Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (i): New methods for protection of tabular data or for other types of results from table and analysis servers

Measuring Disclosure Risk and Data Utility for Flexible Table Generators

Prepared by Natalie Shlomo, Laszlo Antal and Mark Elliot, University of Manchester, United Kingdom

Measuring Disclosure Risk and Data Utility for Flexible Table Generators

Natalie Shlomo, Laszlo Antal and Mark Elliot*

* Social Statistics and CCSR University of Manchester, United Kingdom, natalie.shlomo@manchester.ac.uk

Abstract: Statistical agencies are considering making more use of the internet to disseminate tabular outputs through on-line flexible table generating servers that allow users to define and generate their own tables. The key questions in the development of these servers are what data should be used to generate the tables and what statistical disclosure control (SDC) method should be applied. For flexible table generating, the server has to measure the disclosure risk in the table, apply the SDC method and then 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 generated from original data. Besides disclosure risk, the server should provide measures of information loss comparing the perturbed table to the original table. In this paper, we examine the development of a flexible table generating server and demonstrate an application of measuring risk-utility comparing different SDC methods for census data. We propose measures for disclosure risk and data utility that are based on Information Theory.

Acknowledgement: 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).

1 Introduction

Many statistical agencies are considering the development of flexible table generating servers that allow users to define and generate their own tables. The United States Census Bureau and the Australian Bureau of Statistics (ABS) have developed such servers on their websites for disseminating census tables. Other agencies, such as the Israel Central Bureau of Statistics, have developed a flexible table generator for their Social Survey. Users access the servers via the internet, define their own table from a set of pre-defined variables/categories using drop down lists and then download their table of interest.

The key questions in the development of these servers are: (1) what data should be used in the background for producing the tables, and (2) at what stage should the statistical disclosure control (SDC) method be applied. This paper addresses these two questions within a broad SDC framework but also examines and compares some specific SDC methods using census data as a demonstration of how statistical agencies should undertake a disclosure risk-data utility analysis when considering these questions.

In general, SDC methods employed by statistical agencies are often motivated by country-specific agendas and policy sensitivities and it is difficult to develop universal best practice. However, one important distinction when considering SDC methods for flexible table generators is that the outputs are defined by the users and the amount of disclosure risk may vary in each output. It is this distinction that motivates statistical agencies to examine the Computer Science literature on the theory of guaranteeing privacy in outputs from query-based systems. In that literature, disclosure risk is formally defined as 'differential privacy' where a 'worst case' scenario is allowed for in which the intruder has complete information about all the units in the population database except for one unit of interest. To guarantee differential privacy, perturbative methods based on noise addition under specific parameterizations are introduced (see for example Dinur and Nissim, 2003, Dwork, et al. 2006).

With respect to the question of what data should be used in the background of a flexible table generator server, we can use original microdata which may or may not have undergone SDC methods. Often the original data is aggregated to minimum cell levels below which data cannot be disseminated via the server. With respect to the question at what stage to apply SDC methods in a flexible table generating server, there are two approaches: apply SDC to the underlying data so that all tables generated in the server are deemed safe for dissemination (the pre-tabular SDC method), or produce tables directly from original data and apply the SDC method to the final tabular output (the post-tabular SDC method). Although a sometimes neater and less resource intensive when data is from a single source, the pre-tabular approach is problematic in some cases. For example, for disseminating European Census data all member states would have to agree on a common SDC method since if one member state employs a rounding algorithm whilst another member state employs cell suppression, there will be significant utility loss in a table that is generated based on both member states' data. Moreover, when aggregating data which has been separately disclosure controlled, we compound the effects of the SDC and over-protect the data. For example, aggregating rounded cells not only over-protects the data but also exacerbates the data utility impact by providing counts that are no longer rounded to the nearest rounding base. With the approach of protecting final tabular outputs, SDC methods are not compounded in this way.

