Statistical Disclosure Control Methods for Census Frequency Tables

Natalie Shlomo

Abstract

This paper provides a review of common statistical disclosure control (SDC) methods implemented at Statistical Agencies for standard tabular outputs containing whole population counts from a Census (either enumerated or based on a register). These methods include record swapping on the microdata prior to its tabulation and rounding of entries in the tables after they are produced. The approach for assessing SDC methods is based on a disclosure risk–data utility framework and the need to find the balance between managing disclosure risk while maximizing the amount of information that can be released to users and ensuring high quality outputs. To carry out the analysis, quantitative measures of disclosure risk and data utility are defined and methods compared. Conclusions from the analysis show that record swapping as a sole SDC method leaves high probabilities of disclosure risk. Targeted record swapping lowers the disclosure risk, but there is more distortion to distributions. Small cell adjustments (rounding) give protection to Census tables by eliminating small cells but only one set of variables and geographies can be disseminated in order to avoid disclosure by differencing nested tables. Full random rounding offers more protection against disclosure by differencing, but margins are typically rounded separately from the internal cells and tables are not additive. Rounding procedures protect against the perception of disclosure risk compared to record swapping since no small cells appear in the tables. Combining rounding with record swapping raises the level of protection but increases the loss of utility to Census tabular outputs. For some statistical analysis, the combination of record swapping and rounding balances to some degree opposing effects that the methods have on the utility of the tables.

(Supersedes Working Paper M06/03)

Southampton Statistical Sciences Research Institute
Methodology Working Paper M07/04

Statistical Disclosure Control Methods for Census Frequency Tables

Natalie Shlomo

Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom

Department of Statistics, Hebrew University, Mt. Scopus, Jerusalem, Israel

Summary

This paper provides a review of common statistical disclosure control (SDC) methods implemented at Statistical Agencies for standard tabular outputs containing whole population counts from a Census (either enumerated or based on a register). These methods include record swapping on the microdata prior to its tabulation and rounding of entries in the tables after they are produced. The approach for assessing SDC methods is based on a disclosure risk–data utility framework and the need to find the balance between managing disclosure risk while maximizing the amount of information that can be released to users and ensuring high quality outputs. To carry out the analysis, quantitative measures of disclosure risk and data utility are defined and methods compared. Conclusions from the analysis show that record swapping as a sole SDC method leaves high probabilities of disclosure risk. Targeted record swapping lowers the disclosure risk, but there is more distortion to distributions. Small cell adjustments (rounding) give protection to Census tables by eliminating small cells but only one set of variables and geographies can be disseminated in order to avoid disclosure by differencing nested tables. Full random rounding offers more protection against disclosure by differencing, but margins are typically rounded separately from the internal cells and tables are not additive. Rounding procedures protect against the perception of disclosure risk compared to record swapping since no small cells appear in the tables. Combining rounding with record swapping raises the level of protection but increases the loss of utility to Census tabular outputs. For some statistical analysis, the combination of record swapping and rounding balances to some degree opposing effects that the methods have on the utility of the tables.

Key words: Disclosure risk measures, Data utility measures, R-U confidentiality map

R?sum?

Cet article propose une revue des m?thodes de contr?le de la divulgation statistique (CDS) mises en place par les agences statistiques lors de production de tableaux statistiques d?riv?s de donn?es des recensements. Ceci inclue des techniques de pr?-traitements du type « hybridation » - ?change partiel d'information entre individus - ou des m?thodes d'arrondis effectu?es apr?s la production des tableaux. L'approche des m?thodes CDS pr?sent?e insiste sur la n?cessit? de trouver un ?quilibre entre la gestion du risque de divulgation tout en maximisant la quantit? d'information qui peut ?tre fournie aux utilisateurs. Des mesures quantitatives de risques et de degr? d'utilit? sont propos?s et compar?es. Les conclusions des analyses montrent que la technique d'hybridation peut conduire ? des cas de divulgations pour les tableaux pr?sentant des cellules ? faibles effectifs. La m?me technique utilis?e sur des individus "cibl?s" diminue le risque mais au d?triment des distributions statistiques. La m?thode de l'arrondi prot?ge les tableaux en ?liminant les cellules ? faibles effectifs mais un seul type de variables et g?ographie doivent ?tre publi?s pour ?viter le risque de divulgation par diff?renciation quand les tableaux sont li?s les uns aux autres. L'arrondi al?atoire donne plus de protection contre le risque par diff?renciation mais certaines cellules peuvent ?tre reconstruites par comparaison avec les marges. Les techniques d'arrondis prot?gent contre la perception du risque mieux que l'hybridation.. Combiner hybridation et arrondi augmente le niveau de protection mais augmente la perte de qualit? quant ? l'utilit? des sorties statistiques.

Dans certaines analyses statistiques, les deux approches utilis?es simultan?ment peuvent cependant produire un effet ?quilibr?.

1 Introduction

Disclosure risk occurs when there is a high probability that an intruder can re-identify an individual in released statistical outputs and confidential information may be obtained. In order to protect against disclosure risk, statistical disclosure control (SDC) methods are applied to outputs. Standard outputs include tabular data (frequency counts or aggregated data) and micro-data typically from samples and released under license. This paper provides a review of common SDC methods for protecting standard tabular outputs containing whole population frequency counts from Censuses or register-based data.

Protecting Census tables is more difficult than protecting tabular data from a survey sample. The sampling a priori 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. Moreover, tabular data from samples are typically weighted counts where sampling weights vary between units because of differential selection probabilities and non-response adjustments. Therefore, the number of contributors to a cell is not always known. Small sample counts in tables are often suppressed because of low quality and inefficiency and this solves the problem for SDC. For these reasons, Statistical Agencies put more resources into the protection of tabular data from whole population counts.

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. It is well known that Census 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 Census results are disseminated, they are typically perceived and used by the user community as accurate counts.

The main disclosure risk in a Census context comes from small counts, i.e. ones and twos, since these can lead to re-identification. Indeed, the amount and placement of the zeros in the table determines whether new information can be learnt about an individual or a group of individuals. Therefore, SDC methods for Census tabular data should not only protect small cells in the tables but also introduce ambiguity and uncertainty into the zero values.

