WP.18
ENGLISH ONLY

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

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (ii): Tabular data protection

ASSESSING THE IMPACT OF SDC METHODS ON CENSUS FREQUENCY TABLES

Invited Paper

Prepared by Natalie Shlomo, University of Southampton, United Kingdom

Assessing the Impact of SDC Methods on Census Frequency Tables

Natalie Shlomo*

* Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, UK SO17 1BJ, e-mail: N.Shlomo@Soton.ac.uk

Abstract: Statistical Agencies are faced with increasing demands by users to release more detail and high quality statistical data. This requires examining the trade off between managing disclosure risk below tolerable thresholds and disseminating "fit for purpose" data with as much information as possible. In particular, protecting Census data containing whole population counts is one of the greatest SDC challenges and confidentiality requirements and codes of practice are constantly changing to meet these demands for high quality small area data. The impact of SDC methods on whole population counts causes much information loss and hence the need to evaluate a wide range of SDC methods. In this paper we take an in depth look at one particular large table from the UK 2001 Census with respect to measuring disclosure risk, implementing SDC methods and comparing their impact on information loss through measures based on distortions to distributions, measures of association and other statistical analysis tools.

1 Introduction

Statistical Agencies are facing increasing demands to disseminate more detail and high quality statistical data for small areas based on Census results or administrative sources. The standard mode of dissemination for whole population counts are frequency tables. Protecting these tables is more difficult than protecting tables from a survey sample since the sampling introduces ambiguity into the frequency counts and as a result it is more difficult to identify statistical units without response knowledge nor infer what the true count may be in the population.

This paper provides a review of common Statistical Disclosure Control (SDC) methods for protecting tabular outputs containing whole population counts from Censuses or register-based data. Since more invasive SDC methods are needed to protect against disclosure risk in a Census context, this has a negative impact on the utility of the data. The SDC methods will be compared using quantitative disclosure risk and information loss measures which focus on the effects on statistical analysis (see: Shlomo, 2007 and references therein for more details). The aim is to strike a balance between managing disclosure risk while maximizing the amount of information that can be released to users. The analysis of the SDC methods will be demonstrated on one typical table selected from the UK 2001 Census.

It is well known that Census and register-based data have errors due to data processing, coverage adjustments, non-response and edit and imputation procedures, although much effort is devoted to minimizing these errors. When assessing disclosure risk, it is essential to take into account measurement errors and the

protection that is already inherent in the data. For example, a quantitative measure of disclosure risk should take into account the amount of imputation and adjust parameters of the SDC methods accordingly to be inversely proportional to the imputation rate. This ensures that the data is not overly protected causing unnecessary loss of information. It should be noted that once statistical results are disseminated, they are typically perceived and used by the user community as accurate counts.

SDC methods implemented at Statistical Agencies for Census tables include pre-tabular methods, post-tabular methods and combinations of both. Pre-tabular methods are implemented on the microdata prior to the tabulation of the tables. The most commonly used method is record swapping between a pair of households matching on some control variables (Willenborg and de Waal, 2001 and references therein). This method has been used for protecting Census tables at the United States Bureau of the Census and the Office for National Statistics (ONS) in the United Kingdom. Record swapping can be seen as a special case of a more general pre-tabular method based on a Post-Randomization Method (PRAM) (Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998). This method changes values of categorical variables for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. PRAM can also be carried out in such a way as to ensure marginal distributions. In practice, Statistical Agencies prefer record swapping since the method is easy to implement and explain to users.

Post-tabular methods are implemented on the entries of the tables after they are computed and typically take the form of rounding, either on the small cells or on all entries of the tables. The method of small cell adjustments (random rounding) has been carried out on Census tables at the Australian Bureau of Statistics (ABS) and the UK ONS, and full random rounding has been carried out at Statistics Canada and Statistics New Zealand. A fully controlled rounding option has recently been added to the Tau-Argus SDC software package (Hundepool, 2002) although this has yet to be implemented for full scale Census outputs. Tau-Argus also has cell suppression modules among which we implement the heuristic Hypercube Method (Giessing, 2004) in order to cope with the large Census tables. A new technique for cell perturbation is the Controlled Tabular Adjustment (CTA) (Dandekar and Cox, 2002) which involves “imputing” values for the suppressed cells under additivity and other constraints. This method is still under development and will not be considered further in this paper.