For flexible table generating, the server has to be able to measure the disclosure risk in the original table, apply an SDC method and then iteratively reassess the disclosure risk. There are two main types of disclosure risks in frequency tables: identity disclosure where small cell counts may lead to identification of an individual population unit, and attribute disclosure where rows/columns contain (real) zeros and only a small number of cells are non-zero. This potentially enables a user to obtain attribute information of an individual or group of individuals. Another issue that needs to be guarded against is that 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 the main concern in differential privacy as defined by the Computer Science literature and is dealt with by implementing perturbative methods of SDC. After the table is protected, the server should also be able to calculate a data utility measure on the impact of the SDC method by comparing the perturbed table to the original table.

Section 2 discusses the design of flexile table generating servers. Section 3 summarizes some new developments in disclosure risk and data utility measures based on Information Theory. Section 4 presents an application and comparative study of a risk-utility analysis using census tables. The results of the study are presented in Section 5 with general conclusions in Section 6.

2 Designing Flexible Table Generator Servers

The design of a remote flexible table generating server typically involves many ad-hoc preliminary SDC rules that can easily be programmed within the system to determine a priori tables which should not be released. These SDC rules may include: limiting the number of dimensions in the 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 above a minimum threshold; and ensuring average cell size above a minimum threshold. Despite these preliminary SDC rules, the output tables generated in the server may still have high disclosure risk and require the application of further SDC methods. These methods can be applied on the underlying data of the server prior to tabulation or

applied directly to the final output table generated from original data, or both. The steps that the flexible table generating server should take are:

(1) Determine if the table can be generated according to the ad-hoc SDC rules;
(2) Calculate a disclosure risk measure to determine if an SDC method should be applied;
(3) Apply SDC method;
(4) Recalculate the disclosure risk measure to determine if the table is safe to generate; if yes proceed to step 5; else return to step 3.
(5) Output the final table with a measure of the data utility.

We describe some types of data that might be placed in a flexible table generating server:

Census Data: One application of a flexible table generating server is being developed for the dissemination of census tables within the European Census Hub Project. Each member state is required to produce a fixed set of pre-defined multi-dimensional tables (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. The hypercubes will be used as the underlying data behind the flexible table generating server. The platform will allow comparative tables across member states and the combination of census data from multiple member states. The fixed set of hypercubes allow harmonization of census results and have the additional advantage that they provide some a priori protection against disclosure risk since no data below the level of the cells of the hypercube will be released. Nevertheless, the hypercubes are very large with many zero and small cells and carry considerable residual disclosure risk. We demonstrate an application and comparative study on this application in Sections 4 and 5.

Business Data: Business data are highly sensitive and generally placed within safe data enclaves. Producing synthetic data from statistical models (Reiter, 2005) based on the original data preserves some of the analytical properties and may be placed in a flexible table generating server. Compromises can be made, for example, by allowing the dissemination of only medium and small businesses as is the practice at the ABS. The table server would have to accommodate magnitude tables which are typically released for business statistics and this implies different SDC methods, eg. controlled tabular adjustment (CTA) of Dandekar and Cox, 2002, as well as different types of disclosure risk measures. We will not consider business data further in this paper.

Sample Data: Flexible table generating servers can disseminate weighted survey data from social surveys with little intervention. Statistical agencies generally regard weighted sample counts from social surveys with large and differential sample weights as safe for dissemination. Any cells that have small underlying sample counts are generally suppressed because of low efficiency and this solves the problem of disclosure risk. However, un-weighted (original) sample counts without further SDC methods applied are disclosive in a flexible table generator and should not be disseminated. As shown in Shlomo and Skinner, 2012, differential privacy cannot be guaranteed when using the mechanism of sampling as an SDC method. Under the differential privacy definition, it is assumed that the intruder knows attributes of every unit in the population table except for one target unit, and that the intruder cannot make inference about the target unit when only one of its values is changed. Assume the population database having K cells where F_k is the population total and f_k is the associated sample total in cell k of the table, $k=1,\ldots K$. If we move one individual in the population from cell k to another cell, then we may obtain that $F_k<f_k$ which is impossible and therefore