SDC methods for Census tables that are typically implemented at Statistical Agencies 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). 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 adds "noise" to categorical variables by changing values of categories 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 and because it is a stochastic perturbation, users can make use of the probability transition matrix in their statistical analysis. This method however has yet to be implemented for a large scale Census. In practice, Statistical Agencies prefer record swapping since the method is easy to implement and marginal distributions are preserved exactly on higher aggregations of the data. It should be noted that Statistical Agencies do not typically release parameters of the SDC methods, i.e. swapping rates or probability transition matrices, in order to minimize the chance of deciphering the perturbation process.

Post-tabular methods are implemented on the entries of the tables after they are computed and typically take the form of random rounding, either on the small cells of the tables or on all entries of the tables. The method of small cell adjustments (rounding) has been carried out on the 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. Within the framework of developing the SDC software package, Tau Argus, a fully controlled rounding option has been added (Hundepool, 2002). 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. However, the controlled rounding option is not able to cope with the size, scope and magnitude of Census tabular outputs at this time. Other post-tabular methods include cell suppression or some form of random perturbation on the cells of the Census tables. Cell suppression is not typically used in a Census context because of the large number of tables that need to be consistently suppressed. Cell perturbation based on a stochastic mechanism (for example, the method used in the 1991 UK Census was to add 0, ±1 to each cell count in a table according to prescribed probabilities), is basically the same as record swapping except with the disadvantage that internal cells and marginal totals are inconsistent across Census tables. Therefore these methods will not be considered in this paper.

Few evaluation studies have been carried out on the impact of SDC methods on disclosure risk and the resulting utility and quality of Census tables. Carter (2001) implemented a comparative study on the risk of attribute disclosure for Census tables (i.e. the probability of obtaining a one on a margin of a table) for the methods: random cell perturbation described above, random record swapping and random rounding. The study was based on distributional assumptions on hypothetical population counts and average cell sizes in Census tables. Gomatam, Karr and Sanil (2003) provided an analysis of categorical data swapping on real data sets where parameters of the data swapping were determined by examining the trade-off between balancing the disclosure risk measured by the percent of un-swapped records and utility measured by a distance metric between original and perturbed distributions. Boyd and Vickers (1999) assessed the impact of record swapping on distortions to distributions.

In this paper, we propose quantitative disclosure risk and data utility measures and illustrate how a Statistical Agency should carry out a comprehensive assessment of different SDC methods for Census tabular outputs based on a disclosure risk–data utility framework as described in Willenborg and De Waal (2001) and Duncan, Keller-McNulty, and Stokes (2001). Utility is assessed by analyzing the impact of SDC methods on statistical analysis and new measures are introduced that quantify these effects. Moreover, we demonstrate how SDC methods should be modified and combined in order to increase the utility of the data without increasing disclosure risk. 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 real data sets from the UK 2001 Census.

Section 2 provides a brief outline of the relevant types of disclosure risk in a Census context where many tables are disseminated from a single database containing whole population counts. Section 3 outlines the SDC methods that are examined and Section 4 details the data and Census tables that are used in the analysis. Sections 5 and 6 define the quantitative disclosure risk and data utility measures with results of the assessment of the SDC methods. A discussion and conclusions from the analysis are presented in Section 7.

2 Types of Disclosure Risk in Census Tabular Outputs

Disclosure risk in Census tables include the following:

Individual attribute disclosure - An individual can be identified on the basis of some of the variables spanning the table and a new attribute revealed about the individual, i.e. for tabular data, this means that there is a one in a margin of the table. Identification is a necessary pre-condition for attribute disclosure and therefore should be avoided. In a Census context where many tables are released, an identification made in a lower dimensional table will lead to attribute disclosure in a higher dimensional table. For example, in data taken from the 2001 UK Census, out of 184 persons living in a particular Output Area, unique persons were found on the following sex-age groups: males 50-59, males 85 and over and females 60-64. In another table, these same individuals were further disseminated according to health variables and it was learnt that the single male aged 50-59 and the single female aged

60-64 have good or fairly good health and have no limiting long-term illness, the single male aged 85 and over has poor health and has a limiting long-term illness.

Group attribute disclosure - If there is a row or column that contain mostly zeros and a small number of non-zero cells, then one can learn a new attribute about a group of individuals and also learn about the group of individuals who do not have this attribute. This type of disclosure risk does not require individual identification. For example, all elderly persons above the age of 65 in a particular Output Area have limiting long-term illnesses. All persons below that age do not have long-term illnesses.

Disclosure by differencing – Two tables that are nested may be subtracted one from the other resulting in a new table containing small cells and the above disclosure risk scenarios would apply. For example, a table containing the elderly population in private households may be subtracted from a table containing the total elderly population, resulting in a table of the elderly in communal establishments. This table is typically very sparse compared to the two original tables.

Disclosure by linking tables – Since all Census tables are disseminated from one data set, they can be linked though common cells and common margins thereby increasing the chances for revealing SDC methods and original cell counts. For example, assume an SDC method of random rounding to base 3 and several tables are disseminated containing a particular cell with an original value of 1. If the small cell is rounded down more times than it is rounded up across the tables, then it can be assumed that the original count was a one. Small cell adjustments (rounding) where the marginal totals are obtained by aggregating rounded and non-rounded cells are especially problematic since if there are no small cells in the table, exact marginal totals are obtained. These exact marginal totals can be used to decipher counts on higher dimensional tables which may contain small rounded cells. It should be noted that in a Census context where there are many tables disseminated from a common dataset, there is currently no simultaneous rounding procedure for tables that can be linked across common cells.

Perception of disclosure risk - This type of disclosure risk is particularly important to Statistical Agencies who are concerned that response rates may drop for Censuses and surveys if the public perceive that the Agency is not protecting their confidentiality.

To protect against attribute disclosure, SDC methods should limit the risk of identification and also introduce ambiguity into the zero counts. To avoid disclosure by differencing, often only one set of variables and geographies are disseminated with no possibilities for overlapping categories. To avoid disclosure by linking tables, margins and cells of tables should be consistent. To avoid the perception of disclosure risk, Statistical Agencies often employ transparent and visible SDC methods and resources are directed to ensure that the public is informed about the measures taken to protect confidentiality.

