

WP. 31
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (v): Statistical disclosure control methods for the next census round

**STATISTICAL DISCLOSURE CONTROL FOR
EUROPEAN CENSUS DISSEMINATION**

Supporting Paper

Prepared by Natalie Shlomo (University of Southampton)

Statistical Disclosure Control for European Census Dissemination

Natalie Shlomo*

* Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom, N.Shlomo @soton.ac.uk

1 Introduction

In December 2008, the task force 'EU Methodology for Census Data Disclosure Control' (CENSDC) was set up by Eurostat to address methodological issues in the statistical disclosure control (SDC) of Census tabular outputs. The members of the task force included representatives from NSIs of the Netherlands, Germany, Italy, Portugal, Estonia and the University of Southampton. The aim of the task force was to support EU regulations to formulate a unified dissemination program for Census outputs and in particular to provide users with high quality and comparable Census information. Each member state is required to prepare a set of pre-defined hypercubes containing Census counts: 19 hypercubes for the geography level of LAU2 and over 100 hypercubes for the geography level of NUTS2, cross-classified with as many as six other Census variables which can then be used as building blocks for an online flexible table generating package known as the Eurostat Census Hub Project. The objectives of the task force:

- Take into account national regulations regarding the confidentiality of Census data,
- Review national practices on data confidentiality for Census data,
- Consider the possibilities of application of further methods as available in the scientific literature,
- Identify a methodology for disclosure protection for Census data which could be applied by all countries subject to EU regulation, and that complies with the national regulations on Census data confidentiality,
- Recommend a tool (or provide advice for its development) for the practical implementation of an identified method, taking into account the ongoing developments in the Eurostat Census Hub Project and other dissemination requirements.

In this paper, we present an initial scoping study carried out by the CENSDC task force to assess SDC methods on the Census hypercubes and to provide recommendations to Eurostat and member states for harmonized methods. In addition, the task force assessed the protection afforded to tabular outputs that might be generated through the flexible table generating software through SDC rules and methods that can be applied 'online' on final outputted tables. The analysis is

based on a simulated hypercube according to a specific definition provided by Eurostat.

Section 2 describes the simulated hypercube and Section 3 some initial results for the SDC methods proposed for protecting the Census hypercubes. Section 4 discusses SDC rules and techniques that can be applied 'online' in the flexible table generating software thus improving the utility of the data by implementing SDC methods on the final outputs of the package and not on the building blocks as defined by the hypercubes. Section 5 concludes with a general discussion.

2 Data

To investigate SDC methods for hypercubes containing Census counts, a synthetic population was generated based on 1,500,000 individuals. The hypercube was defined according to Eurostat specifications with the following variables:

NUTS2 Region - two regions of size 845,539 and 645,461 individuals

Gender – 2 categories

Banded age groups – 21 categories

Current Activity Status – 5 categories

Occupation – 13 categories

Educational attainment – 9 categories

Country of citizenship – 5 categories

The total number of cells in the hypercube was 245,700. The cell proportions were obtained from Census tables derived from the 2001 United Kingdom Census. The average cell size for this hypercube is 6.1. However, the distribution of cell counts is quite skewed with a large proportion of zero cells as seen in Table 1.

Table 1: Distribution of Cell Counts in the Synthetic Hyper-cube

Cell Value	Number of Cells	Percentage of Cells
0	226,939	92.36%
1	4,028	1.64%
2	2,112	0.86%
3-5	2,964	1.21%
6-8	1,664	0.68%
9-10	720	0.29%
11	7,273	2.96%
Total	245,700	100.00%

The synthetic hypercube was comparable to real hypercubes that were generated according to the above specification produced by member countries: Italy and Estonia at the NUTS2 region level and had a similar distribution of cell counts.

3 Statistical Disclosure Control Methods

In this section, we focus on methods for protecting the hypercube: one pre-tabular method based on record swapping and two post-tabular methods based on a semi-controlled random rounding and a probabilistic perturbation mechanism. In addition, the task force also examined the option of cell suppression using Tau-Argus but due to the large size of the hypercubes and the need to consistently suppress cells across hypercubes, this option was not a feasible method.