Section 2 describes the table which will be used to illustrate the disclosure risk-data utility assessment and Section 3 the SDC methods applied. In Section 4 we examine the quantitative disclosure risk and information loss measures that will be implemented and carry out an analysis of the SDC methods. A discussion and conclusions are presented in Section 5.

2 Table Description

We examine a typical table extracted from one estimation area (EA) of the unperturbed 2001 UK Census data. The table is disseminated by Output Areas (OA) which are the smallest Census tracts that are published for the UK Census. The number of OAs in the EA is 1,487 and includes on average about 125 households. For each OA, the table is defined as follows (the number of categories is given in the parenthesis): Economic Activity (9) × Sex (2) × Long-Term Illness (2), i.e. a total of 36 categories. The table includes 317,064 individuals between the ages of 16 and 74 in 53,532 internal cells. The average cells size is 5.92 although the table is skewed with very large columns and very small columns. There are 17,915 (33.5%) zeros in the table and 14,726 (27.5%) cells with 1 or a 2.

3 SDC Methods

In this analysis, we will examine the following SDC methods:

3.1 Record Swapping

The most common pre-tabular method of SDC for frequency tables containing whole population counts is record swapping on the microdata prior to tabulation where variables are exchanged between pairs of households. In order to minimize bias, pairs of households are determined within strata defined by control variables, such as a large geographical area, household size and the age-sex distribution of the individuals in the households. In addition, record swapping can be targeted to high-risk households found in small cells of Census tables thereby ensuring that households that are most at risk for disclosure are likely to be swapped. In a Census context, geography variables are often swapped between households.

For this analysis, random record swapping was carried out on households from extracts of the 2001 UK Census at the following swapping rates: 10%, and 20%. The control variables that defined the strata were the number of persons in the household according to sex and three broad age groups and a "hard-to-count" index of the household based on the 1991 UK Census enumeration. The record swapping was carried out within a large geographical area (Local Authority) and households were swapped in and out of small geographical areas (Output Areas). In addition, targeted record swapping was carried out by defining an additional control variable based on a "flag" for the household that had at least one person in a small cell in a range of Census tables. On average, about 0.15% of the households selected for swapping were not swapped because no paired record was found for them. In general, those records would have to be swapped outside the large geographical area.

3.2 Rounding

The most common post-tabular method of SDC for Census tables is based on variations of rounding as follows:

Unbiased Random Rounding: Let $Floor(x)$ be the largest multiple k of the base b such that $bk < x$ for an entry x. In addition, define $res(x) = x - Floor(x)$. For an unbiased rounding procedure, x is rounded up to $(Floor(x) + b)$ with probability $\frac{res(x)}{b}$ and rounded down to $Floor(x)$ with probability $(1 - \frac{res(x)}{b})$. If x is already a multiple of b, it remains unchanged. Each cell is rounded independently in the table, i.e. a random uniform number u between 0 and 1 is generated for each cell. If $u < \frac{res(x)}{b}$ then the entry is rounded up, otherwise it is rounded down. As mentioned, the expectation of the rounding is zero and no bias should remain in the table. However, the realization of this stochastic process on a finite number of cells in a table may lead to overall bias since the sum of the perturbations (i.e. the difference between the original and rounded cell) going down may not equal the sum of the perturbations going up.

The method can be carried out on small cells only. In this case, margins of the tables are obtained by aggregating rounded and non-rounded cells, and therefore tables with the same population base will have different totals. While this provides ambiguity in the marginal totals, the users of Census tables generally object to inconsistent totals across tables. For full random rounding, margins are rounded separately from the internal cells because of the large number of perturbations and therefore tables are not additive.