3 Common SDC Methods for Census Tables

3.1 Record Swapping

The most common pre-tabular method of SDC for frequency tables is record swapping on the microdata prior to tabulation where variables are exchanged between pairs of households. In order to minimize bias, pairs of households are 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 the following reasons:

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

For this analysis, random record swapping was carried out on households from extracts of the 2001 UK Census at the following swapping rates: 1%, 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 (LA)) and households were swapped in and out of small geographical areas (Output Areas (OA)). 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 one of the Census tables under evaluation (see Section 4). 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 (LA).

Table 1 presents advantages and disadvantages of record swapping as a pre-tabular method of SDC for Census tabular outputs.

[PLACE TABLE 1 HERE]

3.2 Rounding

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

Small Cell Adjustments: The method is an unbiased random rounding on small cells only. Let *x* be a small cell and 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. The expected value of the rounded entry is the original entry since:

$$(x - Floor(x)) \times (1 - \frac{res(x)}{b}) + (x - (Floor(x) + b)) \times \frac{res(x)}{b} = 0 \quad .$$

Each small cell is rounded independently in the table, i.e. a random uniform number *u* between 0 and 1 is generated for each cell. If $u < \frac{res(x)}{b}$ then the entry is rounded up, otherwise it is rounded down. 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.

When only small cells are rounded, 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. The confidence interval for the expected differences between perturbed totals and true totals is a function of the number of small cells that are rounded. Figure 1 presents the confidence interval when rounding cells to base 3.

[PLACE FIGURE 1 HERE]

Full Random Rounding: Unbiased random rounding is carried out on all entries in the table. This is implemented as described above for the small cells after first converting the entries *x* to residuals of the rounding base *res(x).* Because of the large number of perturbations in the table, margins are rounded separately from internal cells 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. The advantages and disadvantages of rounding methods for protecting Census tabular outputs are presented in Table 2.

[PLACE TABLE 2 HERE]

For this analysis, we carry out both the small cell adjustments and the full random rounding under the following methods: independent rounding in each cell; semi-controlled to the overall total; semi-controlled to the OA totals in the tables. In addition, we assess the impact of combining the SDC methods based on record swapping and rounding with respect to disclosure risk and utility in the Census tables.

4 Data Used in the Analysis

To carry out the disclosure risk–data utility analysis, extracts of unperturbed 2001 UK census data were obtained from three Estimation Areas (EA):

EA1 - 437,744 persons, 182,337 households, 1,487 Output Areas (OA)

EA2 - 507,049 persons, 216,502 households, 1,755 Output Areas (OA)

EA3 - 523,464 persons, 215,858 households, 1,800 Output Areas (OA)

For each Estimation Area (EA), five standard census tables were defined (the number of categories of the variable is in parenthesis):

(1) Religion(9) × Age-Sex(6) × OA

(2) Travel to Work(12) × Age-Sex(12) × OA

(3) Country of Birth (17) × Sex (2) × OA

(4) Economic Activity (9) × Sex (2) × Long-Term Illness (2) × OA

(5) Health status (5) × Age-Sex (14) × OA

As an example, the characteristics of the five tables for EA1 are presented in Table 3.

[PLACE TABLE 3 HERE]

5 Disclosure Risk Measures

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 inhibit small cells from appearing in tables and therefore a quantitative disclosure risk measure is needed which reflects whether the ones and twos 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, $|C_1 \cup C_2|$ the number of small cells with a value of 1 or 2. The disclosure risk measure is: $DR = \frac{\sum_{i \in C_1 \cup C_2} I(R_i \quad not \quad perturbed \quad or \;\; imputed)}{|C_1 \cup C_2|}$. Table 4 presents results of the disclosure risk measure for two EAs.

[PLACE TABLE 4 HERE]

Based on Table 4, without any disclosure control method, imputation provides some protection to the small cells: 16% of the records in small cells in EA1 (and EA3) had some imputation carried out and 21% in EA2. There is little impact on disclosure risk for the 1% random and targeted record swapping. In either case, there is still about an 80% chance that a small cell in a table (a one or a two) is a true value. This leaves a high probability that small cells can be identified in Census tables. For the other 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 0.8 (i.e. 1-2×0.10). The level of imputation was 0.16 and therefore we obtained a final probability of 0.634. 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 eliminate all small cells in the table and therefore disclosure risk is minimal with respect to attribute disclosure. In addition, ambiguity is introduced into the zeros of the table since small cells can be rounded down to zero in the rounding procedures. It is important to note in contrast to record swapping that the perception of disclosure risk is also minimal since no small cells appear in the tables. Some forms of rounding can be deciphered by linking and differencing tables with common margins. To minimize this risk of disclosure, the following steps are often undertaken at Statistics Agencies:

- Only one set of geographies and variables are disseminated, for example, it is not possible to publish cell counts for ages 16-19 and also ages 15-19 since this leads to disclosure by differencing. Also, population thresholds are determined below which whole tables are suppressed.

- Tables that have undergone stochastic SDC rounding methods are audited. The marginal totals are also rounded in order to avoid linking tables with common margins.

Therefore, for this analysis we assume that the rounding procedures provide good protection and only the dimension of utility is examined in the disclosure risk–data utility framework.

6 Data Utility Measures

Data utility measures can be divided into several subsets according to the statistical analysis that is to be carried out: (1) Measuring distortions to distributions; (2) Impact on the variance of estimates; (3) Impact on measures of association (tests for independence between categorical variables) and other goodness of fit criteria; (4) Impact on ranks and correlations.

This section will demonstrate the use of data utility measures on the 2001 UK Census tables based on different types of analysis. The three EAs used in the analysis have similar results and therefore only representative tables and figures are presented.

6.1 Measuring Distortions to Distributions

6.1.1 Distance Metrics on Internal Cells of the Tables

Distance metrics are used to measure distortions to distributions as a result of applying SDC methods. Some useful metrics were presented in Gomatam and Karr (2003). Since the basic unit for most Census tables are small geographies, i.e. Output Areas (OA), a measure of distortion at this level of geography is preferred. The distance metrics between original and protected distributions of the tables are calculated separately for each OA. The final utility measure is the overall average of the distance metrics across the OAs. In this section we examine distance metrics for distortions to distributions of the internal cells of the tables. Marginal totals are examined in Section 6.1.3.