inference can be made about the target unit. How likely is it to get F_k=f_k in a sample? For sample and population doubles, triples, etc. the chances are extremely small under typical social survey designs. However, the chance of obtaining a population unique from a sample unique are larger and causes considerable disclosure risk. Statistical agencies generally allow this 'slippage' and release data from social surveys often with recoded and banded variables because they do not assume the strict notion of response knowledge as in the differential privacy paradigm and there may be access restrictions on the data through licensing. In addition, population uniques are generally unknown, even to the agency itself. However, allowing open access to the un-weighted sample counts in a web-based flexible table generating server may open the statistical agency to attacks on the data and should be avoided. Skinner and Shlomo, 2012 also point out that many perturbation methods can be made to conform to the guarantee of differential privacy provided that the method is stochastic and allows for all possible perturbations, i.e. the probability transition matrix for the perturbation does not have zero probabilities in the off-diagonals.

3 Information Theory Based Disclosure Risk and Data Utility Measures

For each final output table generated, the server must provide disclosure risk and data utility measures. We propose to use Information Theory (IT) to define these measures since this theory is particularly sensitive to the case of attribute disclosure which is caused by a row/column or table having many (real) zeros and only a few cells that are nonzero. Information Theory based disclosure risk measures are discussed in Shlomo, Antal and Elliot, 2013 and summarized below with extensions.

3.1 An Information Theory Disclosure Risk Measure

A high entropy indicates that the distribution across cells is uniform and a low entropy indicates mainly zeros in a row/column or table with a few non-zero cells. The fewer the number of non-zero cells, the more likely that attribute disclosure occurs. To produce a disclosure risk measure between 0 and 1 for census tables containing whole population counts, we use the entropy $H\left(\frac{F}{N}\right) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N}$ for a frequency vector in a table of size K, $F = (F_1, F_2, \ldots, F_K)$ where $\sum_{i=1}^{K} F_i = N$. Taking into account the bounds of the entropy, the measure is defined as: $1 - H\left(\frac{F}{N}\right) \Big/ log\ K$. The entropy however does not take into account the magnitude of the cell which contributes to identity disclosure. Let A be the set of zeros in the table and $|A|$ the number of zeros in the set. We define a disclosure risk measure as a weighted average of different components, each component being a measure between 0 and 1 as follows:

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

The first measure in (1) is the proportion of zeros which is relevant for attribute disclosure, the more zeros in a table, the more risk of learning new attributes after an identification. The second measure in (1) is the risk based on the entropy which is the core of the overall risk measure. The third measure in (1) allows us to differentiate between tables with different magnitudes. As the population size N gets larger, the third measure converges to zero. The

weights w_1 and w_2 can be chosen depending on the data protector's choice of how important each of the terms are in contributing to the disclosure risk or can be calibrated to obtain the maximal disclosure risk measure.

3.2 Adapting the Disclosure Risk Measure After Perturbation

The disclosure risk measure in (1) does not take into account the application of SDC methods that might be typically carried out on census data containing whole population counts and therefore needs to be adapted to reflect the uncertainty that is introduced into the counts of the table. Random rounding to base 3, for example, eliminates ones and twos by introducing more zeros and threes in the table, and seemingly increases the risk of attribute disclosure. However, these additional zeros and threes are not true counts rather are random according to the SDC mechanism and therefore should decrease the risk of attribute disclosure in the table. The disclosure risk as measured in (1) does not reflect this randomness. In order to take into account the perturbation, we propose to modify the first two terms of the risk measure in (1) as follows:

We generalize the first term of the proportion of zeros in (1) to compare the number of zeros in the original and perturbed table. From (1), A is the set of zeros in the original table and $|A|$ is the number of zeros in the set. Similarly, let B be the set of zeros in the perturbed table and $|B|$ the number of zeros in the set. We denote $A \cup B$ as the union of the sets of zeros in the original and perturbed table and $A \cap B$ as the intersection of the sets of zeros in the original and perturbed table. The revised measure, which takes into account that non-zero cells may be transformed into zero counts and vice versa, is defined as: $\left(\frac{|A|}{K}\right)^{\frac{|A \cup B|}{|A \cap B|}}$.

