WP. 33
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
(Geneva, Switzerland, 9-11 November 2005)

Topic (v): Confidentiality aspects of tabular data, frequency tables, etc.

INFORMATION LOSS MEASURES FOR FREQUENCY TABLES

Invited Paper

Submitted by the Office for National Statistics, University of Southampton, United Kingdom, and the Hebrew University, Israel [1]

[1] Prepared by Natalie Shlomo and Caroline Young.

Information Loss Measures for Frequency Tables

Natalie Shlomo* and Caroline Young**

* Southampton Statistical Sciences Research Institute, University of Southampton, Department of Statistics, Hebrew University, Office for National Statistics
** University of Southampton, Office for National Statistics

Abstract: In order to manage the disclosure risk in frequency tables containing population counts, the tables undergo statistical disclosure control (SDC) methods. This results in information loss. We examine quantitative information loss measures for frequency tables and compare them across different SDC methods. We show examples of the information loss measures on real UK 2001 Census tables after they have been perturbed. We study the relationship between the results of the information loss measures, the perturbation method and the characteristics of the table (sparsity, skewness, uniformity, etc.).

1. Introduction

The Office for National Statistics (ONS) is leading the development of a new Internet service, Neighborhood Statistics (NeSS), which provides access to tables containing administrative and census data for small areas. The object is to supply the information needs for the National Strategy for Neighborhood Renewal and deliver small area statistics for policy evaluation, informing new developments in areas of deprivation and for addressing issues arising in local areas. The statistical disclosure control (SDC) methods at the ONS for protecting NeSS tables containing population counts include post-tabular methods: controlled rounding and cell suppression implemented using the Tau-Argus Statistical Disclosure Control Software (Hundepool (2003)), and stochastic unbiased forms of random rounding and small cell adjustments. Each of these methods modify the original data in the table in order to reduce the disclosure risk resulting in small cells of the tables. Reducing disclosure risk however results in information loss. In this paper we develop and evaluate quantitative information loss measures for determining the impact of the SDC methods on the original table.

Information loss measures can be split into two classes: measures for data suppliers in order to make informed decisions about optimal SDC methods which depend on the characteristics of the tables, and measures for users in order to allow adjustments to be made when carrying out statistical analysis on protected tables. In this paper, we focus on measures for data suppliers who have access to the raw tables and the aim is to choose the best SDC method which minimizes the information loss.

The SDC methods reviewed in this paper all give adequate protection against disclosure by identification since the small cells are eliminated from the tables. However, small cells also result from differencing nested non-coterminous tables. The cell suppression and small cell adjustments do not protect against disclosure by differencing whereas the full rounding methods do. Therefore, in order to obtain the same level of protection for all the SDC methods in this analysis, we assume that only one set of coding of the variables and geographies are disseminated in the tables and that there is no risk of disclosure by differencing.

Section 2 introduces the SDC methods that will be compared in the paper and Section 3 the data used for analysis. Section 4 presents information loss measures with numerical and graphical results on the data. We conclude in Section 5 with a discussion.

2. Data Masking Techniques for Frequency Tables

Some methods for protecting frequency tables against the disclosure risk of small cells in tables are:

2.1 Small Cell Adjustments (SCA)

Small cell adjustments is an unbiased random rounding procedure carried out on the small cells of the tables (ones and twos). 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. 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. For this analysis, we randomly rounded to base 3. For each cell, the mean of the perturbation is 0 and the variance is 2. When only small cells are rounded, the margins of the tables are obtained by aggregating the rounded and non-rounded cells, and therefore tables with the same population base will have different totals due to the stochastic process.

2.2 Full Random Rounding (RaRo)

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, the margins are rounded separately from the internal cells and therefore tables are not additive. We implemented random rounding to base 3 for this analysis.

Although we implemented the small cell adjustments and the full random rounding independently in each cell for this analysis, we note that the random rounding procedure can be improved by controlling for some of the marginal (and overall) totals of the table. A very simple algorithm for semi-controlling the random rounding procedure which preserves row totals and the overall total (or column totals after first transposing the table) is as follows:

1. Convert the entire table so that the entries are residuals of the rounding base.
2. Select first row of the table and randomly sort the entries.