Following the notation of Gomatam and Karr (2003), let D^k represent a table for OA k and let $D^k(c)$ be the cell frequency c in the table. Let $|OA|$ be the number of OAs in the EA. The distance metrics are:

- Hellinger's Distance:

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

- Average Absolute Distance per Cell:

$$AAD(D_{orig}, D_{pert}) = \frac{1}{|OA|}\sum_{k=1}^{|OA|}\frac{\sum_{c\in k}|D^k_{pert}(c) - D^k_{orig}(c)|}{|k|} \text{ where}$$

$$|k| = \sum_c I(c \in k) \text{ the number of non-zero cells in the } k^{th} \text{ OA}$$

The HD distance is based on Information Theory. It is heavily influenced by small cells. The AAD is more intuitive and describes the average absolute difference per non-zero cell of an OA. Other distance metrics based on relative differences are undefined for the case when the original cell count is zero and therefore we do not examine these in this paper. Table 5 presents results of the utility measures for tables in EA1 for the different SDC methods.

[PLACE TABLE 5 HERE]

Based on Table 5, the distance metrics are low for the 1% swapping rate and increase for the higher swapping rates. The utility of the data is compromised for large swapping rates. The measure of *AAD* quantifies by how much non-zero cells are perturbed on average for each OA. For example, for the random record swapping, each non-zero cell is perturbed by about 0.7 for the 10% swap and about 1.0 for the

20% swap. Similarly, for the targeted record swapping, each cell is perturbed by about 0.8 for the 10% swap and about 1.2 for the 20% swap. Targeted record swapping has higher distance metrics which demonstrate that more distortion occurs when the unique records are targeted for swapping. It is important to note that distortions to distributions caused by record swapping are hidden to the user. Census tables are all consistent but counts are perturbed and confidence intervals for the true counts cannot be calculated and provided to the users in order to assist in their analysis.

The small cell adjustments on the original data according to Table 5 cause slightly less distortion to distributions compared to the 10% random record swapping according to the *AAD* measure, but more distortion to distributions according to the *HD* measure. For example, small cell adjustments on tables in EA1 have an *AAD* of 0.629 and an *HD* of 5.272. In comparison, the 10% random record swapping on tables in EA1 have an *AAD* of 0.722 and an *HD* of 3.714. This is due to the fact that the *HD* distance metric is influenced more by small cells in the distributions than the *AAD* distance metric.

When combining rounding procedures 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 the Census tables by introducing ambiguity when differencing and linking tables.

There is little difference when examining internal cells of tables based on these distance metrics between the independent rounding procedures and semi-controlling for the individual OA totals (i.e., small cell adjustments (SCA) compared to benchmarked small cell adjustments (BSCA) and full random rounding (RR) compared to benchmarked full random rounding (BRND)). However, the benchmarked method also preserves some of the additivity of the table and therefore increases utility.

6.1.2 Aggregating Internal Perturbed Cells

In this section, a distance metric is defined for differences in sub-totals that are obtained by aggregating internal perturbed cells. The difference for a sub-total $N^k(C')$ where $N^k(C') = \sum_{c \in C'} D^k(c)$ is: $AD(N^k_{orig}, N^k_{pert}) = N^k_{pert}(C') - N^k_{orig}(C')$.

One of the main uses of lower level geography (OA) tables is to aggregate internal cells in order to obtain sub-totals for non-standard geographies, such as school districts. The lower level tables are typically used as building blocks to construct higher level (non-standard) geographies. The tables at the lower level, however, are highly perturbed and therefore aggregating lower level data compounds the effects of SDC methods.

In order to evaluate the range of the differences between perturbed and original sub-totals (*AD*) for specific Census target variables, the statistical graphing tool of a box plot is used. For unbiased rounding schemes, the average and median of the *AD* measures are centered at zero. The length of the box and the length of the whiskers gives an indication of how wide spread the perturbed sub-totals are from their original sub-totals.

For this analysis, ten consecutive OAs in each EA were aggregated for a specific target variable and the differences between the true sub-totals and the perturbed sub-totals (*AD)* were calculated. Figure 2 presents box plots of the differences in the sub-totals (*ADs)* for EA1 based on the number of Males born in Western Europe within ten consecutive groupings of OAs under the different methods of record swapping.

[PLACE FIGURE 2 HERE]

In Figure 2, there is almost no difference between the aggregated original and perturbed sub-totals for the 1% swapping rate. The targeted 1% record swapping has slightly more differences in the perturbed totals compared to the 1% random record swapping. The 10% and 20% swapping rates have higher differences between original and perturbed sub-totals with wide spread whiskers. The maximum difference reaches as high as ± 15 which is 61% of the average original sub-total of 24.6.

In Figure 3, we examine box plots of the differences between sub-totals (*ADs)* for the rounding methods. It is clear that the boxes are smaller when semi-controlling the rounding procedures for the overall total (controlled small cell adjustments (CSCA) and controlled random rounding (CRND)), but when semi-controlling for each individual OA (benchmarked small cell adjustments (BSCA) and benchmarked random rounding (BRND)) the boxes are about the same as if no controls are carried out. In addition, there appears little difference between small cell adjustments and full rounding of all cells. This is because about 60% of the cell values for this particular target variable across the OAs were small cells and therefore were rounded for both the full rounding and small cell adjustments procedures. In general, we would expect that the differences between original and perturbed aggregated sub-totals would be less for small cell adjustments than with full random rounding.

[PLACE FIGURE 3 HERE]

According to Figure 3, the differences between the aggregated original and perturbed sub-totals for ten consecutive OAs rarely goes beyond ± 10 and therefore compared to record swapping there is less distortion when aggregating perturbed cells for this particular target variable.

6.1.3 Marginal Totals of Tables

In the previous sections, the impact of the SDC methods on internal cells of the tables and on sub-totals that are aggregated from internal cells were examined. In this section, the totals that appear as margins in the table are examined, and in particular the total number of persons in the OA. The distance metric is the Average Absolute Distance per OA:

$AADOA(N_{orig}, N_{pert}) = \frac{1}{|OA|}\sum_{k=1}^{|OA|} | N_{pert}^{k} - N_{orig}^{k} |$ where $|OA|$ is the number of OAs and N^k is the total number of persons in the OA.

