DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 305–324, http://dx.doi.org/10.1515/JOS-2015-0019

Measuring Disclosure Risk and Data Utility for Flexible Table Generators

Natalie Shlomo[1], *Laszlo Antal*[1], *and Mark Elliot*[1]

Statistical agencies are making increased use of the internet to disseminate census tabular outputs through web-based flexible table-generating servers that allow users to define and generate their own tables. The key questions in the development of these servers are: (1) what data should be used to generate the tables, and (2) what statistical disclosure control (SDC) method should be applied. To generate flexible tables, the server has to be able to measure the disclosure risk in the final output table, apply the SDC method and then iteratively reassess the disclosure risk. SDC methods may be applied either to the underlying data used to generate the tables and/or to the final output table that is generated from original data. Besides assessing disclosure risk, the server should provide a measure of data utility by comparing the perturbed table to the original table. In this article, we examine aspects of the design and development of a flexible table-generating server for census tables and demonstrate a disclosure risk-data utility analysis for comparing SDC methods. We propose measures for disclosure risk and data utility that are based on information theory.

Key words: Statistical disclosure control; census tabular data; entropy; Hellinger distance.

1. Introduction

Driven by demand from policy makers and researchers for specialized and tailored census frequency tables, many statistical agencies are considering the development of a web-based software platform where users can generate tables of interest from underlying census microdata through a user-friendly interface. This platform is called a "flexible table-generating server". Users access the server via the internet and generate their preferred set of tables from predefined variables or categories using drop-down lists. These tables can then be downloaded to the personal computers of the users. The United States Census Bureau and the Australian Bureau of Statistics have developed such servers on their websites to disseminate census frequency tables.

When generating flexible tables, the server should be able to provide a measure of disclosure risk for the original table, apply a statistical disclosure control (SDC) method and then reassess disclosure risk and the impact on data utility following the SDC method. These steps must be carried out "on the fly" within the server for each generated output table. SDC is a set of statistical practices which aim to ensure that no individual population

[1] University of Manchester, Social Statistics, Humanities Bridgeford Street, Manchester M13 9PL, United Kingdom. Emails: natalie.shlomo@manchester.ac.uk, laszlo.antal@postgrad.manchester.ac.uk, and mark.elliot@manchester.ac.uk

Acknowledgments: The project is funded by the EU 7th framework infrastructure research grant: 262608, Data Without Boundaries (DwB) and the ONS-ESRC funded PhD studentship (Ref. ES/J500161/1).

unit can be reidentified from anonymised data nor any new information learnt about any specific individual (with certainty). SDC is an active research area. For reviews of this area, see Willenborg and de Waal (2001), Doyle et al. (2001), Duncan et al. (2011) and Hundepool et al. (2012).

There are two main types of disclosure risks in census frequency tables: identity disclosure, where small cell counts may lead to the identification of an individual in the population, and attribute disclosure, where new information may be learnt about an individual or group of individuals. Attribute disclosure in frequency tables occurs when rows or columns of a table contain (real) zeroes and only one or two cells are nonzero. This enables an "intruder" to first make an identification based on a margin total and subsequently reveal new information according to other variables spanning the table. Another type of disclosure risk that needs to be guarded against is disclosure by differencing. The differencing of tables generated through the server can lead to residual tables that are more susceptible to the above disclosure risks and even to the reconstruction of individual records. This is typically dealt with by applying perturbative methods of SDC, which raises the level of uncertainty of true counts in the tables and hence of the difference between counts across tables. After the table is protected, a data utility measure must also be calculated by comparing the perturbed table to the original table.

The need to measure disclosure risk "on the fly" for census frequency tables produced via a flexible table-generating server motivated the research and development of a new global disclosure risk measure. Until now, disclosure risk measures for tabular data have been defined at the cell level and not for the entire table. We propose a new disclosure risk measure based on information theory as shown in Antal et al. (2014) and also relate this theory to a data utility measure.

The key issues when developing a web-based flexible table generating server addressed in this article are: (1) what underlying data should be used in the background for generating the output tables, and (2) at what stage should the SDC method be applied. In addition, the article provides a comparison study of some common SDC methods which may be used to protect census tables within a flexible table-generating server and demonstrates how statistical agencies should undertake a disclosure risk-data utility analysis to inform decisions about SDC methods and their parameterization. In general, SDC methods employed by statistical agencies are often motivated by country-specific agendas and policy sensitivities and it is difficult to develop a universal best practice. However, one important distinction when considering SDC methods for flexible table-generating servers is that the outputs are defined by users and the amount of disclosure risk may vary in each output.

Section 2 presents aspects to consider in the design of a flexible-table generating server, including the underlying data for generating output tables and the stage when SDC methods may be applied. In Section 3, some common SDC methods for census frequency tables are described. Section 4 introduces a new global disclosure risk measure based on information theory and a related data utility measure that can be calculated "on the fly" for each output table generated in the server. In Section 5, a comparison study is carried out on generated census output tables from a flexible table-generating server. The comparison study will be informed by a disclosure risk-data utility analysis on the generated tables perturbed by the SDC methods described in Section 3 based on the

measures outlined in Section 4. A discussion and concluding remarks are presented in Section 6.

2. Designing a Flexible Table-Generating Server

In this section, we describe the design of an online flexible table-generating server and discuss the following issues: the underlying data that may be used as input to the server, the stage at which SDC methods can be applied, and preliminary SDC rules to determine *a priori* whether the requested table can be generated or not.

2.1. Underlying Input Data to the Server

The underlying data to use as input for a web-based flexible table-generating server can be based on the original microdata or disclosure-controlled microdata. The input data is largely determined by the source and content of the data as well as the SDC method that will be applied to the final output tables (if any). Microdata arising from social surveys with small sampling fractions have a lower disclosure risk than microdata arising from censuses containing whole population counts, and therefore are more appropriate for use in their original form. Output tables generated from survey microdata where only weighted counts are released are generally considered to be of low disclosure risk with no further need for an application of SDC methods. Census (and administrative data) containing whole populations and particularly those containing sensitive data, such as health statistics or business microdata, are more problematic. In microdata containing the whole population, individuals (or businesses) can easily be identified leading to the disclosure of attributes. In this case, the underlying input data should be protected prior to the generation of tables.

