculating the Goodman Kruskal's Gamma statistic[7] one finds that the perturbation noise *does not follow any particular spatial pattern*, i.e. the hypothesis of no association between the small area and perturbation count variables cannot be rejected (gamma=-0.0014, with a 95% confidence interval between -0.025 and 0.022, including zero).

[7]https://www.rdocumentation.org/packages/DescTools/versions/0.99.37/topics/GoodmanKruskalGamma

The risk measure (d1, d2, d3) are described by their cell distributions (see Figure 3) and by their cumulative distributions (as in Figure 4).

Figure 4. Distributions of risk measures

	val	d1	d2	d3
1:	Min	0.000	0.000	0.000
2:	Q10	0.000	0.000	0.000
3:	Q20	0.000	0.000	0.000
4:	Q30	0.000	0.000	0.000
5:	Q40	1.000	0.001	0.014
6:	Mean	0.751	0.056	0.061
7:	Median	1.000	0.003	0.027
8:	Q60	1.000	0.005	0.036
9:	Q70	1.000	0.009	0.053
10:	Q80	1.000	0.028	0.091
11:	Q90	2.000	0.100	0.172
12:	Q95	2.000	0.333	0.318
13:	Q99	2.000	1.000	0.414
14:	Max	2.000	2.000	0.732

Figure 5. Cumulative distributions of risk measures

$cumdistr_d1

	cat	cnt	pct
1:	0	1788	0.3803446
2:	1	4084	0.8687513
3:	2	4701	1.0000000

$cumdistr_d2

	cat	cnt	pct
1:	[0,0.02]	3665	0.7796214
2:	(0.02,0.05]	3987	0.8481174
3:	(0.05,0.1]	4237	0.9012976
4:	(0.1,0.2]	4415	0.9391619
5:	(0.2,0.3]	4457	0.9480961
6:	(0.3,0.4]	4518	0.9610721
7:	(0.4,0.5]	4555	0.9689428
8:	(0.5,Inf]	4701	1.0000000

$cumdistr_d3

	cat	cnt	pct
1:	[0,0.02]	2078	0.4420336
2:	(0.02,0.05]	3235	0.6881515
3:	(0.05,0.1]	3838	0.8164220
4:	(0.1,0.2]	4304	0.9155499
5:	(0.2,0.3]	4440	0.9444799
6:	(0.3,0.4]	4537	0.9651138
7:	(0.4,0.5]	4681	0.9957456
8:	(0.5,Inf]	4701	1.0000000

Contrasting with cell suppression method

In order to show the extent of loss of information which would be the result of applying a method which does not change the micro or aggregate data, we have run the R-package *sdcTable*[8] which implements the cell suppression method by identifying sensitive cells of first and secondary (induced) orders and makes use of the *sdcHierarchies*[9] package as well. Given the education, gender, small areas dimension hierarchies (a total of 5319 cells), one would only be able to publish 4317 cells, while 552 primary sensitive cells and 450 additional cells would be suppressed.

Conclusions

The most critical stages in applying and evaluating an SDC method are: the identification of risk variables and the risk-utility analysis. As shown in [1], the former is a rather subjective process which is based on legal, cultural and information types of conditions. The latter is the object of an interesting statistical problem, i.e. evaluating the effect of multivariate transformations (as implicitly defined by both methods employed here) on multivariate data distributions. The analytical and empirical results in [7] are the most promising in defining unique measures for both risk and utility which in turn can be used to define the parameters of the optimum regime of the employed SDC method. As an implementation note and work in progress, one may mention that the tests described here could be extended by combining record swapping and cell-key perturbation methods, especially for the high risk cells and by further refining the parameter choice procedure.

[8] https://sdctools.github.io/sdcTable

[9] https://cran.r-project.org/web/packages/sdcHierarchies/index.html

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] E*U legislation on the 2021 population and housing censuses*, Printed by Publications Office of the European Union in Luxembourg, 2019 edition.
[3] Shlomo, N & Antal, L. & Elliot, M. *Measuring Disclosure Risk and Data Utility for Flexible Table Generators,* Journal of official statistics, 31, 2015.
[4] Shlomo, N., *Statistical Disclosure Control Methods for Census Frequency Tables*, International Statistical Review / Revue Internationale de Statistique, 75(2), 199-217, 2007.
[5] Leaver, V., *Implementing a method for automatically protecting user-defined Census tables*, Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain, 2009.
[6] Marley, J.K., Leaver, V. L., *A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis*, Int. Statistical Inst.: Proc. 58th World Statistical Congress, Dublin (Session IPS060), 2011.
[7] Shlomo, N., *Methods to assess and quantify disclosure risk and information lossunder statistical disclosure control*, A contributing article to the National Statistician's Quality Review into Privacy and Data Confidentiality Methods, Government Statistical Service, 2018.
[8] Antal, L., Enderle, T., Giessing, S., Harmonised protection of census data in the ESS, *Statistical disclosure control methods for harmonised protection of census data*, 2017.
[9] Willenborg, L. and de Waal, T., *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, 155 , Springer Verlag, New York, 2001.
[10] Duncan, G., Keller-McNulty, S., and Stokes, S. *Disclosure Risk vs. Data Utility: the R-U Confidentiality Map,* Technical Report LA-UR-01-6428, Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory, 2001.
[11] SDC UKCDMAC Subgroup Papers, *Statistical Disclosure Control (SDC) Methods Short-listed for 2011 UK Census Tabular Outputs*, NSO.
[12] Shlomo, N. and Young, C., *Statistical Disclosure Limitation Methods Through a Risk-Utility Framework,* In PSD'2006 Privacy in Statistical Databases, (Eds. J. Domingo-Ferrer and L. Franconi), Springer LNCS 4302, pp. 68-81, 2006.
[13] Antal, L., Shlomo, N., and Elliot, M., *Measuring Disclosure Risk with Entropy in Population Based Frequency Tables*, In PSD'2014 Privacy in Statistical Databases, (Eds. J. Domingo-Ferrer), Germany: Springer LNCS 8744, 62-78, 2014.
[14] Antal, L., Shlomo, N., and Elliot, M. (2015) Disclosure Risk Measurement with Entropy in Two-Dimensional Sample Based Frequency Tables. Joint UNECE/Eurostat work session on statistical data confidentiality, Helsinki, October 2015, https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_14_Session_1_-_Univ._Manchester.pdf

Hagtíðindi Greinargerðir
Statistical Series *Working papers*
Vol. 105. • No. 6. 2 September 2020
ISSN 1670-4770 Umsjón *Supervision* Violeta Calian • Violeta.Calian@Statice.is
www.statice.is www.hagstofa.is

Sími *Telephone* +(354) 528 1000
Bréfasími *Fax* +(354) 528 1099