The marginal totals of tables that have undergone small cell adjustments are obtained by aggregating the rounded and non-rounded cells. The full random rounding procedure however rounds the marginal totals separately from the internal cells and therefore all the marginal totals are original rounded totals but the tables are not additive. To assess the impact on the loss of additivity, we aggregated the rounded internal cells of the full rounding procedures and compared them to the true OA totals.

Table 6 presents the average absolute distance per OA (*AADOA*) metric for the record swapping and rounding procedures for the Travel to Work Table in EA1. To avoid confusion, a "*" is used to denote the fact that the marginal totals for the full random rounding methods are obtained by aggregating internal cells in order to assess the non-additivity of the table, although the actual margins in the table would be the rounded original total. The benchmarked random rounding to the OA totals (BRND) reflect the difference between the rounded total and the original total.

[PLACE TABLE 6 HERE]

In Table 6, the benchmarked rounding methods to the OA totals (BSCA and BRND) have small average absolute distances per OA total (*AADOA*) as expected which is due to rounding within the base of the true OA total. This would be the same distance metric for the other full rounding procedures (RR and CRND) but without the additivity of the tables. The extent of the non-additivity of the tables for the full random rounding (RR*) and the controlled random rounding to the overall total (CRND*) is reflected in the large average absolute distance per OA (*AADOA*) of about 7 (3.2% of the average OA total). In contrast, the small cell adjustment methods (SCA and CSCA) aggregate rounded and non-rounded cells and therefore tables are additive. However, different totals appear for the same population base in different tables. The average absolute distance per OA (*AADOA*) is about 6 (2.7% of the average OA total) for the small cell adjustments methods (SCA and CSCA).

It is interesting to note that in Tables 5 and 6 of Section 6.1.1 we obtained that the average absolute distance per cell (*AAD*) was slightly smaller for the small cell adjustments compared to the random record swapping: for small cell adjustments the *AAD* is 0.629 and for the 10% and 20% random record swapping the *AAD* is 0.722 and 1.036 respectively. However, small cell adjustments aggregate rounded and non-rounded cells to obtain an OA total and therefore the impact on the average absolute distance per OA (*AADOA*) is much larger compared to the random record swapping: for small cell adjustments the *AADOA* is 5.973 and for the 10% and 20% random record swapping the *AADOA* is 1.625 and 2.433 respectively.

6.1.4 R-U Confidentiality Map for Record Swapping

In this section, an R-U Confidentiality Map (Duncan, et al., 2001) is presented for the different record swapping scenarios. For the rounding procedures it is assumed that the disclosure risk arising from small cells in tables is minimal and therefore we only analyze the dimension of utility after applying the SDC methods.

Figure 4 presents an empirical R-U confidentiality map for the record swapping methods on tables in EA2 based on the disclosure risk measure *DR* and the distance metric *AAD*.

[PLACE FIGURE 4 HERE]

Based on Figure 4, the 1% swapping rates for both methods of record swapping have high utility but also very high disclosure risk (about 80% of the small cells in the table (ones and twos) are true values after taking into account the level of imputation). The 10% targeted record swapping has about the same disclosure risk as the 20% random record swapping (about 45% of the small cells are true values). However, more utility in the data is gained with the 10% targeted record swapping compared to the 20% random record swapping.

6.2 Impact on Variance of Estimates

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 Output Area (OA) level of geography in the table. The overall utility measure is obtained by the percent difference between the average variance across all of the OAs for the original tables and the same average variance for the perturbed tables.

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

Table 7 present results of the percent differences in the variance of the average cell counts (*RDV)* based on the different scenarios of record swapping and rounding procedures for the Census tables in EA2 and EA3. The same results are obtained when semi-controlling for totals in the rounding procedures and therefore only the independent small cell adjustments (SCA) and the independent full random rounding (RR) are presented in Table 7.

[PLACE TABLE 7 HERE]

In Table 7, a clear pattern emerges of decreasing variances of the average cell counts as higher swapping rates are introduced, i.e. the cell counts are "flattening" out. The random record swapping has a slightly larger reduction in the variance of the average cell count compared to the targeted record swapping. However, the opposite effect occurs with the rounding procedures and the variance of the average cell count is increasing although with less magnitude than the swapping methods. Therefore, when combining rounding procedures with record swapping we see that opposing effects on the variance are canceling out and we obtain less reduction in the variance of the average cell counts compared to the variance obtained by record swapping alone.

6.3 Impact on statistical analysis

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 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: $CV = \sqrt{\frac{\chi^2 / n}{\min(R-1),(C-1)}}$.

The utility measure is the percent relative difference:

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

Table 8 presents results of the percent relative difference in the Cramer's V Statistic *(RCV)* based on the different scenarios of record swapping and rounding procedures for a Census table in EA2 and EA3 defined by: OA (1,755 categories for EA2 and 1,800 categories for EA3)×Sex (2 categories) on the rows and Economic Activity (9 categories)×Long-Term Illness (2 categories) on the columns. The Cramer's V statistic was calculated for both the original table and the perturbed table. As would be expected for this type of analysis in a standard statistical package, the expected cell frequency e_{ij} is calculated by aggregating internal cells for both the small cell adjustments and the full random rounding procedures. A large Cramer's V represents a high level of association between the rows and the columns of the two-way table.

[PLACE TABLE 8 HERE]

Table 8 demonstrates the loss in association and attenuation when swapping records across geographical areas. The two-way Census table based on economic activity × long-term illness and OA × sex is leaning more towards independence since the counts are "flattening" out in the table. With higher swapping rates the loss in association is more severe. Targeted record swapping which was carried out on unique records in the table has less of an impact on the loss of association compared to the random record swapping. We also see in Table 8 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. As seen in Table 7, when combining rounding procedures with record swapping, there are opposing effects on Cramer's V and therefore the percent relative difference (*RCV*) is getting smaller for the higher swapping rates compared to the *RCV* on the rounding procedures alone.