For a flexible table-generating server of census tables, one method for producing the underlying input data is to aggregate the microdata into a very large multi-dimensional frequency table, called a hypercube, where no data of individuals can be disseminated below the level of a cell value in the hypercube. For example, users may only be able to disseminate frequency counts of age in 5-year age bands and not counts for single years. This approach was taken by Eurostat for the dissemination of census tables from European Member States. A flexible table-generating server for European census tables is being developed through the European Census Hub Project. Each Member State is required to produce a set of predefined hypercubes containing their country's census counts: 19 hypercubes at the geography level of LAU2 and over 100 hypercubes at the geography level of NUTS2, cross-classified with as many as six other census variables in each hypercube. NUTS2 is a European subregional geography and LAU2 are small municipalities or equivalent. Researchers are able to use the considerable number of multidimensional hypercubes and their wealth of census data made available through the European Census Hub to generate tables of interest beyond what would have been available previously using standard table-extraction software. The flexible table-generating server will allow comparative tables across Member States and the combining of census data from multiple Member States. The hypercubes have the additional advantage that they provide some limited protection against disclosure risk since no data below the level of the cell values of the hypercube can be disseminated.

However, the hypercubes themselves still have considerable disclosure risk since they are very large and sparse with many zero and small cell counts. Therefore, there will still be the need to apply an SDC method to protect output tables generated from the flexible table-generating server.

2.2. *Application of SDC Methods*

SDC methods for protecting output tables generated from a flexible table-generating server can be applied either on the underlying input data so that all tables generated are deemed safe for dissemination (the pretabular SDC approach), or applied directly to the final output table generated from the original data (the post-tabular SDC approach) or a combination of both. Although sometimes neater and less resource intensive when data is from a single source, the pretabular SDC approach is problematic for the dissemination of European Census data for two reasons. Firstly, all Member States would have to agree on a common SDC method in order to provide consistent hypercubes across all Member States. For example, if one Member State employs a rounding method whilst another Member State employs cell suppression, there will be significant quality issues in a table that is generated based on both Member States' data. Secondly, when aggregating data which have been separately disclosure controlled, the effects of the SDC methods are compounded and the data may be overprotected. For example, aggregating cells that have already been rounded not only overprotects the data but also exacerbates the data utility impact by providing counts that are no longer rounded to the nearest base. With the second approach of protecting only the final tabular output, SDC methods are not compounded in this way. We investigate the pretabular and post-tabular approaches in the comparison study presented in Section 5.

2.3. *Preliminary SDC Rules*

The design of a web-based flexible table-generating server typically involves many *ad hoc* preliminary SDC rules which determine *a priori* if generated tables can be released or not. These SDC rules may include:

- Limiting the number of dimensions in the output tables.
- Ensuring consistent and nested categories of variables to avoid disclosure by differencing.
- Ensuring minimum population thresholds.
- Ensuring that the percentage of small cells is below a maximum threshold.
- Ensuring average cell size above a minimum threshold.

The steps in a flexible table-generating server are:

(1) Determine whether the table can be released according to the preliminary SDC rules.
(2) Calculate a disclosure risk measure to determine if an SDC method should be applied to the final output table.
(3) Apply the SDC method.

(4) Recalculate the disclosure risk measure to determine if the table is safe to generate; if yes proceed to Step 5, otherwise do not release the table.
(5) Output the final table with a measure of data utility.

According to the steps of a flexible table-generating server, it is clear that analytical expressions of disclosure risk and data utility that can be calculated "on the fly" within the server are necessary.

3. Statistical Disclosure Control Methods

In this section, we describe some common SDC methods which have been used to protect census frequency tables: a pretabular SDC method of record swapping is used in the United States and the United Kingdom, a post-tabular method of random rounding is used in New Zealand and Canada, and a post-tabular probabilistic perturbation mechanism has recently been implemented in Australia.

3.1. Record Swapping

Record swapping is based on the exchange of values of variable(s) between similar pairs of population units (often households). In order to minimize bias, pairs of population units are determined within strata defined by control variables. For example, when swapping households, control variables may include: a large geographical area, household size, and the age-sex distribution of individuals in the households. In addition, record swapping can be targeted to high-risk population units found in small cells of census tables. In a census context, geographical variables related to place of residence are often swapped. Swapping place of residence has the following properties: (1) it minimizes bias based on the assumption that place of residence is independent of other census target variables conditional on the control variables; (2) it provides more protection for census tables since place of residence is a highly visible variable which can be used to identify individuals; (3) it preserves marginal distributions within a larger geographical area. For more information on record swapping, see Dalenius and Reiss (1982), Fienberg and McIntyre (2005), and Shlomo (2007).

3.2. Semi-Controlled Random Rounding

A post-tabular method of SDC for census frequency tables is unbiased random rounding. Let *Floor(x)* be the largest multiple bk of the base b such that $bk < x$ for any value of x. In this case, $res(x) = x\text{-}Floor(x)$. For an unbiased rounding procedure, x is rounded up to $Floor(x) + b$ with probability $res(x)/b$ and rounded down to $Floor(x)$ with probability $(1 - (res(x)/b))$. If x is already a multiple of b, it remains unchanged.

In general, each cell is rounded independently in the table, that is, a random uniform number u between 0 and 1 is generated for each cell. If $u \leq (res(x)/b)$ then the entry is rounded up, otherwise it is rounded down. This ensures an unbiased rounding scheme, that is, the expectation of the rounding perturbation is zero. However, the realization of this stochastic process on a finite number of cells in a table will not ensure that the sum of the perturbations will exactly equal zero. To place some control in the random rounding procedure, we use a semi-controlled random rounding algorithm for selecting entries to round up or down as follows: 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). The expected number is rounded to the nearest integer. Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This procedure ensures that rounded internal cells aggregate to the controlled rounded total.