The stochastic rounding methods are transparent and users can take the rounding into account when carrying out statistical analysis. The random rounding procedure (for all cells or only on small cells) is typically carried out independently for each cell based on a random draw, i.e. sampling with replacement. The algorithm however can be improved by preserving the stochastic unbiased properties but placing more control in the selection of the entries to round up or down. First the expected number of entries that are rounded up is predetermined (for the entire table or for each row/ column of the table). Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This process ensures a bias of zero and the rounded internal cells aggregate to the controlled rounded total. For this analysis, we carried out the full random rounding to base 3 and to base 5 under the following methods: independent rounding in each cell and semi-controlled to the overall total. In addition, we assess the impact of combining the SDC methods of record swapping and random rounding with respect to disclosure risk and information loss in the Census table.

Controlled Rounding: We implemented the controlled rounding feature in Tau-Argus on the Census table. The procedure uses linear programming techniques to round entries up or down and in addition ensures that all rounded entries add up to the rounded totals. It should be noted that the method is not unbiased and cells can jump a base in order to meet the constraints of the program. We implemented the method to base 3 and base 5.

Cell Suppression: Cell Suppression in Tau-Argus for frequency tables is determined by a minimum threshold for identifying the primary suppressions. Secondary suppressions are then chosen in order to avoid the recalculation of the primary suppressions through the margins. For an optimal selection of secondary suppressions, one can minimize a target function within a linear program framework subject to constraints of protection intervals for each suppressed cell. Because of the size of Census tables, we implemented the heuristic Hybercube method where the minimum threshold for primary suppressions was 3.

4 Analysis of SDC Methods

4.1 Disclosure Risk

The main type of disclosure risk arises from small cells in tables (or small cells appearing in potential slithers of differenced tables) as well as the amount and placement of the zeros. This can lead to identification and attribute disclosure when many tables are disseminated from one database.

Pre-tabular methods of disclosure control, and in particular record swapping, will not prevent small cells and therefore a quantitative disclosure risk measure is needed which reflects whether the small counts in tables are true values. The quantitative disclosure risk measure for assessing the impact of record swapping is the proportion of records in small cells that have not been perturbed. The perturbation comes from two sources: record swapping and imputation. In general, imputed records are viewed as protected records and therefore we need to take them into account in the quantitative risk measures. Imputation is typically carried out for item non-response, unit non-response and for Census coverage adjustments.

Let R_i represent the record i, I the indicator function having a value 1 if true and 0 if false, C_1 the set of cells with a value of 1, C_2 the set of cells with a value of 2, $n_{C_1 \cup C_2}$ the number of small cells with a value of 1 or 2. The disclosure risk measure is: $DR1 = \dfrac{\sum_{i \in C_1 \cup C_2} I(R_i \quad not \quad perturbed \quad or \quad imputed)}{n_{C_1 \cup C_2}}$. Table 1 presents

results of the disclosure risk measure *DR1* for the table in the analysis under record swapping.

Original	Random Swap		Targeted Swap	
	10%	20%	10%	20%
0.83	0.65	0.54	0.49	0.33

Table 1 Percentage of Records in Small Cells not Swapped or Imputed (*DR1*)

Based on Table 1, without any disclosure control method, imputation provides some protection to the small cells: 17% of the records in small cells in the table had some imputation carried out. For both swapping rates (10% and 20%), lower levels of disclosure risk are obtained, especially if records to be swapped are targeted from among unique records. In general, the probability that a small cell is indeed a true value for random record swapping is about (1-2×swapping rate). For example, for the 10% random record swapping in EA1, the probability of a true small value is approximately 0.8 (i.e. 1-2×0.10). The level of imputation was 0.17 and therefore we obtained a final probability of 0.65. The targeted record swapping at higher swapping rates gives better protection by lowering the probability of a true small value.

Post-tabular forms of rounding or cell suppression eliminate all small cells in the table and therefore disclosure risk is minimal with respect to attribute disclosure. In addition, in contrast to record swapping, the perception of disclosure risk is also minimal since no small cells appear in the tables.