To control the rate of convergence of this term to zero we may replace the power term $\frac{|A \cup B|}{|A \cap B|}$ with a square root: $\sqrt{\frac{|A \cup B|}{|A \cap B|}}$.

For the entropy based term, we assume that the possible cell values in the table are: 0,1,2,…,L and the frequency of frequencies of these values is denoted by: $(n_0, n_1, n_2, ..., n_L)$ where for $i = 0,1,...L$, $n_i = \sum_{k=1}^{K} I(value \ \ of \ \ k^{th} \ \ cell \ \ is \ \ i)$ and I is the indicator function. Assume that the table is perturbed according to a perturbation mechanism, for example using a probability transition matrix **P** which is an $(L+1)\times(L+1)$ matrix containing conditional probabilities: $p_{ij} = P(perturbed \ \ cell \ \ value \ \ is \ \ j \,|\, original \ \ cell \ \ value \ \ is \ \ i)$ for cell values from 0 to L (usually a cap L is put on the cell value). Let the frequency of frequencies of the perturbed values be denoted by: $(n_0', n_1', n_2', ..., n_L')$. For an observed perturbed value j, j=0,1,…,L, the sum of the cells of value j in the perturbed table can be estimated by the sum of the proportion of the original counts of j that are not changed: $(j \cdot n_j) \cdot p_{jj}$ and the sum of the proportion of other counts i, $i \neq j$ that are changed to value j: $\sum_{i \neq j} (i \cdot n_i) \cdot p_{ij}$, so the expected sum of cells of category 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 cells of value j by the term: $\sum_{i=0}^{L}(i \cdot n_i) \cdot p_{ij} \Big/ n'_j$ thereby distributing the expected total value across all of the cells of value j. As an example, assume the SDC method of random rounding. We replace the zero cells in the perturbed table with: $\left[0 \cdot n_0 + 1 \cdot n_1 \cdot \frac{2}{3} + 2 \cdot n_2 \cdot \frac{1}{3}\right] \Big/ n'_0$. Similarly, for the observed cell values of size three, we replace these with the term: $\left[1 \cdot n_1 \cdot \frac{1}{3} + 2 \cdot n_2 \cdot \frac{2}{3} + 3 \cdot n_3 + 4 \cdot n_4 \cdot \frac{2}{3} + 5 \cdot n_5 \cdot \frac{1}{3}\right] \Big/ n'_3$. The procedure takes into account the uncertainty in the cell value of the perturbed table and ensures the same overall total of the original and adjusted vector of counts. After replacing the values in the perturbed table, we calculate the entropy as shown in (1).

3.3 Adapting the Disclosure Risk Measure After Sampling

Under simple random sampling and a sample fraction of π, the expected value for the sample count in cell k is $f_k = \pi F_k$ and therefore depending on the population count and the sample fraction, there should be little difference in the entropy term. The difference in the entropy term would result from the introduction of random zeros which would also affect the first term in the disclosure risk measure of (1). Since population counts are generally unknown and only sample counts are observed, we can use the probabilistic modelling framework of Skinner and Shlomo, 2008 to estimate population parameters and use these to estimate the disclosure risk measure in (1). Let $\mathrm{f} = \{ f_k \}$ denote a K-way frequency table, which is a sample from a population table $\mathrm{F} = \{ F_k \}$, where $k = (k_1,...,k_K)$ indicates a cell and f_k and F_k denote the frequency in the sample and in the population cell k, respectively. Denote by n and N the sample and population size, respectively.. A common assumption in the frequency table literature is $F_k \sim Poisson(\lambda_k)$, independently, where $\sum_k F_k = N$ is a random parameter. Binomial (or Poisson) sampling from F_k means that $f_k \mid F_k \sim Bin(F_k, \pi)$ independently. By standard calculations we then have: $f_k \sim Poisson(\lambda_k \pi)$ and, $F_k \mid f_k \sim f_k + Poisson(\lambda_k(1-\pi))$, where $F_k \mid f_k$ are conditionally independent. Following the approach of Skinner and Shlomo, 2008 (and references therein) we use log linear models to estimate population parameters. The sample counts $\{ f_k \}$ are used to fit a log-linear model: $log\ \mu_k = x'_k \beta$ where $\mu_k = \lambda_k \pi$ in order to obtain estimates for the parameters: $\hat{\lambda}_k = \hat{\mu}_k / \pi$. Under simple random sample, the maximum likelihood (MLE) estimator $\hat{\beta}$ may be obtained by solving the score equations: $\sum_k [f_k - exp(x'_k \beta)] x_k = 0$. For the entropy term in (1), we replace F_k by $\hat{\lambda}_k$. For the proportion of zeros in the first term in (1), we estimate the number of zero cells in the population by $\sum_k exp(-\hat{\lambda}_k)$.