3. For those entries having $res(x)$, select first $\frac{res(x)}{b}$ of the entries and round upwards, the rest of the entries round downwards. Repeat for all $res(x)$.
4. Sort entries back into their proper order.
5. Repeat on next row.

2.3 Controlled Rounding (CR(3))

Controlled rounding is a complex procedure carried out in Tau-Argus which is intended to be used as an ONS standard tool for disclosure control on frequency tables. In particular, it is largely supported by NeSS for protecting administrative register based tables disseminated over the internet. The procedure uses sophisticated linear optimization programming techniques to round entries, where the constraint is the equality of the rounded margins to the sum of the interior rounded cells. The algorithm for controlled rounding can also be carried out on the small cells only of the table, thus preserving totals and marginal distributions while only perturbing the small cells. All tables were controlled-rounded to base three for this analysis.

2.4 Suppression (S-A and S-WA)

Tau-Argus can also be used to suppress sensitive cells in frequency tables. Sensitive cells are defined as having counts of a one or a two. To apply secondary suppressions, the Hypercube method was chosen since a solution could be obtained for all tables in this analysis. This ensured a fair assessment of performance across tables. Note that due to the nature of the Hypercube method, occasionally some relatively large counts in the tables were secondary suppressed (Geissing (2003)).

In order to assess information loss from the perspective of what a user might do with suppressed cells in a table prior to analyzing the data, we implemented two very simple methods of imputation for the suppressed cells. Note that more sophisticated techniques for filling in missing data were not carried out in this analysis, but will be developed in future research. Let m_{kj} be a cell count in a two way table $k = 1,...,K$ rows and $j = 1,...J$ columns. For NeSS and Census tables at the ONS, rows are typically geographies: outputs areas or wards. The columns are defined by cross-classified variables, for example sex*long term illness*economic activity. Let the marginal totals be defined as: $m_{k.}$ and $m_{.j}$. The margins appear in the table without perturbation unless they have a small value and are primary suppressed. In that case, we define the margin to take a value of 1 for the following imputation schemes. Let z_{kj} be an indicator taking on the value of 1 if the cell was suppressed (primary or secondary) and a 0 otherwise.

In the first method (S-A), the user replaces all suppressed cells of row k with an average total of the suppressed values, i.e. $\frac{m_{k.} - \sum_{j=1}^{J} m_{kj}(1 - z_{kj})}{\sum_{j=1}^{J} z_{kj}}$. In the second method (S-WA), we use a weighted average to replace suppressed cells in a row *k*.

The weights are based on the average cell size of the columns j: $w_j = \frac{m_{.j}}{J}$. A low frequency column will result in a smaller imputed cell frequency and a high frequency column will result in a larger imputed cell frequency. Each suppressed cell in row k is replaced by: $\frac{w_j \times (m_{k.} - \sum_{j=1}^{J} m_{kj}(1 - z_{kj}))}{\sum_{j=1}^{J} w_j z_{kj}}$.

3. Data Used

For the purpose of this analysis we used three 2001 Census tables from one Estimation Area of the UK in the Southwest part of the country. The area included 437,744 persons in 182,337 households in 70 wards (on average 6,250 persons to a ward for this Estimation Area). The tables were the following:

(1) Tenure(3) * Age (7) * Health(4) * Ward
(2) Ethnicity (17) * Ward
(3) Economic Activity (9) * Sex (2) * Long-Term Illness (2) * Ward

The Economic Activity table only includes employed persons. The different SDC methods as described in Section 2 were implemented on the tables.

Table 1 provides summary statistics for each of the tables. Tenure is the largest table in terms of number of cells but also the sparsest with many small cells. The Ethnicity table contains large cell counts for the ethnic 'white' group defined in one column of the table and this is reflected in high skewness and high standard error of the cell counts. In comparison, the Employment table consists of both large and small cell counts.