Due to the large number of perturbations under random rounding, margins are typically rounded separately from internal cells and tables are not additive. When using semicontrolled random rounding this alleviates some of the problems of nonadditivity since one of the margins and the overall total will be preserved. Another problem with random rounding is the consistency of the rounding across same cells that are generated in different tables. It is important to ensure that the cell value is rounded consistently, otherwise the true cell count can be learnt by generating many tables containing the same cell and observing the perturbation patterns. Fraser and Wooton (2005) propose the use of *microdata keys* which can solve the consistency problem. First, a random number (which they call a key) is defined for each record in the microdata. When building a census frequency table, records in the microdata are combined to form a cell defined by the spanning variables of the table. When these records are combined to a cell, their keys are also aggregated. This aggregated key serves as the seed for the rounding and therefore same cells will always have the same seed and result in consistent rounding.

Further research is needed to ensure both the additivity and consistency properties for random rounding. For simple tables of the type that would be generated in a flexible table-generating server, controlled rounding algorithms can be applied to ensure additivity on remaining totals without distorting the unbiasedness of the rounding (see Willenborg and De Waal 2001).

3.3. *Stochastic Perturbation*

A more general method than random rounding is stochastic perturbation, which involves perturbing the internal cells of a table using a probability transition matrix and is similar to the postrandomisation method that is used to perturb categorical variables in microdata (see Gouweleeuw et al. 1998). In this case, it is the cell counts in a table that are perturbed. More details can be found in Fraser and Wooton (2005) and Shlomo and Young (2008).

Let $\mathbf{P}$ be a $(L+1)\times(L+1)$ transition matrix containing conditional probabilities: $p_{ij} = P(\textit{perturbed cell value is } j \,|\, \textit{original cell value is } i)$ for cell values from 0 to L, where L is a cap on the cell values and any cell value above the cap will 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 let $\mathbf{v}$ be the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/\mathrm{K}$ where K is the number of cells in the table. In each cell of the table, the cell value i is changed or not changed according to the prescribed transition probabilities in matrix $\mathbf{P}$ and the result of a draw of a random multinomial variate u with parameters $p_{ij} j = 0, 1, \ldots, L$. If the jth value is selected, value i is moved to value j. When $i = j$, no change occurs.

Placing the condition of invariance on the probability transition matrix $\mathbf{P}$ (i.e., $\mathbf{tP = t}$) means that the marginal distribution of the cell values are approximately preserved under

the perturbation. As described in the random rounding procedure in Subsection 3.2, in order to obtain the exact marginal distribution a similar strategy for selecting cell values to change can be carried out. For each cell value *i,* the expected number of cells that need to be changed to a different value *j* is calculated according to the probabilities in the transition matrix. We then randomly select (without replacement) the expected number of cells *i* and carry out the change to *j*.

To preserve exact additivity in the table, an iterative proportional fitting algorithm can be used to fit the margins of the table after the perturbation according to the original margins. This results in cell values that are not integers. Exact additivity with integer counts can be achieved for simple tables by controlled rounding to base 1 using Tau-Argus, for example (Salazar-Gonzalez et al. 2005). Cell values can also be rounded to their nearest integers resulting in "close" additivity because of the invariance property of the transition matrix. Finally, the use of microdata keys as described in Subsection 3.2 can also be adapted to this SDC method to ensure the consistent perturbation of same cells across different tables by fixing the seed for the perturbation.

4. Information Theory-Based Disclosure Risk and Data Utility Measures

For each output table generated, the flexible table-generating server must provide analytical expressions of disclosure risk and data utility that can be calculated "on the fly" within the server. As mentioned in Section 1, one of the major causes of disclosure risk in census frequency tables is attribute disclosure caused by rows/columns that have many zero cells and only one or two populated cells. A row/column with a uniform distribution of cell counts would have little attribute disclosure risk, whilst a degenerate distribution of cell counts would have high attribute disclosure risk. Moreover, a row/column with large counts would have less risk of reidentification compared to a row/column with small counts.

There is no single global-level disclosure risk measure for census frequency tables that measures attribute disclosure and identity disclosure. In planning for the 2011 UK Census, the Office for National Statistics assessed attribute disclosure by producing many census tables and calculating the proportion of those columns/rows where only one or two cells were populated and the rest of the cells were zero. They also provided a measure based on the proportion of small cells across the tables. These measures do not provide an accurate quantification of the disclosure risk for a specific table. To obtain an analytical expression of disclosure risk for the entire table (or row/columns), it is natural to use information theory, specifically the entropy.

4.1. An Information Theory Disclosure Risk Measure

As described in Antal et al. (2014), a disclosure risk measure for a census frequency table should have the following properties: (a) small cell values have higher disclosure risk than large values; (b) uniformly distributed frequencies imply low disclosure risk; (c) the more zero cells in the census table, the higher the disclosure risk; (d) the risk measure should be bounded by 0 and 1. Using information theory, we develop an analytical expression of disclosure risk that meets these properties.

Information theory is covered comprehensively in Cover and Thomas (2006). One of the most important measures is the entropy. Let X be a discrete random variable having a

distribution $P = (p_1, p_2, \ldots, p_K)$. The entropy is defined as:

$$H(X) = H(P) = -\sum_{i=1}^{K} p_i \cdot \log p_i$$

If $p_i = 0$ for a category i, the respective term in the sum will be considered 0, since $\lim_{x \to 0} x \log x = 0$. It follows that $H(P) \geq 0$, since $-p_i \cdot \log p_i \geq 0$ with $H(P) = 0$ *iff* the probability mass is concentrated on one point. Therefore, the smaller the entropy *H(P)*, the more likely that attribute disclosure can occur. Under the uniform distribution $U_K = ((1/K), (1/K), \ldots, (1/K))$, we obtain the maximum entropy: $H(U_K) = \log K$ and minimum attribute disclosure risk.