3.4 An Information Theory Data Utility Measure

For the data utility measure we use the distance metric defined by the Hellinger Distance where $(\sqrt{G_1}, \sqrt{G_2}, ..., \sqrt{G_K})$ represent the square roots of the perturbed counts:

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

Since SDC methods applied to tables will generally have the same overall total N due to controlled methods of perturbation, we can compare the Hellinger Distance across the SDC methods as it is bounded by 0 and approximately $\sqrt{N}$.

4 Application of a Flexible Table Generating Server

There are several options for deciding what data should be used as input in a flexible table generating server. For example, one can choose to use the original microdata, disclosure controlled microdata or hypercubes (with or without minimal cell sizes). The input data is largely determined by the data type and content as well as the SDC method that will be applied on the final output tables (if any). This application is based on the European census tables where each member state is producing a set of predefined hypercubes to be used as input into a flexible table generating server. We simulate a census hypercube with an underlying population of 1,500,000 individuals for two NUTS2 regions. The variables defining the hypercube follow the Eurostat specification for one of the hypercubes:

- NUTS2 Region - 2 regions
- 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

From the UK Census 2001, we calculate cell proportions from available published tables, multiplied the proportions by the 1,500,000 individuals in the population and calculated all cross-classified proportions of the table through iterative proportional fitting to produce the final synthetic hypercube. The hypercube used in the simulation study had 245,700 cells. The distribution of cell counts is skewed with a large proportion of zero cells as seen in Table 1. The distributions in the synthetic hypercube were compared to those obtained from real hypercubes produced by member states Italy and Estonia at the NUTS2 region level according to the above specification and similar distributions were obtained.

Table 1: Distribution of Cell Counts in the Synthetic Hypercube

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 and over	7,273	2.96%
Total	245,700	100.00%

As an example of a possible final output table that can be generated in a flexible table generating server, we assume that the ad-hoc SDC rules would allow the user to generate a table of up to three dimensions with one additional variable defining the population. For our

output table in the comparative study, we select the population as those in the first NUTS2 region and define the 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 original cell counts of the final output table are shown in Table 2. It is clear by the small cell counts and many zero cells that this final output table will need the application of SDC, either by applying the SDC method to the input hypercube or applying the SDC method to the final output table. We also draw a 1:50 sample from this population and produced a sample based table as shown in Table 2.

Table 2: Distribution of Cell Counts in the Generated Table for Population and 1:50 Sample: Banded Age Group*Education*Occupation for NUTS2=1

Cell Value	Population		1:50 Sample	
	Number of Cells	Percentage	Number of Cells	Percentage
0	1534	62.4%	1837	74.8%
1	44	1.8%	105	4.3%
2	35	1.4%	49	2.0%
3	27	1.1%	43	1.8%
4	20	0.8%	31	1.3%
5 and over	797	32.4%	392	16.0%
Total	2457	100.0%	2457	100.0%

We applied the following pre-tabular SDC methods on the population hypercube:

- Random record swapping (see Fienberg and McIntyre, 2005) at the individual level where 5% of the individuals were selected in each NUTS2 region; the selected individuals were paired randomly with other individuals in different LAU2 geographies within the NUTS2 region, and the LAU2 geographies swapped between them. This produced a total of 10% of individuals in each NUTS2 region having their LAU2 geography variable swapped.
- Full random rounding to base 3 semi-controlled to the two NUTS2 totals in the hypercube (see Shlomo, 2007). We also apply semi-controlled random rounding to base 3 on the final output table generated from the original hypercube for the comparison study in Section 5.
- Stochastic perturbation based on an invariant probability matrix with controls in the overall totals of the two NUTS2 regions (See Shlomo and Young, 2008). We carry out the perturbation on cells of values in the range 0-10; all cells above a value of 11 were not perturbed. The invariant perturbation matrix used in this study is presented in Table 3.

Table 3: Invariant Perturbation Matrix used to Perturb Hypercube

Cell	Perturbed Cell Value										
Value	0	1	2	3	4	5	6	7	8	9	10
0	0.998	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.080	0.760	0.080	0.047	0.024	0.004	0.002	0.001	0.001	0.000	0.000
2	0.080	0.153	0.686	0.047	0.024	0.005	0.002	0.001	0.001	0.000	0.000
3	0.000	0.148	0.078	0.703	0.027	0.031	0.007	0.002	0.002	0.001	0.001
4	0.000	0.103	0.054	0.037	0.725	0.022	0.024	0.020	0.006	0.005	0.005
5	0.000	0.023	0.014	0.055	0.029	0.783	0.031	0.025	0.023	0.009	0.008
6	0.000	0.013	0.007	0.012	0.032	0.032	0.814	0.029	0.026	0.025	0.010
7	0.000	0.005	0.003	0.005	0.035	0.034	0.037	0.797	0.029	0.027	0.027
8	0.000	0.005	0.003	0.005	0.013	0.039	0.042	0.036	0.798	0.030	0.030
9	0.000	0.005	0.003	0.005	0.013	0.017	0.046	0.039	0.034	0.807	0.032
10	0.000	0.005	0.003	0.005	0.013	0.017	0.021	0.043	0.037	0.034	0.823

5 Results of Comparative Study

We report in Table 4 the disclosure risk measure in (1) and the Hellinger Distance in (2) for the small generated table based on SDC methods implemented on the input hypercube (record swapping, semi-controlled random rounding and stochastic perturbation). In addition, we report the measures when implementing the SDC method of semi-controlled random rounding applied directly on the output table generated from the original hypercube as well as on the sample-based table. The final disclosure risk measure (1) is calculated 'on the fly' within the flexible table generating server without the need to see the table beforehand. In order to emphasize the risk of small counts (ones and twos) which may still remain in the table for some of the SDC methods, we split the entropy risk measure as shown in (1) into two parts, small counts up to 3 and larger counts 4 and more, and provide different weights for each part. For this study, we use the weights: $w_1 = 0.1$, $w_{2Part1} = 0.7$, $w_{2Part2} = 0.1$ and $w_3 = 0.1$ which provides the largest weight to the entropy risk measure based on small counts. Regarding the number of small cells of size 1 and 2, there were a total of 6,140 small cells in the hypercube (2.5%). The stochastic perturbation changed only 6.9% of the small cells, the random rounding to base 3 changed 100% of the small cells and the random record swapping changed 16.2% of the small cells. The sample based table had 2 sample uniques that were population uniques out of 105 sample uniques.

For the disclosure risk measure on the sample-based small table, we estimate population parameters through the log-linear model approach using an all two-way interactions model on the three dimensions of the table. The number of estimated zeros in the population based on the log-linear model is 1,620 compared to the true 1,534 zero counts in the original table and the 1,846 zero counts in the sample table. The entropy risk measure defined as the second term in (1) for the full population counts is 0.318. Based on the sample counts, with the additional random zeros, this term increased to 0.323. By 'smoothing' out the disclosure risk using the log-linear modeling, the term is 0.319. For the sample based table, the Hellinger Distance is measured by the distance between the estimated population parameters and the true population counts.