		Table		
		Tenure	**Ethnicity**	**Employment**
Number of Person in Table		433,817	433,817	317,064
Number of Cells		5,880	1,120	2,520
Average cell size and Standard Error		73.8 (3.3)	387.3 (51.3)	125.8 (6.6)
Average cell size in row	**Minimum**	0.0	0.1	0.0
	Maximum	171.4	899.9	309.3
Average cell size in column	**Minimum**	0.2	2.8	3.0
	Maximum	943.7	5,729.4	1,411.7
Percentage of Zero Cells		26%	23%	17%
Percentage of Small Cells		12%	9%	9%

Table 1: Summary Statistics of Tables

4. Information Loss Measures

Information loss measures can be divided into several subsets according to the statistical aspect that is to be measured: Measures for distortion to distributions; Impact on the variance of estimates; Impact on measures of association; Statistical hypothesis tests for bias; Impact on statistical analysis, i.e. Goodness of fit criteria, Rank correlations.

4.1 Measuring distortion to distributions

Information loss measures that measure distortion to distributions are based on distance metrics between the original and perturbed cells. Some useful metrics were presented in Gomatam and Karr (2003). Since the basic unit of most of the Census and NeSS tables is a geography, i.e. ward, we calculate a distance metric for each ward separately in the table and then take the overall average across all of the wards for the information loss measure. When comparing the average distance metric across wards, we need to take into account the level of dispersion as expressed by the standard error (confidence interval).

Changing the notation from the previous section, let D^k represent a table for a ward k and let $D^k(c)$ be the cell frequency c in the table. Let $|W|$ be the number of wards in the estimation area. The distance metrics are:

- Hellinger's Distance:

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

- Relative Absolute Distance:

$$RAD(D_{pert}, D_{orig}) = \frac{1}{|W|}\sum_{k=1}^{|W|}\sum_{c\in k}\frac{|D^k_{pert}(c) - D^k_{orig}(c)|}{D^k_{orig}(c)}$$

- Average Absolute Distance per Cell:

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

the number of cells in the k^{th} ward.

These distance metrics can also be calculated for sub-totals and totals of the tables. In this report, we use a distance metric defined by the individual perturbations for sub-totals: $PA(N^k_{pert}, N^k_{orig}) = N^k_{pert}(C') - N^k_{orig}(C')$ where $N^k(C') = \sum_{c\in C'} D^k(c)$ is a sub-total for group C'. Table 2 presents results of the information loss measures based on average distance metrics across wards for the tables and their confidence intervals.

		SCA	RaRo	CR(3)	S - A	S - WA
Tenure	**HD**	1.97 (±0.16)	2.07 (±0.15)	1.78 (±0.15)	0.79 (±0.30)	1.20 (±0.14)
	RAD	10.79 (±1.55)	14.27 (±1.77)	10.64 (±1.29)	7.04 (±3.83)	8.36 (±1.57)
	AAD	0.16 (±0.03)	0.70 (±0.08)	0.52 (±0.06)	0.16 (±0.14)	0.15 (±0.05)
Ethnicity	**HD**	0.55 (±0.13)	0.72 (±0.12)	0.59 (±0.10)	0.79 (±0.89)	0.34 (±0.20)
	RAD	1.59 (±0.0.47)	2.38 (±0.59)	1.98 (±0.47)	12.06 (±20.76)	1.42 (±0.57)
	AAD	0.12 (±0.04)	0.69 (±0.08)	0.56 (±0.06)	3.13 (±5.80)	0.20 (±0.12)

		SCA	RaRo	CR(3)	S - A	S - WA
Employment	**HD**	0.93 (±0.15)	1.09 (±0.12)	0.93 (±0.10)	0.61 (±0.43)	0.45 (±0.11)
	RAD	3.24 (±0.65)	4.28 (±0.64)	3.56 (±0.52)	5.58 (±0.56)	1.87 (±0.49)
	AAD	0.12 (±0.02)	0.75 (±0.07)	0.59 (±0.06)	0.47 (±0.66)	0.14 (±0.09)

Table 2: Average Distance Metrics Between Original and Perturbed Tables per Ward (95% confidence intervals in parentheses)