The entropy of the frequency vector in a table of size K, $F = (F_1, F_2, \ldots, F_K)$ where $\sum_{i=1}^{K} F_i = N$ is:

$$H(P) = H\left(\frac{F}{N}\right) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N} \tag{1}$$

To produce a disclosure risk measure between 0 and 1, we define the risk measure as:

$$1 - \frac{H\left(\frac{F}{N}\right)}{\log K}. \tag{2}$$

The disclosure risk measure in (2) ensures property (b) since the term will tend to zero as the frequency distribution is more uniform, and ensures property (d) since the measure is bounded between 0 and 1. However, the disclosure risk measure does not take into account the magnitude of the cells counts or the number of zero cells in the table (or row/column of the table) and does not preserve properties (a) and (c). Therefore, an extended disclosure risk measure is proposed in (3) and is defined as a weighted average of three different terms, each term being a measure between 0 and 1.

$$\begin{aligned} R(F, w_1, w_2) = {} & w_1 \cdot \left[\frac{|A|}{K}\right] + w_2 \cdot \left[1 - \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K}\right] \\ & - (1 - w_1 - w_2) \cdot \left[\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e\sqrt{N}}\right] \end{aligned} \tag{3}$$

where A is the set of zeroes in the table and $|A|$ the number of zeros in the set, K, N and F as defined above and w_1, w_2 are arbitrary weights: $0 \leq w_1 + w_2 \leq 1$.

The first measure in (3) is the proportion of zeros which is relevant for attribute disclosure and property (c). The third measure in (3) allows us to differentiate between tables with different magnitudes and accounts for property (a). As the population size N gets larger in the table, the third measure tends to zero. The weights w_1 and w_2 should be chosen depending on the data protector's choice of how important each of the terms are in contributing to disclosure risk. Alternatively, one can avoid weights altogether by taking the L_2- norm (see Subsection 4.3) of the three terms of the risk measure in (3) as follows: $\left(\left(\left(\sum_{i=1}^{3} |x_i|^2\right)^{1/2}\right) \Big/ \sqrt{3}\right)$ where x_i represents term i, $i = 1, 2, 3$ in (3).

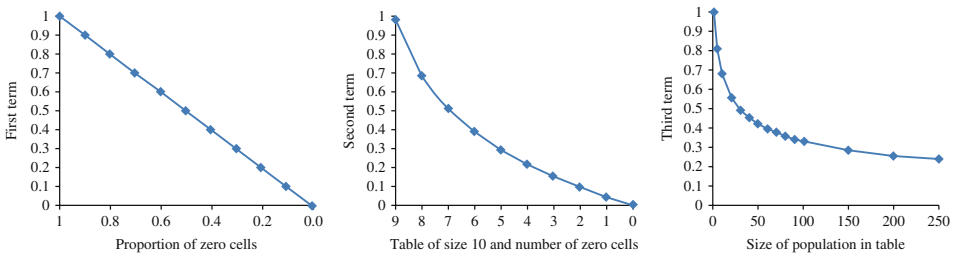

Fig. 1. The three components of the proposed disclosure risk measure in (3)

Figure 1 provides a graphical interpretation of each of the three terms of the proposed disclosure risk measure in (3). The figure on the left shows the first term of the disclosure risk measure as a function of the proportion of zero cells (although a table of all zeros would not be permissible in a flexible table-generating server). The figure in the middle shows the second term based on the entropy in (2) where we demonstrate with a table of ten cells and move from a uniform distribution to a degenerate distribution by accumulating zero cells and spreading the total to the remaining cells. The figure on the right shows the third term of the disclosure risk measure as the size of the population of the table increases.

The final disclosure risk measure (3) is an analytical expression and can be calculated "on the fly" in the flexible table-generating server without the need to see the generated table beforehand. In order to emphasize the risk of identity disclosure arising from small counts (ones and twos), we split the entropy measure as shown in (2) into two parts, small counts up to six and larger counts of seven and more, and provide different weights for each part. For the comparison study in Section 5, the following weights were chosen: $w_1 = 0.1$, $w_{2Part1} = 0.7$, $w_{2Part2} = 0.1$ and $w_3 = 0.1$ where the largest weight is attributed to the entropy term based on small counts. These weights were motivated by the empirical work carried out at the Office for National Statistics on SDC methods for the 2011 UK census tabular outputs, where attribute disclosure and small counts were of the highest concern.

4.2. Modifying the Disclosure Risk Measure After Perturbation

The disclosure risk measure in (3) does not take into account the application of SDC methods and therefore needs to be modified to reflect the uncertainty that is introduced into the counts of the table. Random rounding, for example, eliminates cells of size one and two by introducing more cells of size zero and three in the table, and seemingly increases the risk of attribute disclosure. However, these additional cells of size zero and three are not true counts and the risk of attribute disclosure should decrease. The disclosure risk as measured by the entropy in (2) (and the second term in (3)) does not reflect this uncertainty on whether the cell count is a true value or not. Therefore, we introduce an additional property for the disclosure risk measure following on from those defined in Subsection 4.1: (e) the disclosure risk measure following the application of an SDC method must be less than the original disclosure risk measure. In order to ensure property (e), we propose to modify the first two terms of the disclosure risk measure in (3) after the application of an SDC method as follows:

Modifying the First Term in (3):

The first term in (3) based on the proportion of zero cells can be generalized to compare the number of zero cells in the original and perturbed table. From (3), A is the set of zero cells in the original table and $|A|$ is the number of zero cells in the set. Similarly, let B be the set of zero cells in the perturbed table and $|B|$ the number of zero cells in the set. Denote $A \cup B$ as the union of the sets of zero cells and $A \cap B$ as the intersection of the sets of zero cells in the original and perturbed table. The revised first term in (3), which takes into account that nonzero cells may have been perturbed into zero cells and vice versa, is defined as: $(|A|/K)^{|A \cup B|/|A \cap B|}$. If there are no zero cells in the original table and hence $A \cap B = 0$, then the first term in (3) will remain equal to 0 following perturbation. For example, assume in a table there is a fraction of 0.10 zero cells and following perturbation a fraction of 0.20 zero cells and all original zero cells remain as zero in the perturbed table. In this case, the power term will be 2 and the risk measure following perturbation is reduced to 0.01 from the original 0.10. The modification of the first term in (3) is always less than the original term if nonzero cells are perturbed to zero cells and vice versa, and thus property (e) is ensured.