When assessing the impact of analysis on multi-dimensional tables, through for example log-linear models, the same effects on the goodness of fit criteria occur as in the two dimensional table for Cramer's V. The swapping methods homogenize the counts and lower the level of association while rounding procedures artificially increase dependencies. These effects cancel out somewhat when combining the SDC methods.

Another tool for statistical inference is Spearman's Rank Correlation. This is a technique that tests the direction and strength of the relationship between two variables. The statistic is based on ranking both variables from the highest to the lowest and calculating a correlation statistic. An important assessment for analyzing the impact of SDC methods on statistical data is to test whether the rankings of values within the variables are distorted.

In the following example, two target variables are used from EA2: the number of full time employed females with no long term illness and the number of unemployed females with no long term illness. Each of the target variables are sorted across the 1,755 OAs of EA2 according to their size. The first target variable is very large with no small cells while the second target variable is sparse with many small cells. After sorting each target variable, the values across the OAs are grouped into 20 equal groupings (v^{orig}). This procedure is repeated for the perturbed target variables (v^{pert}).

The utility measure is: $RC = \frac{100 \times \sum_{k=1}^{|OA|} I(v_k^{orig} \neq v_k^{pert})}{|OA|}$ where I is the indicator function and is 1 if the statement is true and 0 otherwise, and $|OA|$ is the number of OAs in EA2.

Table 9 presents results of the percentage of values that have changed groupings due to the SDC methods for each of the two target variables in EA2.

[PLACE TABLE 9 HERE]

In Table 9, the more sparse the target variable the higher the *RC* measure. This is because of the high impact on rankings of values of variables when there are many small values that are perturbed (zeros, ones and twos). For the large target variable of full time females with no long term illness, there are no small cells to perturb. Therefore, there is no effect when carrying out small cell adjustments and only a 10% difference in groupings for the full random rounding. For the small target variable of unemployed females with no long term illness, the percentage of values that jump between groupings is about 50% due to the rounding of the small cells. The record swapping methods however have a greater impact on changes to the rankings of the variables, ranging from about 60% for the 10% swapping methods and 70% for the

20% swapping method for the large target variable and even higher percentages for the small target variable. Combining rounding methods with record swapping produces mixed effects where the percentages of the RC measure increase when combining rounding with the 10% swapping methods but decrease when combining with the 20% swapping methods.

7. Discussion

In this analysis, we examined two 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. In addition, we assessed the impact when combining the 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 severe distortion to the data. Small cell adjustments give protection to Census tables by eliminating small cells but only one set of variables and geographies can be disseminated in order to avoid disclosure by differencing nested tables. Full random rounding offers more protection against disclosure by differencing but similar to small cell adjustments, protected cells can be deciphered by linking tables on common margins. The overall distortion on internal cells of tables is slightly less severe with the rounding procedures compared to the swapping methods, but the effects on the marginal totals and the non-additivity of Census tables is more damaging. Semi-controlling the rounding procedures to the overall total or benchmarking to geographical totals increases the utility of the tables by preserving some of the additivity. In addition, 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 distorts rankings while rounding procedures introduce more dependencies, increase measures of association and have less impact on distortions to rankings. These effects that were found in the record swapping and rounding procedures are consistent across all tables containing counts or proportions and not only those examined in 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, for example, by combining methods that have opposing effects that may cancel out and benchmarking totals. 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.

8 References

Boyd, M and Vickers, P. (1991). Record Swapping – A Possible Disclosure Control Approach for the 2001 UK Census. *UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Thessaloniki , March 1999.

Carter, R. (2000). Notes on Disclosure control Procedures for Tabular Outputs from the UK 2001 Census. *ONS Internal Report.*

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001). Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. *Technical Report LA-UR-01-6428*, Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory

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

Gomatam, S., Karr, A. and Sanil, A. (2003). A Risk-Utility Framework for Categorical Data Swapping. *Technical Report Number 132*, National Institute of Statistical Sciences.

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, pp. 463-478.

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

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

Table 1 : Advantages and Disadvantages of Record Swapping as a Pre-Tabular SDC method for Census Tabular Outputs

Advantages	Disadvantages
Consistent tables	High proportion of high-risk (unique) records left unperturbed
Preserves marginal distributions at higher aggregated levels	Errors (bias) in data, joint distributions distorted
Some protection against disclosure by differencing nested tables	Effects of perturbation hidden and cannot be accounted for in the analysis of the data
Less edit failures when swapping geographies	Method not transparent to users (perception of disclosure risk)

Figure 1: Confidence Intervals for Random Rounding to Base 3

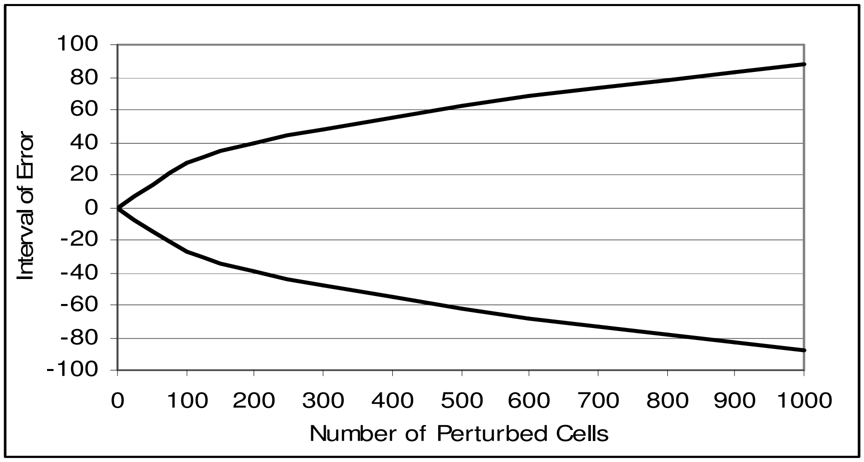


Table 2 : Advantages and Disadvantages of Small Cell Adjustments and Full Random Rounding on Census Tabular Outputs