Another disclosure risk measure comparable across all the SDC methods is the percentage of true zeros out of the total number of zeros (perturbed and not-perturbed) in the table. The more ambiguity introduced into the zero counts, the more the table is protected. Let C_0^{orig} be the number of true zero counts and C_0^{pert} the number of perturbed zero counts. The disclosure risk measure is defined as: $DR2 = \frac{C_0^{orig}}{C_0^{orig} + C_0^{pert}}$. Table 2 presents the *DR2* measures for the SDC methods evaluated for the Census table.

Record Swapping				Rounding		Cell Suppression
10%		20%		Base 3	Base 5	
Random	Targeted	Random	Targeted			
0.92	0.86	0.89	0.81	0.69	0.58	1.00

Table 2 Proportion of true zeros (*DR2*) in Table 4 of EA1

Based on Table 2, a zero in the table will be a true zero about 90% of the time for the record swapping whereas this proportion is greatly reduced when random or controlled rounding to base 3 or base 5. The cell suppression does not introduce any

ambiguity in the zero counts since these are not usually suppressed and the users know the true zeros.

Some forms of rounding can be deciphered by linking and differencing tables with common margins. To minimize this risk of disclosure, Statistics Agencies often disseminate only one set of geographies and variables, ensure minimum population thresholds and carry out auditing to evaluate the protection levels.

4.2 Information Loss

In this analysis we look at four main topics for measuring information loss: distortion to distributions, the impact on a measure of association (Cramer's V) for 2 dimensional tables, the impact on the variance of the cell counts and a "between" variance that is used in an ANOVA. All of the results are presented in Table 3.

When assessing information loss for cell suppression we need to implement a method of imputation for the suppressed cells which would typically be carried out by an average user. The simplest case would be to replace the suppressed cells by the average information loss in each row or column. More formally:

Let m_{ij} be a cell count in a two way table $i = 1,..., I$ rows and $j = 1,...J$ columns. Let marginal totals be defined as: $m_{i.}$ and $m_{.j}$. The margins appear in the table without perturbation unless they have a small value and are suppressed. In that case, we define the margin to take a value of 1 for the imputation scheme. Let z_{ij} be an indicator taking on the value of 1 if the cell was suppressed (primary or secondary) and a 0 otherwise. Each suppressed cell in row i is replaced by $\hat{y}_i = \frac{m_{i.} - \sum_{j=1}^{J} m_{ij}(1 - z_{ij})}{\sum_{j=1}^{J} z_{ij}}$. For example: Two cells are suppressed in a row where the known marginal total is 500. The total obtained by adding up non-suppressed cells is 400, and therefore the total information loss in the row is 100. Each of the two suppressed cells is replaced with a value of 50.

Information loss will be defined as follows:

- **Distance Metric**

We examine distortions to the internal and marginal cells of the Census table. Since the basic unit for most Census tables are small geographies, i.e. OAs, a measure of distortion at this level of geography is preferred. The distance metric between original and protected cells of the table (including zero cells) are calculated separately for each OA. The final utility measure is the overall average of the distance metric across the OAs.

Following the notation of Gomatam and Karr (2003), let D^k represent a table for OA *k*, $D^k(c)$ be the cell frequency *c* in the table and n_{OA} the number of OAs, i.e. $n_{OA} = 1{,}487$. We define the Hellinger's Distance metric as follows:

$$HD(D_{orig}, D_{pert}) = \frac{1}{n_{OA}} \sum_{k=1}^{n_{OA}} \sqrt{\sum_{c \in k} \frac{1}{2} \left(\sqrt{D^k_{pert}(c)} - \sqrt{D^k_{orig}(c)} \right)^2}$$

In addition, we examine distortions to the marginal totals of the Census table for both rows and columns. Denote by M the margin under consideration, n_M the number of categories in the margin and N^l the total number of persons in the *l-th* category of margin M. The Hellinger's Distance metrics is:

$$HDM(N_{orig}, N_{pert}) = \sum_{l=1}^{n_M} \sqrt{\frac{1}{2} \left(\sqrt{N^l_{pert}} - \sqrt{N^l_{orig}} \right)^2}$$