Modifying the Second Term in (3):

Assume that the possible values in the table are: $0, 1, 2, \ldots, L$ and the frequency of frequencies of these values is denoted by: $(n_0, n_1, n_2, \ldots, n_L)$. The table is perturbed according to a probability transition matrix (for example, the probability transition matrix $\mathbf{P}$ defined in Subsection 3.3). Let the frequency of frequencies of the perturbed values be denoted by: $(n'_0, n'_1, n'_2, \ldots, n'_L)$. For an observed perturbed value j, $j = 0, 1, \ldots, L$, the expected total from the cells of value j can be estimated by the proportion of the original values of j that are not changed: $(j \cdot n_j) \cdot p_{jj}$ and the proportion of other values i, $i \neq j$ that are changed to value j: $\sum_{i \neq j} (i \cdot n_i) \cdot p_{ij}$, so the expected total from cells of value j after perturbation is: $\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij}$.

To reflect the uncertainty of the counts in the perturbed table, we replace the observed perturbed cells of value j by the expected total from cells of value j distributed evenly across all cells having the perturbed value j: $\left(\left(\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij} \right) / (n'_j) \right)$. As an example, assume the SDC method of random rounding to base 3. We replace the zero cells in the perturbed table with: $[0 \cdot n_0 + 1 \cdot n_1 \cdot (2/3) + 2 \cdot n_2 \cdot (1/3)]/n'_0$. This reflects the fact that zero cells in the perturbed table are not true zeroes; rather, a proportion of them arise from the perturbation of cells of values one and two to zero cells under the probability mechanism, and it is unknown which zero cells are true zero cells and which zero cells are a result of the perturbation. Similarly, for the perturbed cell values of size three, we replace these with the term: $[1 \cdot n_1 \cdot (1/3) + 2 \cdot n_2 \cdot (2/3) + 3 \cdot n_3 + 4 \cdot n_4 \cdot (2/3) + 5 \cdot n_5 \cdot (1/3)]/n'_3$.

For the pretabular method of record swapping, we use a probability transition matrix applied at the cell level of the table for calculating the expectations as explained above, although it is possible that a perturbed table will be equal to the original table if the swapping variable is not involved in generating the table. The expected total from cells of value j in the table after record swapping is: $\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij}$, where p_{ij} is a probability

transition matrix with the swap rate on the diagonal and all off-diagonals have equal probability constrained to the sum of the row probabilities being equal to 1. This means that we assume that every cell in the table can be perturbed according to the swap rate and reflects the assumption that an intruder would not know which variables were swapped.

The modification of the entropy term in (2) replaces observed perturbed counts with their expectations according to the probability transition matrix. In particular, true zero cells which did not contribute to the entropy in the original table are now replaced by their expected values. This should lead to a more even distribution of cell counts in the calculation of the entropy and to a general reduction in the disclosure risk measure in (2) following perturbation. As a final adjustment and to further guarantee property (e), we multiply the resulting entropy-based disclosure risk measures in (2) by a multiplier based on the average of the diagonal probabilities of the probability transition matrix. This multiplier reflects a global level of uncertainty introduced into the perturbed cell counts.

4.3. An Information Theory Data Utility Measure

To assess the distance between two distributions, we use the L_2-norm which, when applied to the difference of two vectors, preserves the properties of a distance metric (non-negativity, coincidence axiom, symmetry and triangle inequality). Measuring the distance infers that the smaller the distance, the more information is left in the table. For an arbitrary vector $x = (x_1, x_2, \ldots, x_K)$ the L_2-norm of x is defined as:

$$\| x \|_2 = \left(\sum_{i=1}^{K} |x_i|^2 \right)^{1/2}.$$

Let $P = (p_1, p_2, \ldots, p_K)$ be the original probability distribution of cell counts and $Q = (q_1, q_2, \ldots, q_K)$ the perturbed probability distribution of cell counts. Define: $\sqrt{P} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K})$ and $\sqrt{Q} = (\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_K})$. These are not (necessarily) probability distributions but have the property that as vectors, their L_2- norms are 1.

The Hellinger Distance is defined as the L_2 -norm:

$$HD(P, Q) = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{P} - \sqrt{Q} \right\|_2$$

and is bounded by 0 and 1.

In the case of frequency distributions from census tables, where $F = (F_1, F_2, \ldots, F_K)$ is the vector of original counts and $G = (G_1, G_2, \ldots, G_K)$ is the vector of perturbed counts, and $\sum_{i=1}^{K} F_i = N$ and $\sum_{i=1}^{K} G_i = M$, the Hellinger distance is defined as:

$$HD(F, G) = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{F} - \sqrt{G} \right\|_2 = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i} \right)^2} \tag{4}$$

The Hellinger distance is grounded in Information Theory and takes into account the magnitude of the cells since the difference between square roots of two "large" numbers is smaller than the difference between square roots of two "small" numbers, even if these pairs have the same absolute difference. Naturally, while the lower bound remains zero,

the upper bound of this distance metric changes:

$$HD(F,G) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i}\right)^2} = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(F_i + G_i - 2 \cdot \sqrt{F_i \cdot G_i}\right)}$$

$$= \frac{1}{\sqrt{2}} \cdot \sqrt{N + M - 2 \cdot \sum_{i=1}^{K} \sqrt{F_i \cdot G_i}} \leq \sqrt{\frac{N+M}{2}}.$$

Since the SDC methods described in Section 3 produce approximately the same overall population total N due to controlled methods of perturbation, the Hellinger distance is bounded by 0 and $\sqrt{N}$. For the comparison study in Section 5, we use the expression of $1 - (HD(F,G)/\sqrt{N})$ as the data utility measure, which is bounded between 0 and 1, 0 representing low utility and 1 representing high utility.

5. A Comparison Study

In this section we present a flexible table-generating server for census tables where we proceed with the European Census Hub approach of defining a large hypercube as the underlying data input to the server. We compare the application of SDC methods described in Section 3 to four generated output tables and examine the properties of the disclosure risk and data utility measures presented in Section 4.