HD is based on information theory and less intuitive than the other distance metrics. From Table 2, we see that HD doesn't pick up differences between small cell adjustments (SCA) and full rounding (RaRo and CR(3)) as the other measures do. This is because HD is more influenced by small cells than the other metrics. For the Tenure Table and Ethnicity Table, the distance metrics show consistency with respect to the order of the information loss according to the SDC methods. The Employment Table has a slightly mixed order of information loss for the SDC methods depending on the distance metric. This shows that several metrics should be used when assessing bias due to SDC methods and that the impact on the distance metrics are driven by the characteristics of the table.

The minimum distance metric for the Employment Table and Ethnicity Table is obtained by cell suppression with imputed weighted averages (S-WA). This method is nearest to obtaining the original table since there is less error when imputing for suppressed small and large cells. For the Tenure Table, the minimum distance metric is suppression with simple averages (S-A). This is because of the uniformity of the table. The maximum distance metric for the Employment Table and the Tenure Table is random rounding (RaRo). For the Ethnicity Table, the maximum distance metric is cell suppression with imputed simple averages (S-A) because of the fact that the table is highly skewed with one very large column. When a cell in this column is suppressed, the simple average imputation does not take into account the differential cell sizes. Note that controlled rounding (CR(3)) is always better than the random rounding (RaRo), and small cell adjustments (SCA) is on the whole doing better than both.

Users typically want to aggregate tables of lower level geographies in order to obtain statistics for non-standard higher level geographies. The lower level tables however have many small cells and are therefore greatly perturbed because of the SDC methods applied. This leads to more information loss when aggregating sub-totals. In order to evaluate the range of the perturbations for sub-totals of specific target variables obtained by aggregating lower level geographies, we use the statistical graphing tool of a box plot on the differences between the perturbed sub-total and the original sub-total (*PA)*. For unbiased SDC methods, we expect the average and median to be centered around zero. The length of the box and the length of the whiskers gives an indication of how widespread the perturbed totals are from the original totals. Figure 1 presents the box plot of the differences between the original and perturbed sub-totals *(PA's)* for the number of unemployed females with long term illness after aggregating the variable for three consecutive wards across the Estimation Area.

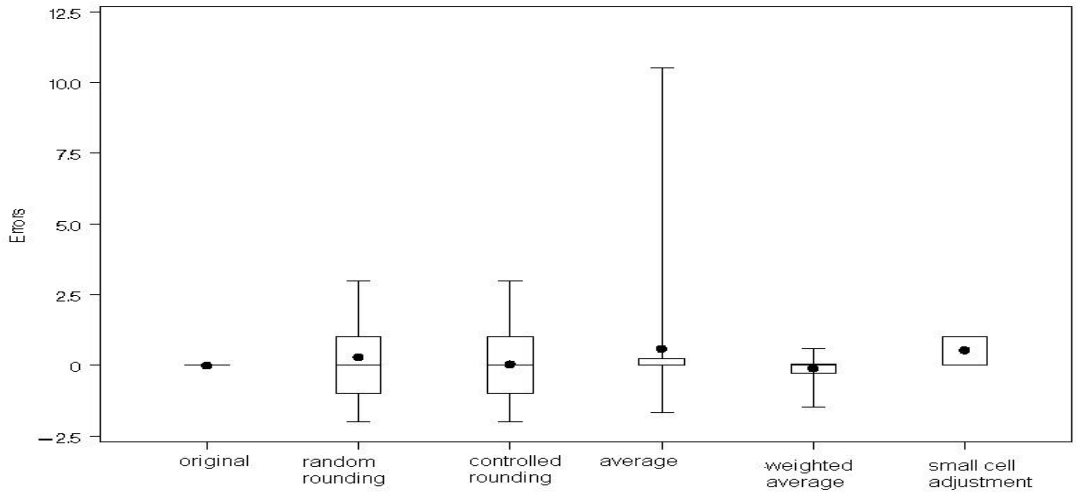

Figure 1: Box Plot of PA's for the Number of Unemployed Females with Long Term Illness in Three Consecutive Wards
Average Original Total in combined 3 wards = 14.4

Focusing on one target variable, we see the impact of the different SDC methods on the sub-totals. Full rounding (RaRo and CR(3)) have the same effects with respect to the differences in the perturbed and original sub-totals. Suppression with weighted averages (S-WA) has less information loss than simple averages (S-A). Note that the *PA*'s for S-A can differ by 76% of the average original total in three consecutive wards. The small cell adjustments (SCA) result in less information loss than the other rounding procedures since only small cells are affected.