3.1 Pre-tabular Method

The most common pre-tabular method of SDC for Census frequency tables is record swapping on the microdata prior to tabulation where variables are exchanged between pairs of households. In order to minimize bias, pairs of households are typically determined within strata defined by control variables, such as a large geographical area, household size and the age-sex distribution of the individuals in the households. In addition, record swapping can be targeted to high-risk households found in small cells of Census tables thereby ensuring that households that are most at risk for disclosure are likely to be swapped. For more information on record swapping, see Dalenius and Reiss, 1982, Fienberg and McIntyre, 2005 and Shlomo, 2007.

In a Census context, geography variables are often swapped between households for the following reasons:

- Given household characteristics, other Census variables are likely to be independent of geography and therefore it is assumed that less bias will occur. In addition, because of the conditional independence assumption, swapping geography will not necessarily result in inconsistent and illogical records. By contrast, swapping a variable such as age would result in many inconsistencies with other variables, such as marital status and education level.
- At a higher geographical level and within control strata, the marginal distributions are preserved.
- The level of protection increases by swapping variables which are highly “matchable” such as geography.
- There is some protection for disclosure risk from differencing two tables with nested geographies since record swapping introduces ambiguity into the true cell counts. This is true for other variables, for example nested age bands.

For this study, we had to carry out the random record swapping at the individual level since Census microdata was unavailable. In addition, to keep the study simple, a random sample of 5% of the individuals were selected in each NUTS2 region. The selected individuals were paired randomly with other individuals within different geographies at the LAU2 level, and the LAU2 geographies swapped between them. Therefore, a total of 10% of the individuals in each NUTS2 region had their LAU2 geography variable swapped.

3.2 Post-tabular Methods

3.2.1 Random Rounding

The most common post-tabular method of SDC for Census frequency tables is based on unbiased random rounding. The entries of the table x are first converted to residuals of the rounding base b , $res(x)$. Let $Floor(x)$ be the largest multiple k of the base b such that $bk < x$ for an entry x . In this case, $res(x) = x - Floor(x)$. For an unbiased rounding procedure, x is rounded up to $(Floor(x) + b)$ with probability $\frac{res(x)}{b}$ and rounded down to $Floor(x)$ with probability $(1 - \frac{res(x)}{b})$. If x is already a multiple of b , it remains unchanged.

In general, each small cell is rounded independently in the table, i.e. a random uniform number u between 0 and 1 is generated for each cell. If $u < \frac{res(x)}{b}$ then the entry is rounded up, otherwise it is rounded down. This ensures an unbiased rounding scheme and the expectation of the rounding perturbation is zero and no bias should remain in the table. However, the realization of this stochastic process on a finite number of cells in a table may lead to overall bias since the sum of the perturbations (i.e. the difference between the original and rounded cell) going down may not equal the sum of the perturbations going up. Because of the large number of perturbations in the table, margins are typically rounded separately from internal cells and therefore tables are not additive.

To place some controls in the random rounding procedure, the following algorithm can be used for selecting the entries to round up or down: First the expected number of entries of a given $res(x)$ that are to be rounded up is predetermined (for the entire table or for each row/ column of the table). Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This process ensures a bias of zero and the rounded internal cells aggregate to the controlled rounded total.

Another problem with random rounding is the consistency of the rounding across same cells that are aggregated in different tables. The consistency can be solved by

the use of microdata keys. For each record in the microdata, a random number (i.e., a key) is defined which when combined with other records to form a cell of a table defines the seed for the rounding. Records that are aggregated into same cells will always have the same seed and therefore a consistent rounding (see Shlomo and Young, 2008).

For this analysis, we carry out full random rounding to base 3 semi-controlled to the two NUTS2 totals in the hypercube.