5.1. Preparing the Underlying Hypercube and Generating Output Tables

For the comparison study, we generate a hypercube with an underlying population of size 1,500,000 individuals for two NUTS2 regions. The variables defining the hypercube follow one of Eurostat's specifications for a hypercube required by all Member States as follows:

- NUTS2 Region – 2 regions
- Gender – 2 categories
- Banded age groups – 21 categories
- Current employment status – 5 categories
- Occupation – 13 categories
- Educational attainment – 9 categories
- Country of citizenship – 5 categories

From the UK Census 2001, cell proportions from published tables were calculated and cross-classified using iterative proportional fitting. We then multiplied the proportions by our population size of 1,500,000 individuals to produce the final hypercube. The hypercube used in the comparison study has 245,700 cells. The distribution of cell counts is skewed with a large proportion of zero cells as seen in Table 1.

The distribution of cell counts in the hypercube as shown in Table 1 was comparable to the hypercube based on real census data produced by the United Kingdom according to the above specification.

Table 1. Distribution of cell counts in the generated hypercube

Cell value	Number of cells	Percentage of cells
0	226,939	92.4
1	4,028	1.6
2	2,112	0.9
3–5	2,964	1.2
6–8	1,664	0.7
9–10	720	0.3
11 and over	7,273	3.0
Total	245,700	100.0

In the flexible table-generating server of our comparison study, we apply a set of preliminary SDC rules for generating tables and allow a maximum of three dimensions with one additional variable to define the population of the table. Four different-size output tables are generated from the input hypercube as follows (number of categories of each variable are in parenthesis):

(1) Selected population: NUTS2 = 1, table spanned by: Banded age group (21) * Educational Attainment (9) * Occupation (13).
(2) Selected population: NUTS2 = 2, table spanned by: Gender (2) * Banded age group (21) * Country of citizenship (5)
(3) Selected population: Gender = 1, table spanned by: Current activity status (5) * Occupation (13) * Educational attainment (9)
(4) Selected population: Banded age group = 10, table spanned by: NUTS2 (2) * Occupation (13) * Educational attainment (9)

Table 2 contains details of the four generated output tables that are used in the comparison study: the total size of the population, the number of cells and the average cell size in each table as well as the distribution of cell counts.

Table 2. Details of four generated tables to be used in the comparison study

Details	Table 1	Table 2	Table 3	Table 4
Total Population	854,539	645,461	736,355	96,656
Number of cells	2,457	210	585	234
Average cell size	347.8	3,073.6	1,258.7	413.1
Number of	%	%	%	%
Zeroes	1,534 (62.4)	49 (23.3)	275 (47.0)	84 (35.9)
Ones	44 (1.8)	14 (6.7)	16 (2.7)	9 (3.9)
Twos	35 (1.4)	2 (1.0)	9 (1.5)	4 (1.7)
Threes	27 (1.1)	5 (2.4)	3 (0.5)	6 (2.6)
Fours	20 (0.8)	4 (1.9)	9 (1.5)	1 (0.4)
Fives	17 (0.7)	1 (0.5)	5 (0.9)	4 (1.7)
Sixes and over	780 (31.8)	135 (64.3)	268 (45.8)	126 (53.9)

From Table 2, output Table 1 is the largest table with the largest proportion of zero cells. Output Tables 2 and 4 are similar in the number of cells but the size of the population is considerably smaller in output Table 4, resulting in a larger proportion of zero cells and a smaller proportion of cells of value one. Output Table 3 is a midsize table. It is clear from the small cell counts and many zero cells that the generated output tables require the application of SDC methods in the flexible table-generating server.

In the comparison study we provide an example of how a statistical agency might go about assessing different SDC methods for a flexible table-generating server of census tables through disclosure risk and data utility measures. In the pretabular approach of protecting the input hypercube prior to generating tables, we apply three SDC methods as follows:

- Full random rounding of the hypercube to base 3 semicontrolled to the two NUTS2 totals.
- Random record swapping carried out by first constructing microdata of individuals from the hypercube where each cell is duplicated to the number of individuals in the cell. A random sample of five percent of individuals is selected in each NUTS2 region, then randomly paired with individuals in the opposite NUTS2 region and their geographical variables swapped. This produced a total swap rate of ten percent of individuals having their NUTS2 regions swapped. Following the record-swapping procedure, the hypercube is reconstructed.
- Stochastic perturbation on the hypercube is based on an invariant probability transition matrix with controls in the overall totals of the two NUTS2 regions. The perturbation is carried out on cells of values in the range 0–10; all cells above a value of 10 have the same probabilities of perturbation depending on their residual value to base 5. The probability transition matrix for each NUTS2 region used in this study is presented in Table 3.

The pretabular disclosure-controlled hypercubes are used as input to the flexible table-generating server and the four output tables generated under each SDC method. The comparison results also include the case where a post-tabular SDC method of semicontrolled random rounding to base 3 is applied directly to the four output tables that are generated from the original unperturbed hypercube. The SDC methods are compared through the disclosure risk and data utility measures described in Section 4.

5.2. *Results of the Comparison Study*

To compare the pretabular SDC methods applied to the original hypercube (record swapping, semicontrolled random rounding and stochastic perturbation), we first assess the impact of the perturbation on the small cells in the generated output tables. Table 4 presents the number of small cells of size 1 and 2 in the original hypercube and in each of the four output tables defined in Subsection 5.1, and the percentage of those cells that were altered under the SDC methods. Record swapping generally provided the least number of small cells perturbed except for output Table 4, where the swapped variable NUTS2 is used as a spanning variable of the table. Output Table 3 did not include the swapped NUTS2 variable and hence all cells in the table contain original cell counts. Random rounding eliminates all small cells of size 1 and 2 and provides more protection compared

Table 3. Probability transition matrix used to perturb the hypercube in each NUTS2

		Perturbed Counts											Residual of count to base 5 is:				
Cell Value		0	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5
Original Counts	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0.5000	0.1250	0.1250	0.1250	0.1250	0	0	0	0	0	0	0	0	0	0
	2	0	0.1250	0.5000	0.1250	0.1250	0.1250	0	0	0	0	0	0	0	0	0	0
	3	0	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0.0167	0	0	0	0	0	0	0	0
	4	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0	0	0
	5	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0	0
	6	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0
	7	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0
	8	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0
	9	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0
	10	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0
Residual of count	1	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0
to base 5 is:	2	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167
	3	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167
	4	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.0167	0.9000	0.0167
	5	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.9000