Table 4: Disclosure Risk and Data Utility for the Generated Table

	Disclosure Risk	**Hellinger Distance from Original Table**
Original Table	**0.352**	**-**
1:50 Sample based Table	**0.425**	**59.054**
Perturbed Input		
Record Swapping	**0.351**	**6.469**
Semi-controlled Random Rounding	**0.237**	**7.970**
Stochastic Perturbation	**0.230**	**14.120**
Perturbed Output		
Semi-Controlled Random Rounding	**0.233**	**5.902**

Based on Table 4, the disclosure risk of the sample based table is greater than the original table and the original table with SDC perturbation methods applied. One reason is that the

table was split for the calculation of the entropy term of the risk measure in (1) although the log-linear model (the smoothing) was carried out across the whole table. In general, we would expect the sample based risk measure to be equal or less than the risk measure for the original data and further work needs to be carried out to reduce the risk of the sample based table. On the other hand, sampling is considered a non-perturbative method of SDC and disclosure risk remains high with population uniques in the table. Sample counts in particular have high disclosure risks and it is recommended that only weighted sample counts be disseminated. The Hellinger Distance is also much greater in the sample based table reflecting the high level of information loss when disseminating only a 1:50 sample.

Comparing the tables of whole population counts with perturbation in Table 4, it is clear that the method of record swapping when applied to the input hypercube did little to reduce the disclosure risk in the final output table. This was due to the fact that most of the small cells remained unperturbed in the final table. On the other hand, record swapping provides the smallest distance metric (highest data utility) between the original and perturbed table compared to the other pre-tabular methods. From among the input perturbation methods on the hypercube, the stochastic perturbation provided the most protection against disclosure but at the cost of a low data utility with the highest distance metric between the original and perturbed table. Removing the small cells entirely and rounding the other cells provided lower disclosure risk as seen in the measures for the semi-controlled random rounding but had less of an impact on the data utility. Comparing the pre-tabular and post-tabular semi-controlled random rounding procedure, we see slightly lower disclosure risk based on the post-tabular rounding but much improvement in data utility since the SDC method is not compounded by aggregating rounded cells. The semi-controlled random rounding on the final output table would be the preferred method based on the results of the study.

6 Concluding Remarks

In this paper, we have examined flexible table generating servers and demonstrated how a comparative study can be carried out to assess applications of SDC methods at different stages of generating tables within the server. We have seen that a post-tabular SDC method on the final output table produces nearly the same amount of disclosure risk reduction as pre-tabular perturbative SDC methods whilst achieving the best level of data utility. However, the aim of this study was not primarily to evaluate specific SDC methods for Census tables, rather we aimed to demonstrate how such a disclosure risk and data utility analysis should be carried out. To this end, we have proposed new measures for disclosure risk and data utility based on Information Theory which are particularly suited for assessing disclosure risk arising from attribute disclosure in tables and can easily be embedded in a flexible table generating server. Further research is needed to refine and improve the post-tabular SDC methods whilst preserving additivity and consistency of user-defined tables. More extensive empirical studies with real (rather than synthetic) data and the various SDC methods tested across their respective parameter spaces are also needed. Finally, further research is needed on the theoretical properties of the Information Theory based disclosure risk and data utility measures, in particular for un-weighted sample based tables.

References

Dandekar, R.A. and Cox L. H. (2002). Synthetic Tabular Data: An Alternative to Complementary Cell Suppression. *Manuscript, Energy Information Administration*, U. S. Department of Energy.

Dinur, I. and Nissim, K. (2003). Revealing Information While Preserving Privacy. *PODS 2003*, pp. 202-210.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS Vol. 3876, 265-284.

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

Reiter, J.P. (2005), Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society*, A, Vol.168, No.1, 185-205.

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

Shlomo, N. Antal, L. and Elliot, M. (2013). Disclosure Risk and Data Utility in Flexible Table Generators**.** *Proceedings of the NTTS2013 Conference*, Brussels, March 5-7. http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_68.pdf

Shlomo, N. and Skinner. C.J. (2012). Privacy Protection from Sampling and Perturbation in Survey Microdata. *Journal of Privacy and Confidentiality*, Vol. 4, Issue 1.

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