3.2.2 Stochastic Perturbation

A more general method to rounding can be carried out by perturbing the internal cells of the hypercube using a probability mechanism based on a probability transition matrix similar to the method that is used in PRAM (Gouweleeuw, Kooiman, Willenborg, and De Wolf, 1998). Let \mathbf{P} be a $L \times L$ transition matrix containing conditional probabilities:

$$p_{ij} = p(\text{perturbed cell value is } j \mid \text{original cell value is } i)$$

for cell values from 0 to L (usually a cap is put on the cell values and any cell value above the cap would have the same perturbation probabilities). Let \mathbf{t} be the vector of frequencies of the cell values where the last component would contain the number of cells above cap L and \mathbf{v} the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/K$, where K is the number of cells in the table. In each cell of the table, the cell value is changed or not changed according to the prescribed transition probabilities in the matrix \mathbf{P} and the result of a draw of a random multinomial variate u with parameters q_{ij} ($j=1, \dots, L$). If the j -th value is selected, value i is moved to value j . When $i = j$, no change occurs.

Placing the condition of invariance on the transition matrix \mathbf{P} , i.e. $\mathbf{tP} = \mathbf{t}$, means that the marginal distribution of the cell values are approximately preserved under the perturbation and we ensure a zero bias in the overall total. As described in the random rounding procedure, in order to obtain the exact overall total, a “without” replacement strategy for selecting the cell values to change can be carried out. For each particular cell value, we calculate the expected number of cells that need to be changed to another value according to the probabilities in the transition matrix. We then randomly select (without replacement) the cell values and change their values.

To preserve exact additivity in the table, an IPF algorithm can be used to fit the margins of the table after the perturbation. This results in cell values that are not integers. Exact additivity with integer counts can be achieved by controlled rounding to base 1 using for example Tau-Argus (see Salazar-Gonzalez, Bycroft, and Staggemeier, 2005). Cell values can also be rounded to their nearest integers resulting in ‘close’ additivity because of the constraints on the marginal distribution of the cell counts due to the invariance property of the transition matrix. Finally,

the use of microdata keys can ensure the consistent perturbation of cells across hypercubes (see Shlomo and Young, 2008).

For this particular study, we carry out a stochastic perturbation based on invariant PRAM with controls in the overall totals of the two NUTS2 regions. We carry out the perturbation on cells of values 1,2,3...10 only. All large cells above a value of 11 were not perturbed.

4 Results

4.1 Disclosure risk

A measure to quantify disclosure risk in frequency tables is the number of small cells of size 1 and 2 that are not changed by the SDC method. As can be seen in Table 1, there were a total of 6,140 small cells in the hypercube (2.5%). The stochastic perturbation changed 46.6% of the small cells, the random rounding to base 3 changed 100% of the small cells and the random record swapping changed only 16.2% of the small cells.

Another measure that can be used to quantify disclosure risk is based on the entropy. The entropy obtains a minimum value of zero if all cells are zero except for one cell with a 100% count, and a maximum value if all cells have an equal value. Therefore, entropy measures the degree to which attribute disclosure might be a problem due to the placement of zeros in rows/columns of a table.

For each sub- group z defined by NUTS2*Gender*Banded age groups, the following table of counts was calculated: Current Activity Status * Occupation *Educational Attainment* Country of citizenship. Let $D_z(c)$ represent the table of counts for a particular cell c and sub-group z . The cell probability for cell c in sub-group z is defined as: $p_z(c) = D_z(c) / \sum_c D_z(c)$ and the entropy: $-\sum_c p_z(c) \log(p_z(c))$. The

measure is the median of the entropy across all sub-groups z . The original table and the tables resulting from the post-tabular methods of random rounding and stochastic perturbation all had a median entropy of 2.36. The median entropy however for record swapping was 2.34 showing the slight ‘smoothing’ of the cell counts in the hypercube.

4.2 Information Loss

Information loss can be measured by comparing perturbed cell values and original cell values using a distance metric. Let the table of interest be denoted D and $D(c)$ the cell count in table D of cell c . We define the relative absolute distance: $RAD = |D(c_{pert}) - D(c_{orig})| / D(c_{orig})$ across the non-zero cells of the original table.