Advantages	Disadvantages
Methods clear and transparent to users	Stochastic methods are easier to decipher through linking tables, so tables need to be audited prior to release
Stochastic methods can be accounted for in statistical analysis	
Small Cell Adjustments	
Protection for high-risk (unique) cells against identification	Inconsistent totals between tables since margins aggregated from rounded and non-rounded cells
Only small cells are affected by the rounding	No protection against disclosure by differencing so only one set of geographies and other variables disseminated
	Inconsistent and non-rounded marginal totals makes it easier to decipher
Full Random Rounding	
Protection for high- risk (unique) cells against identification	Margins rounded separately and tables are not additive
Protects against disclosure by differencing nested tables	

Table 3: Table Characteristics for EA1

	Table 1	Table 2	Table 3	Table 4	Table 5
Number of Individuals	437,744	320,621	437,744	317,064	433,817
Number of internal cells	80,298	214,128	50,558	53,532	83,272
Average cell size	5.45	1.50	8.66	5.92	5.21
Number of zeros	47,433 (59.1%)	139,337 (65.1%)	26,475 (52.4%)	17,915 (33.5%)	34,161 (41.0%)
Number of small cells	10,137 (12.6%)	41,114 (19.2%)	14,611 (28.9%)	14,726 (27.5%)	22,988 (27.6%)

Table 4: Percentage of Records in Small Cells of Tables that were Not Swapped or Imputed for two EAs

Method	EA1			EA2		
	Original – 84.2%			Original – 79.1%		
	1%	10%	20%	1%	10%	20%
Random	82.0%	63.4%	43.6%	77.0%	57.9%	38.4%
Targeted	80.6%	45.9%	18.0%	75.7%	43.0%	16.9%

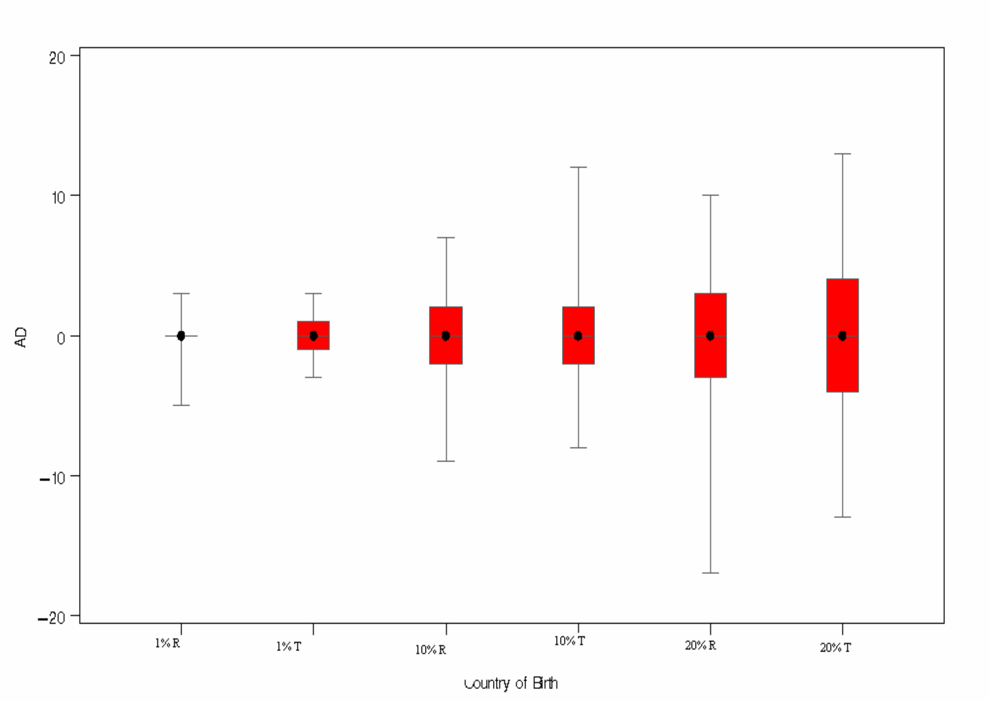
Table 5: Average Distance Metrics Between Original and Perturbed Internal Cells of Tables for EA1

	Method		HD	AAD
Original	SCA		5.272	0.629
	BSCA		5.394	0.653
	RR		5.411	1.021
	BRND		5.467	1.045
Random	1%	Original	1.044	0.136
	10%	Original	3.714	0.722
		SCA	6.305	1.114
		RR	6.425	1.248
	20%	Original	5.238	1.036
		SCA	7.173	1.315
		RR	7.285	1.402
Targeted	1%	Original	1.376	0.160
	10%	Original	4.787	0.845
		SCA	6.791	1.165
		RR	6.895	1.298
	20%	Original	6.372	1.173
		SCA	7.800	1.383
		RR	7.900	1.468

[*] SCA – small cell adjustments, BSCA – benchmarked SCA to OA total, RR – random rounding; BRND – benchmarked RR to OA totals

Figure 2: Box Plot of ADs for the Number of Males Born in Western Europe in Ten Consecutive OAs of EA1 for Record Swapping

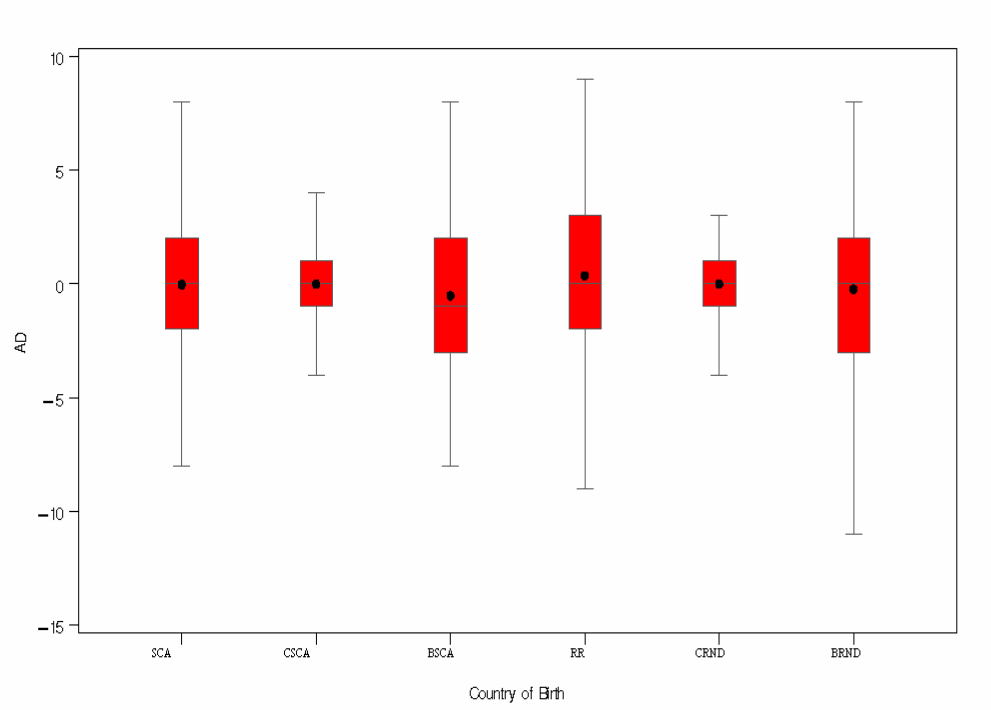
Average Original Sub-total in 10 OAs = 24.6