- **Impact on Measures of Association**

A very important statistical tool that is frequently carried out on contingency tables is the Chi-Square test for independence based on the Pearson Chi-Squared Statistic χ^2 which tests the null hypothesis that the criteria of classification, when applied to a population, are independent. The Pearson Statistic for a two-dimensional table $i = 1,..,I$ and $j = 1,..,J$ is defined as: $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ where under the null hypothesis of independence: $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$, $n_{i.}$ is the marginal row total and $n_{.j}$ is the marginal column total.

In order to assess the impact of the SDC methods on tests for independence, the Pearson statistic obtained from a perturbed contingency table is compared to the Pearson statistic obtained from the original contingency table. In particular, we focus on the measure of association, Cramer's V defined as: $CV = \sqrt{\frac{\chi^2 / n}{\min(I-1),(J-1)}}$.

The information loss measure is the percent relative difference:

$$RCV(D_{orig}, D_{pert}) = \frac{100 \times (CV(D_{pert}) - CV(D_{orig}))}{CV(D_{orig})} .$$

For this analysis, the rows of the table are the OAs and the columns of the table are the Economic Activity × Sex × Long-Term Illness indicator. It should be noted that random rounding rounds the margins separately from the internal cells and tables are not additive. Nevertheless, using a standard statistical package, the expected cell

frequency e_{ij} is calculated by aggregating internal cells and not obtained from known margins. A large Cramer's V represents a high level of association between the rows and the columns of the two-way table.

- **Impact on Variance of Cell Counts**

SDC methods impact on the variances that are calculated for estimates based on the frequency tables. The focus in this analysis is on the variance of the average cell count calculated at the OA level of geography in the table. The overall information loss measure is obtained by the percent difference between the average variance across all of the OAs for the original table and the same average variance for the perturbed table.

Let: $V(D_{orig}) = \frac{1}{n_{OA}} \sum_{k=1}^{n_{OA}} \frac{1}{n_k - 1} \sum_{c \in k} (D_{orig}^k(c) - \overline{D}_{orig}^k)^2$ where n_k is the number of columns, i.e. $n_k = 36$, and $V(D_{pert})$ similarly calculated. The utility measure is the percent relative difference: $RDV(D_{orig}, D_{pert}) = \frac{100 \times (V(D_{pert}) - V(D_{orig}))}{V(D_{orig})}$.

- **Impact on "Between" Variance**

We assess the impact of SDC methods on the goodness of fit criterion R^2 of a regression analysis or ANOVA and in particular on the "between" variance which is used as a component in R^2. For example, in an ANOVA, we test whether a continuous dependent variable has the same means within groupings defined by categorical explanatory variables. The goodness of fit criterion R^2 is based on a decomposition of the variance of the mean of the dependent variable. The total sum of squares *SST* can be broken down into two components: the "within" sum of squares *SSW* which measures the variance of the mean of the target variable within groupings which are defined by combining explanatory variables and the "between" sum of squares *SSB* which measures the variance of the mean of the target variable between the groupings. R^2 is the ratio of *SSB* to *SST*. By perturbing the statistical data, the groupings may lose their homogeneity, *SSB* becomes smaller, and *SSW* becomes larger. In other words, the proportions within each of the groupings shrink towards the overall mean. On the other hand, *SSB* may become artificially larger showing more association within the groupings than in the original variable.

We define information loss based on the "between" variance of a proportion on cell *c*: Let $P_{orig}^k(c)$ be a target proportion for cell *c* in OA *k*, i.e. $P_{orig}^k(c) = \frac{D_{orig}^k(c)}{\sum_{c \in k} D_{orig}^k(c)}$

and let $P_{orig}(c) = \frac{\sum_{k=1}^{n_{OA}} D_{orig}^{k}(c)}{\sum_{k=1}^{n_{OA}} \sum_{c \in k} D_{orig}^{k}(c)}$ be the overall proportion across all the OAs of the table. The "between" variance for the proportion is defined as: $BV(P_{orig}(c)) = \frac{1}{n_{OA} - 1} \sum_{k=1}^{n_{OA}} (P_{orig}^{k}(c) - P_{orig}(c))^2$ and the information loss measure is:

$$BVR(P_{pert}(c), P_{orig}(c)) = \frac{100 \times (BV(P_{pert}(c)) - BV(P_{orig}(c)))}{BV(P_{orig}(c))}.$$

For this analysis, we chose the proportion of full-time male students with no long-term illness.