Table 4. Number of small cells of size 1 and 2 in original hypercube and generated tables, and percentage of those cells that were perturbed

	Original hypercube	Table 1	Table 2	Table 3	Table 4
Number of cells of size 1 and 2	6140	79	16	25	13
Percentage perturbed:					
Record swapping	26.9	15.2	12.5	0	30.8
Stochastic perturbation	33.2	29.1	25.0	36.0	23.1
Random rounding	100	100	100	100	100

to record swapping and the stochastic perturbation. It is well known, however, that random rounding has the risk of being able to reveal original cell values, especially when the sum of rounded cells does not add up to the rounded marginal totals. However, ensuring the consistency of the rounding across same cells in different tables and controlling some of the marginal totals lowers the risk of being able to reveal original cell values.

Table 5 presents the disclosure risk measure in (3) and the Hellinger distance in (4) for the output tables defined in Subsection 5.1 generated on the pretabular disclosure-controlled hypercubes according to the SDC methods: record swapping, semicontrolled random rounding and stochastic perturbation. In addition, we report the measures for the case where the SDC method of semicontrolled random rounding is applied directly to the output tables that were generated from the original hypercube to compare the pretabular and post-tabular approach for this SDC method.

To modify the second term in the disclosure risk measure in (3) following the SDC methods as described in Subsection 4.2, we used the following multipliers: for record swapping, the average diagonal probability of the probability transition matrix is 0.9; for the stochastic perturbation, the average diagonal probability of the probability transition matrix is 0.75 for the small counts and 0.9 for the large counts; for the random rounding to base 3, we use the multiplier of 0.33.

From Table 5, we see that the disclosure risk measures are all smaller for the perturbed tables compared to the original tables, even for the case of record swapping in output Table 3 where the perturbed table is identical to the original table since the perturbed NUTS2 variable was not included as a spanning variable of the table. The utility measures are all high, showing that all SDC methods can provide tables that are fit for purpose for users.

In general, it is clear that the method of record swapping when applied to the input hypercube did little to reduce disclosure risk in the final output tables in the comparison study. However, the disclosure risk measure is always slightly smaller than the disclosure risk measure of the original table to reflect the uncertainty in the table based on the assumption that an intruder cannot be certain which variables were swapped. The data utility measure based on the Hellinger distance for output Table 3 under record swapping is 1.00, since the perturbed table is equal to the original table. The data utility measure under record swapping was low for the two output Tables 1 and 2 where the perturbed

Table 5. Disclosure risk and data utility (Hellinger distance) for the generated tables

	Disclosure risk $R(F, w_1, w_2)$ in (3)	Data utility $1 - (HD(F, G)/\sqrt{N})$ in (4)
Table 1		
Original	0.318	-
Perturbed input		
Record swapping:	0.282	0.988
Semicontrolled random rounding	0.137	0.991
Stochastic perturbation	0.239	0.995
Perturbed output:		
Semicontrolled random rounding	0.135	0.993
Table 2		
Original	0.248	-
Perturbed input:		
Record swapping	0.191	0.972
Semicontrolled random rounding	0.070	0.996
Stochastic perturbation	0.210	0.998
Perturbed output:		
Semicontrolled random rounding	0.072	0.996
Table 3		
Original	0.339	-
Perturbed input:		
Record swapping	0.295	1.000
Semicontrolled random rounding	0.130	0.994
Stochastic perturbation	0.254	0.996
Perturbed Output:		
Semicontrolled random rounding	0.127	0.996
Table 4		
Original	0.298	-
Perturbed input:		
Record swapping	0.271	0.987
Semicontrolled random rounding	0.105	0.991
Stochastic perturbation	0.229	0.994
Perturbed output:		
Semicontrolled random rounding	0.105	0.992

NUTS2 variable was used to select the population for these tables. The data utility measure under record swapping for output Table 4 was slightly higher, since in this case NUTS2 was a variable spanning the table and hence did not change the overall total of the table.

The stochastic perturbation carried out on the input hypercube outperformed record swapping with smaller disclosure risk measures and higher data utility measures (except for output Table 3). The stochastic perturbation has a higher disclosure risk compared to semicontrolled random rounding, since a large percentage of small cells are unchanged by the perturbation, but it has higher data utility.

The semicontrolled random rounding outperformed all other methods with respect to the lowest disclosure risk, since there are no small cells in the tables and attribute disclosure risk is reduced by the introduction of random zeros. However, the data utility measure based on the Hellinger distance was slightly lower compared to the stochastic perturbation method as mentioned above. There was little difference between the disclosure risk measures comparing the pretabular semicontrolled random rounding on the input hypercube to the post-tabular semicontrolled random rounding applied directly to the output tables generated from the original hypercube. However, there is an increase in the data utility measure when applying the post-tabular semicontrolled random rounding, especially for the large output Table 1 and midsize output Table 3.

Figure 2 presents a disclosure risk-data utility map of the four generated tables where RS is record swapping, SP is the stochastic perturbation, RR is the semicontrolled random rounding on the input hypercube and RRP is the semicontrolled random rounding applied directly to the generated output tables. The data utility measure is the Hellinger distance in (4). The upper right-hand quadrant of the map represents high disclosure risk and high utility and the lower left-hand quadrant represents low disclosure risk and low data utility.