4.2 Impact on Variance of Estimates

SDC methods will have an impact on the variances that are calculated for estimates based on the frequency tables. We first examine the variance of the cell counts across the geographies (wards) before and after the SDC methods as follows: For each ward *k*, we calculate: $V(D^k_{orig}) = \frac{1}{|k|-1}\sum_{c\in k}(D^k_{orig}(c) - \overline{D}^k_{orig})^2$ where $\overline{D}^k_{orig} = \frac{\sum_{c\in k} D^k_{orig}(c)}{|k|}$ and $|k| = \sum_c I(c \in k)$ the number of cells in the k^{th} ward. Next we take the average of the variance across all the wards: $V(D_{orig}) = \frac{1}{|W|}\sum_{k=1}^{|W|} V(D^k_{orig})$. This is repeated for the perturbed table. The information loss measure is: $VR(D_{pert}, D_{orig}) = 100 \times \frac{V(D_{pert}) - V(D_{orig})}{V(D_{orig})}$. Table 3 presents results of the measure *VR* for the different SDC methods on the three Census tables after removing outlying wards that had very small cells.

VR	SCA	RaRo	CR(3)	S - A	S- WA
Tenure	0.003%	0.009%	0.006%	-1.278%	-0.179%
Ethnicity	0.003%	- 0.160%	-0.168%	-2.298%	-0.069%
Employment	0.006%	0.003%	0.138%	-0.266%	-0.111%

Table 3: Percent Relative Difference of Average Variance of Cell Counts (VR) between Original and Perturbed

For all tables, cell suppression with imputed simple averages (S-A) and weighted averages (S-WA) result in smaller overall variance compared to the original tables. This indicates that these SDC methods, especially the S-A method, are producing more uniform cell counts. The stochastic methods of rounding (SCA and RaRo) have little impact on the cell counts for the Tenure and Employment Tables. The Ethnicity Table, which has one large column and very many sparse columns, have more uniform small cells based on full rounding procedures (RaRo and CR(3)), and therefore a smaller overall variance is obtained.

Another variance that we will focus on is the "between" variance used in regression (ANOVA) analysis for a specific target variable. A typical statistical analysis would be to carry out a regression analysis and model a target variable based on a set of explanatory variables (geography, sex, age, etc.). For a regression analysis the goodness of fit criterion is expressed by the measure R^2. This measure is based on a decomposition of the variance of the target variable. For categorical explanatory variables, 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 target variable within the groupings defined by the combination of the explanatory variables and the "between" sum of squares SSB which measures the variance 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 are shrinking 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: Let $P^k_{orig}(c)$ be a target proportion for a cell c in ward k, i.e. $P^k_{orig}(c) = \dfrac{D^k_{orig}(c)}{\sum_{c \in k} D^k_{orig}(c)}$ and let $P_{orig} = \dfrac{\sum_{k=1}^{|W|} D^k_{orig}(c)}{\sum_{k=1}^{|W|}\sum_{c \in k} D^k_{orig}(c)}$ be the overall proportion. The "between" variance is defined as: $BV(P_{orig}) = \dfrac{1}{|W|-1}\sum_{k=1}^{|W|}(P^k_{orig}(c) - P_{orig})^2$ and the information loss measure is: $BVR(P_{pert}, P_{orig}) = \dfrac{BV(P_{pert})}{BV(P_{orig})}$.