Based on Table 3, the greatest impact on distortions to cell counts for both internal and marginal cells is the random rounding to base 5. Putting some control in the random rounding procedure seems to cause slight improvements and indeed the full controlled rounding has less distortions to cell counts for both base 3 and base 5. As expected, rounding to base 3 has less distortions compared to rounding to base 5. The cell suppression with the simple imputation method has the least distortions since marginal totals and original cell counts above the value of 3 (not secondary suppressed) are disseminated without any perturbation. Record swapping has less distortion to distributions than the rounding methods. The distortions are greater as the swapping rates increase. Targeted record swapping has larger distance metrics for the internal cells but not necessarily for the OA margins since records were swapped across OAs in both cases. It should be noted that for the margin based on sex, long-term illness and economic activity, the record swapping did not cause any distortion. This is likely due to the control variables that were used for selecting pairs to swap geographies. In general, there is more distortion when unique records are targeted for swapping. When combining a rounding procedure with record swapping, all distance metrics are higher. The increased distortion to distributions therefore needs to be weighed against the extra protection that record swapping may provide to Census tables by introducing ambiguity when differencing and linking tables.

Table 3 also demonstrates the loss in association and attenuation when swapping records across geographical areas. The two-way Census table examined is leaning more towards independence since the counts are "flattening" out in the table (this is seen by the negative sign of the *RCV* measure). With higher swapping rates the loss in association is more severe. Targeted record swapping has less impact on the loss of association compared to the random record swapping. We also see in the table that the rounding procedures have the opposite effect. By eliminating small cells through the rounding procedures and introducing more zeros into the table, the level of association based on the observed cell counts has artificially increased. This effect

Method	HD	HDM (Columns)	HDM (Rows)	RCV Orig. Cramer's V (0.121)	RDV Orig. Avg. Variance (188.3)	BVR Orig. Between Variance (0.000233)
RR Base 3	2.03	1.48	6.36	11.58	0.52	11.4
RR Base 3 (controlled to total)	2.04	2.27	5.19	11.88	0.54	13.1
Controlled Rounding Base 3	1.95	0.07	1.53	9.97	0.39	12.9
RR Base 5	3.02	3.39	9.87	27.52	1.64	36.6
RR Base 5 (controlled to total)	3.03	3.26	5.43	27.65	1.62	39.4
Controlled Rounding Base 5	2.58	0.09	3.20	26.93	1.27	34.5
Cell Suppression	0.42	0	0	0.22	-0.04	-0.64
Swap Random 10%	1.39	0	2.46	-3.65	-1.31	-4.82
Swap Random 20%	1.98	0	3.59	-6.27	-2.10	-8.25
Swap Targeted 10%	1.58	0	2.38	-1.93	-0.59	-3.49
Swap Targeted 20%	2.19	0	3.16	-4.37	-1.51	-7.61
Swap 10% RR Base 3	2.53	2.17	6.90	10.39	-0.78	9.45
Swap 20% RR Base 3	2.91	1.86	7.16	7.66	-1.57	5.60

Table 3 Results of Information Loss Measures on Census Table

however is less severe with the controlled rounding method. When combining rounding procedures with record swapping, there are opposing effects on Cramer's V and therefore the *RCV* is smaller compared to the *RCV* on the rounding procedures alone.