Table 2 contains the *RAD* for marginal tables and some bivariate tables of the hypercube.

Table 2: Distance Metrics on Marginal and Bivariate Tables in the Hypercube

Table	Stochastic Perturbation	Random Rounding	Nuts2 Swapping
Marginal Tables			
Nuts2	0	0	0
Gender	0	0	0
Age group	0.064	0.029	0
Activity Status	0.002	0.002	0
Occupation	0.067	0.017	0
Education	0.006	0.001	0
Country of Citizenship	0.956	0	0
Bivariate Tables			
Gender*Nuts2	0.001	0.001	0
Age group*Nuts2	0.116	0.085	0.388
Activity Status*Nuts2	0.004	0.006	0.045
Occupation*Nuts2	0.382	0.159	0.717
Education*Nuts2	0.020	0.011	0.418
Country of Citizenship*Nuts2	1.874	0.440	0.895
Gender*Age group	0.164	0.088	0
Activity Status*Age group	0.343	0.196	0
Occupation*Age group	8.022	6.618	0
Education*Age group	4.029	1.884	0
Country of Citizenship*Age group	18.257	16.360	0

With respect to the post-tabular methods, the random rounding has lower distance metrics compared to the stochastic perturbation and provides the most protection on small cells. However, other things to consider in post-tabular methods is the lack of overall additivity and consistency of cells across hypercubes which may lead to

‘unpicking’ the protected cell values. As described in Section 3, new algorithms for carrying out post-tabular perturbation methods are currently under research and development and may alleviate some of the problems of additivity and consistency across tables. The stochastic perturbation is flexible since the data protector defines the specifications of the probabilities in the transition matrix and which cell values to perturb. It is more difficult to ‘unpick’ this method. As seen in this application of the stochastic perturbation, some original small cells remained unperturbed in the hypercube though further tweaking of the probabilities in the transition matrix may reduce this number.

Record swapping had the highest disclosure risk but marginal tables and bivariate tables not involving the swapping variable have no differences at the higher NUTS2 geography since it was used as a control variable in the swapping. However, the bivariate tables that involve the swapping variable have distorted joint distributions as seen by the high distance metrics across cross- classifications of NUTS2 with the other variables of the hypercube. These distance metrics are higher than the distance metrics obtained from the post-tabular methods.

Another measure of information loss is the impact on the Cramer’s V statistic for the bivariate tables defined in Table 2 based on cross- classifying the NUTS2 variable with each of the other variables of the hypercube. The post tabular methods did not show any reduction in the Cramer’s V statistic probably due to the high volume of zeros in the table even before the perturbation. However, there was a consistent reduction of about 11% on the Cramer’s V statistic under the method of record swapping for the bivariate tables. This result is consistent with findings that record swapping ‘smooths’ out the counts in the tables and therefore tends towards a model of independence (see Shlomo, 2007).

5 Online SDC in the Flexible Table Generating Package

Since member states typically have different methods and different standards for protecting Census hypercubes, the task force addressed the question of whether SDC methods can be carried out ‘on the fly’ through the flexible table generating package itself. This would increase the utility of the generated user-defined Census tables since the SDC methods would only be applied on the final outputted table and not on the building blocks that generate the tables since it is well known that aggregating perturbed building blocks exacerbates the impact of the SDC methods.

Some ad-hoc SDC rules can easily be applied in the software package, such as:

- Limit the number of dimensions in the tables,
- Ensure consistent and nested categories of variables to avoid disclosure by differencing,

- Ensure minimum population thresholds,
- Ensure that the percentage of small cells is above a minimum threshold,
- Ensure average cell size above a minimum threshold.