The statistical agency needs to decide on a tolerable disclosure risk threshold above which they are not prepared to release a table. As an example, the disclosure risk-data utility map shows that for a tolerable disclosure risk threshold of up to 15 percent, the

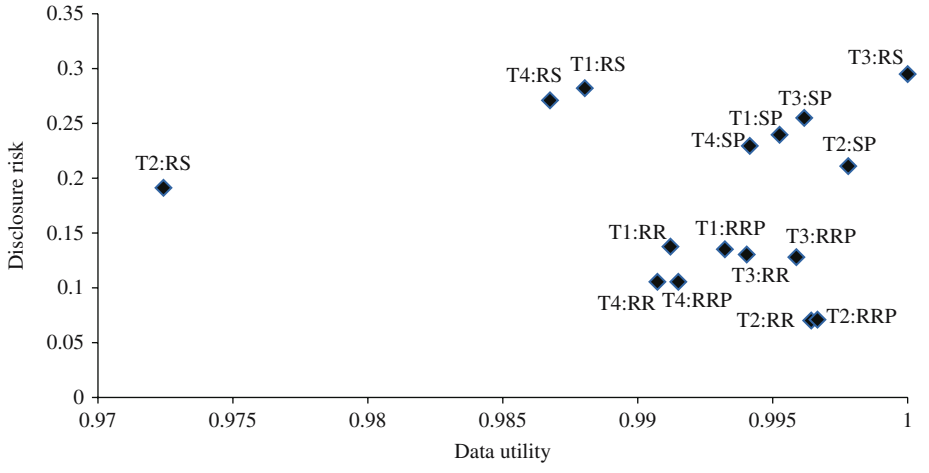

Fig. 2. Disclosure risk – data utility map for generated tables (output Table 1 (T1) to output Table 4 (T4)): RS – record swapping, SP – stochastic perturbation, RR – semicontrolled random rounding on input hypercube, RRP – semicontrolled random rounding on generated tables

output tables where semicontrolled random rounding was applied directly to tables that were generated from the original hypercube have the highest data utility as they are on the farthest right-hand side of the map.

6. Concluding Remarks

In this article, we have compared pretabular SDC methods applied to a large hypercube (record swapping, stochastic perturbation and semicontrolled random rounding) and a semicontrolled random rounding applied directly to output tables generated from the original hypercube. For the pretabular SDC methods, record swapping had little impact on reducing disclosure risk and also had lower data utility. Semicontrolled random rounding offered more protection as all cell values in the table not a multiple of base b are perturbed, and by preserving the consistency of cells across tables, it is more difficult to undo the rounding to reveal original cell values. The stochastic perturbation had the best overall data utility, but entailed higher disclosure risks compared to the semicontrolled random rounding. Finally, we have seen that the post-tabular SDC method of semicontrolled random rounding applied directly to the generated output tables produced nearly the same amount of disclosure risk reduction as the pretabular semicontrolled random rounding applied to the input hypercube, but had a higher level of data utility.

The aim of the comparison study was not primarily to evaluate specific SDC methods or indeed determine their optimum parameterization, but rather to demonstrate how such a disclosure risk and data utility analysis should be carried out by a statistical agency when disseminating census data. To this end, we have proposed new global measures of disclosure risk and data utility based on information theory that are particularly suited for assessing disclosure risk arising from attribute and identity disclosure in census frequency tables and can easily be embedded in a web-based flexible table-generating server. The proposed modifications to the disclosure risk measure following the application of an SDC method show that we can reflect the level of uncertainty added to the tables and therefore reduce the disclosure risk. Further research is needed to refine and improve post-tabular SDC methods whilst preserving additivity and consistency of user-defined tables. More extensive empirical studies are needed that involve real data and the testing of SDC methods across their respective parameter spaces.

Another key aspect of the SDC problem in a flexible table-generating server is the management of users and governance processes. The server can be freely available on the statistical agency's website for all users or restricted via licensing and passwords to only approved users. For the former case, it is clear that SDC rules and methods would have to be highly protective to guard against the fact that users can query the same table multiple times in an attempt to undo SDC methods and reveal original cell counts. Clearly, perturbative SDC methods, preserving the additivity and consistency of same cells across different tables, and high thresholds for dissemination would be required. For the latter case, less protection would be needed, allowing for higher-quality data, but protocols would then need to be in place to handle multiple overlapping queries from the same user, the management of users and their expectations.

7. References

Antal, L., N. Shlomo, and M. Elliot. 2014. "Measuring Disclosure Risk with Entropy in Population Based Frequency Tables." In *PSD'2014 Privacy in Statistical Databases*, edited by J. Domingo-Ferrer, 62–78. Berlin: Springer.

Cover, T.M. and J.A. Thomas. 2006. *Elements of Information Theory*, 2nd ed. New York: Wiley.

Dalenius, T. and S.P. Reiss. 1982. "Data Swapping: A Technique for Disclosure Control." *Journal of Statistical Planning and Inference* 7: 73–85.

Doyle, P., J.I. Lane, J.M.M. Theeuwes, and L. Zayatz. 2001. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: Elsevier Science B.V.

Duncan, G., M.J. Elliot, and J.J. Salazar. 2011. *Statistical Confidentiality: Principles and Practice*. New York: Springer.

Fienberg, S.E. and J. McIntyre. 2005. "Data Swapping: Variations on a Theme by Dalenius and Reiss." *Journal of Official Statistics* 9: 383–406.

Fraser, B. and J. Wooton. 2005. "A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, November 9–11. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf (accessed April 2015).

Gouweleeuw, J., P. Kooiman, L.C.R.J. Willenborg, and P.P. De Wolf. 1998. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." *Journal of Official Statistics* 14: 463–478.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.P. de Wolf. 2012. *Statistical Disclosure Control*. Chichester: John Wiley & Sons.

Salazar-Gonzalez, J.J., C. Bycroft, and A.T. Staggemeier. 2005. "Controlled Rounding Implementation." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, November 9–11. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.36.pdf (accessed April 2015).

Shlomo, N. 2007. "Statistical Disclosure Control Methods for Census Frequency Tables." *International Statistical Review* 75: 199–217. Doi: http://dx.doi.org/10.1111/j.1751-5823.2007.00010.x.

Shlomo, N. and C. Young. 2008. "Invariant Post-tabular Protection of Census Frequency Counts." In *In PSD'2008 Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and Y. Saygin, 77–89. Berlin: Springer.

Willenborg, L.C.R.J. and T. de Waal. 2001. *Elements of Statistical Disclosure Control*. New York: Springer.

Received July 2013
Revised October 2014
Accepted November 2014