[*] 1%R – 1% random record swapping, 1%T – 1% targeted record swapping, 10%R – 10% random record swapping, 10%T – 10% targeted record swapping, 20%R – 20% random record swapping, 20%T – 20% targeted record swapping

Figure 3: Box Plot of ADs for the Number of Males Born in Western Europe in Ten Consecutive OAs of EA1 for Rounding Methods

Average Original Total in 10 OAs = 24.6



[*]SCA – small cell adjustments, CSCA – controlled SCA to overall total, BSCA – benchmarked SCA to OA total, RR – random rounding, CRND-controlled RR to overall total, BRND – benchmarked RR to OA totals

Table 6: Average Absolute Distance per OA Total (AADOA) for the Travel to Work Table in EA1 for Rounding and Record Swapping Procedures (Average OA Total=215.6)

Method	Average Distance per OA Total (*AADOA*)
10% random swap	1.625
10% targeted swap	1.391
20% random swap	2.433
20% targeted swap	2.113
Small cell adjustments (SCA)	5.973
Controlled SCA to overall total (CSCA)	5.981
Benchmarked SCA to OA totals (BSCA)	0.908
Random rounding* (RR)	6.991
Controlled RR to overall total* (CRND)	7.178
Benchmarked RR to OA totals (BRND)	0.877

* Marginal totals obtained by aggregating rounded internal cells

Figure 4: R-U Confidentiality Map for Record Swapping Methods on Tables in EA2

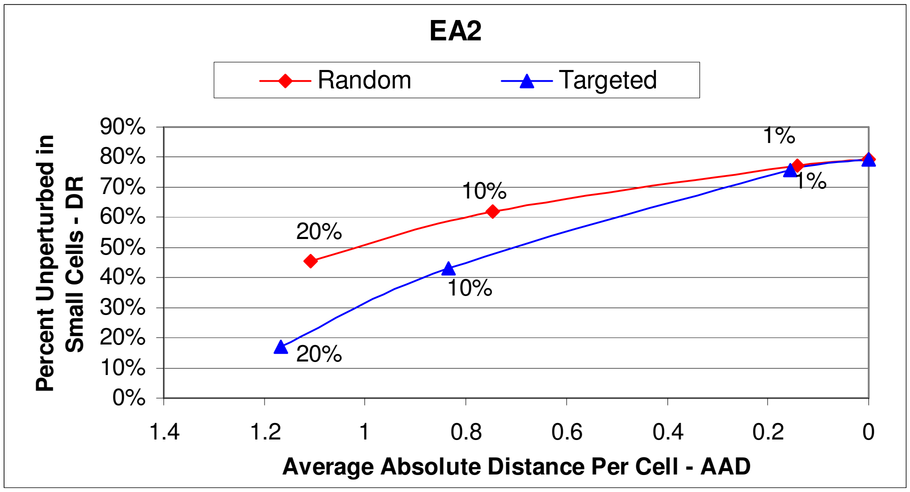


Table 7: Percent Relative Difference in Variance of Cell Counts (RDV) Between Perturbed and Original Tables

Swapping	EA2			EA3		
Method	Variance=235			Variance=363		
	Original	SCA	RR	Original	SCA	RR
Original	0.0	0.4	0.7	0.0	0.2	0.4
10% Random	-11.6	-11.2	-11.0	-11.6	-11.4	-11.2
20% Random	-17.8	-17.5	-17.3	-17.8	-17.6	-17.4
10% Targeted	-11.2	-10.9	-10.7	-11.4	-11.2	-11.0
20% Targeted	-17.4	-17.1	-16.9	-17.5	-17.3	-17.2

*SCA – small cell adjustments, RR – random rounding

Table 8: Percent Relative Difference in Cramer's V (RCV) Between Perturbed and Original Two-way Table (OA × Sex and Economic Activity × Long-Term Illness)

Swapping	EA2			EA3		
Method	Cramer's V= 0.1562			Cramer's V= 0.1695		
	Original	SCA	RR	Original	SCA	RR
Original	0.0	12.0	13.5	0.0	10.6	12.1
10% Random	-2.2	9.8	11.4	-2.8	8.2	9.6
20% Random	-4.2	7.7	9.7	-4.4	6.0	7.4
10% Targeted	-1.8	10.4	12.4	-1.5	9.3	10.7
20% Targeted	-3.8	9.3	10.2	-3.8	7.4	8.9

*SCA – small cell adjustments, RR – random rounding

Table 9: Percent Change of Values Between Groupings (RC) for Full Time Females with No Long Term Illness and Unemployed Females with No Long Term Illness in EA2

Swapping Method	Full Time Females with NLTI N=76,398			Unemployed Females with NLTI N=3,772		
	Original	SCA	RR	Original	SCA	RR
Original	0.0	0.0	10.0	0.0	48.3	54.5
10% Random	59.3	60.6	54.1	70.1	73.3	70.7
20% Random	71.1	69.6	65.4	82.3	79.1	77.6
10% Targeted	61.6	61.6	52.1	65.5	71.5	69.2
20% Targeted	72.2	72.2	71.4	83.3	77.8	76.4

*SCA – small cell adjustments, RR – random rounding