In spite of the ad hoc SDC rules above, it is likely that some small cells may still remain in the generated table. As widely discussed in the Computer Scientist literature (see, for example, Dinur and Nissim, 2003), the only way to guarantee the confidentiality of respondents under flexible query systems is by adding noise. Therefore, one can apply ‘on the fly’ semi controlled random rounding or the stochastic perturbation methods as described in Section 3.2.

As an example, assume a scenario where we limit the number of dimensions for outputted tables in a flexible table generating software to include a geography and three other Census variables. In this example, for NUTS2 region equal to 1, we define a table as: Banded age group*Education*Occupation. This table contains 2,457 cells with 854,539 individuals, giving an average cell size of 347.8 individuals. The table however produces a very skewed distribution of cell counts as seen in Table 3. .

Table 3: Distribution of Cell Counts in the Generated Table Banded Age Group*Education*Occupation for NUTS2=1

Cell Value	Number of Cells	Percentage of Cells
0	1534	62.43%
1	44	1.79%
2	35	1.42%
3	27	1.10%
4	20	0.81%
5+	797	32.44%
Total	2457	100.00%

There are over 3.2% of small cells in the generated table although this number can be reduced by employing some of the other ad hoc SDC rules mentioned above. We can apply a post-tabular method of SDC such as semi-controlled random rounding to base 3 or we may even consider implementing the semi-controlled random rounding to base 3 on the small cells only of the table. New algorithms to carry out the random rounding as described in Section 3.2.1 would ensure consistency of cells across generated tables and 'closeness' to additivity. Another option, given the small dimensions of the table that can be generated in the software package is to use the controlled rounding feature in Tau-Argus(see Salazar-Gonzalez, Bycroft, and Staggemeier, 2005). All of these methods can be implemented ‘on the fly’ before outputting the table to the user.

6. Discussion

In this paper, we described an initial scoping study by the CENSDC task force to assess applications of SDC on pre-defined hypercubes containing Census counts with the aim of providing recommendations to member states for a uniform and valid SDC method for protecting hypercubes. These protected hypercubes could then be used in the flexible table generating package of the Census Hub Project and would allow users to tailor and generate their own tables from among all member states. Based on the scoping study, it was clear that the hypercubes as defined by Eurostat were too large to handle most SDC methods and also had very skewed distributions of cell counts. The recommendation of the task force at this stage was for Eurostat to reduce the size of the hypercubes required by member states.

Applying ad hoc SDC rules and a post-tabular method within the flexible table generating package would relieve the NSIs of member states of having to protect the hypercubes. Indeed, the Census Hub Project can be developed in such a way that the hypercubes never have to physically leave the NSI, rather the information is accessed remotely according to the definitions provided by the user for their table of interest. To rely solely on 'online' SDC methods for protecting the generated tables means that some form of random noise, i.e. stochastic rounding or perturbation, needs to be applied to the final outputted table. This would improve the quality of the outputted table since it does not exacerbate the SDC methods arising from aggregating perturbed building blocks. Further research needs to be directed to improving the stochastic post-tabular methods to improve additivity and consistency of the tables.

References:

- Dalenius, T. and Reiss, S.P. (1982) Data Swapping: A Technique for Disclosure Control. *Journal of Statistical Planning and Inference*, 7, 73-85.
- Dinur, I. and Nissim, K. (2003) Revealing Information While Preserving Privacy. *PODS 2003*, pp. 202-210.
- Fienberg, S.E. and McIntyre, J. (2005) Data Swapping: Variations on a Theme by Dalenius and Reiss. *Journal of Official Statistics*, 9, 383-406.
- Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998) Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, 14, 463-478.
- Salazar-Gonzalez, J.J., Bycroft, C. and Staggemeier, A.T. (2005) Controlled Rounding Implementation. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva.

Shlomo, N. (2007) Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review*, Vol. 75, Number 2, pp. 199-217.

Shlomo, N. and Young, C. (2008) Invariant Post-tabular Protection of Census Frequency Counts. In *PSD'2008 Privacy in Statistical Databases*, (Eds. J. Domingo-Ferrer and Y. Saygin), Springer LNCS 5261, pp. 77-89.