These same conclusions are seen with respect to the impact on the variance of the average cell counts (*RDV)* and the "between" variance (*BVR*). We obtain a clear pattern of decreasing variances (as noted by the negative values) as higher swapping

rates are introduced, i.e. the cell counts are "flattening" out for the *RDV* and the proportions within groups defined by OAs are moving towards the overall mean for the *BVR*. The targeted record swapping has slightly less reduction in the two variances compared to the random record swapping. As seen for the analysis on the Cramer's V above, the opposite effect occurs with the rounding procedures and the two variances are increasing. There is a slight increase in the variances with the semi-controlled rounding but less of an increase for the full controlled rounding. The impact on the variance when combining rounding procedures with record swapping depends on the direction and magnitude of the variances of each procedure separately, although it is clear that the opposing effects are cancelling out.

5 Discussion

In this analysis, we examine some common approaches of SDC for Census tabular outputs: pre-tabular methods based on variations of record swapping and post-tabular methods based on forms of rounding and cell suppression. In addition, we assessed the impact when combining SDC methods.

From this analysis, it was shown that using record swapping as a sole SDC method for Census tables results in high probabilities that small cells in tables are true values and can be identified. Targeted record swapping lowers the disclosure risk but there is more distortion to distributions with respect to distance metrics. Higher swapping rates raise the level of protection but also cause more distortion to the data. The overall distortion on cell counts is higher with the rounding procedures compared to the swapping methods. Placing controls in the rounding procedure preserves additivity and causes less distortions to cell counts and therefore raises the utility of the tables. It should be noted that rounding procedures protect against the perception of disclosure risk compared to record swapping where the effects are hidden to users. Combining rounding with record swapping raises the level of protection but increases the loss of utility to the Census tables. For some statistical analysis, the combination of record swapping and rounding may balance to some degree opposing effects that the methods have on the utility of the tables. For example, record swapping "flattens" out cell counts, reduces measures of association and homogenizes the data while rounding procedures introduce more dependencies, increase measures of association and raises the levels of dispersion. These conclusions found for the record swapping and rounding procedures are consistent across all tables containing whole population counts and not just the particular Census table that was used for this analysis.

We have demonstrated in this paper how a Statistical Agency should carry out an assessment of SDC methods by examining both sides of the SDC decision problem: managing disclosure risk while maximizing the utility and quality of the outputs. The final decision on what SDC methods to employ depends on whether the disclosure

risk is below tolerable thresholds and if the utility of the outputs meets the demands for "fit for purpose" data by the user community. SDC methods should be combined, adapted and modified in order to ensure higher utility in the outputs. A correct balance must be found between the use of non-perturbative transparent SDC methods and perturbative SDC methods which have hidden effects and introduce bias that cannot be accounted for. Clear guidance and quality measures need to be disseminated with the Census tables in order to inform users of the impact of the SDC methods and how to analyze disclosure controlled statistical data.

Future dissemination strategies for Censuses will include more use of flexible table generating software where users can design and generate their own Census tables. Therefore, the development of SDC methods needs to be directed to these types of online dissemination strategies. Improved GIS systems may advance the research for developing SDC methods that protect nested geographies thus allowing more flexibility for online dissemination. Finally, more reliance on safe settings, remote access and license agreements provides alternative SDC strategies which limit the access to the data to sponsored researchers, especially when dealing with highly disclosive Census sample microdata and Origin-Destination tables.

References

Dandekar, R.H. and Cox, L. (2002) Synthetic Tabular Data – an Alternative to Complementary Cell Suppression. *Unpublished manuscript*.

Giessing, S. (2004) Survey on Methods for Tabular Data Protection in Argus. In Domingo-Ferrer, J. and Torra, V. (eds.): *Privacy in Statistical Databases, LNCS 3050,* Springer-Verlag.

Gomatam, S. and Karr, A. (2003) Distortion Measures for Categorical Data Swapping. *Technical Report Number 131*, National Institute of Statistical Sciences.

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

Hundepool, A. (2002) The CASC Project. In Domingo-Ferrer, J. (eds.): *Inference Control in Statistical Databases: From Theory to Practice, LNCS* 2316, Springer-Verlag.

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

Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, 155. New York: Springer-Verlag.