The target variables in this example are the proportion of full time and part time employed males and females with no long term illness out of the total number of employed persons and the explanatory variable defining the groupings are the wards

as obtained in the Employment Table. Table 4 presents the results of the measure *BVR* for the different SDC methods. The overall proportion out of the total in the table is in parentheses.

BVR	SCA	RaRo	CR(3)	S – A	S- WA
Part Time Males NLTI (1.9%)	0.47	6.12	3.52	1.96	0.37
Full Time Males NLTI (31.2%)	0.99	1.61	3.07	0.90	1.42
Part Time Females NLTI (11.1%)	1.01	3.62	1.01	1.11	0.51
Full Time Females NLTI (15.7%)	0.88	1.98	0.58	0.46	0.46

Table 4: Percent Difference of "Between" Variance (*BVR*) for the Proportion of Full and Part-Time Employed Males and Females With No Long-Term Illness Within Groupings Defined by Wards

This information loss measure is showing mixed results, sometimes showing more homogenizing of the target proportions between the wards (BVR less than one) and sometimes showing less. It appears that the full rounding procedures (RaRo and CR(3)) have larger "between" variances and more differences in the proportions across the wards. It's interesting to note that cell suppression with imputed averages (simple or weighted) has conflicting effects on the "between" variance of the target proportions. The small cells that are modified due to the SDC methods in particular have an impact on the proportion of the target variable in the ward, since a small number adjusted down or adjusted up can produce either a proportion of 0 or even a proportion of 1 for some wards. Therefore, the effects on the "between" variance are heavily influenced by the way the small cells are perturbed. The small cell adjustments seem to have the least impact on the "between" variances. This information loss measure therefore is difficult to interpret since no consistent pattern emerges and it seems to be driven by the realization of the SDC methods on the target proportions. Future work will investigate this information loss measure further and ways of improving it.

4.3 Impact on Measures of Association

Another statistical analysis that is frequently carried out on contingency tables are tests for independence between categorical variables that span the table. The test for independence for a two-way table is based on a Pearson Chi-Squared Statistic $c^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ where o_{ij} is the observed count and $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ is the expected count for row i and column j. If the row and column are independent then c^2 has an asymptotic chi-square distribution with *(R-1)(C-1)*and for large values the test rejects the null hypothesis in favor of the alternative hypothesis of association. We use the measure of association, Cramer's V: $CV = \sqrt{\frac{c^2 / n}{\min(R-1),(C-1)}}$ and define information loss by the percent relative difference between the original and perturbed table: $RCV(D_{pert}, D_{orig}) = 100 \times \frac{CV(D_{pert}) - CV(D_{orig})}{CV(D_{orig})}$. Table 5 presents the information loss measures *RCV* for the three Census tables. We produced two way tables from each of the Census tables where the rows are the wards cross

classified with the demographic variables and the columns the cross classified target variables. For example, for the Employment Table, the rows are ward*sex and the columns are economic activity* long term illness.

RCV	SCA	RaRo	CR(3)	S - A	S- WA
Tenure	0.26%	0.29%	0.27%	0.20%	-0.13%
Ethnicity	0.11%	0.11%	0.00%	48.27%	-0.33%
Employment	0.10%	0.13%	0.06%	2.36%	-0.09%

Table 5: Percent Relative Difference in Cramer's V (*RCV)* Between Perturbed and Original Two-way Tables

All methods except for suppressed cells with imputed weighted averages (S-WA) indicate that the perturbed tables have artificially more association than the original table. The skewed Ethnicity Table is particularly affected when imputing simple averages for the suppressed cells (S-A) as seen in Table 5 since if a large cell is secondary suppressed along with a very small cell, they are both replaced with the simple average resulting in a distribution that is more "flat". This apparently raises the level of association with the geography variable in the table. The weighted averages (S-WA) however seem more consistent with the true values and there is a slight loss of association.

4.4 Statistical Hypothesis Tests for Bias

We first carry out an exact Binomial Hypothesis Test to check if the realization of the random rounding procedures on the tables followed the Binomial rounding scheme. The null hypothesis is: $H_0 : p = 2/3$. The realized proportions and p-values are presented in Table 6 where small p-values means that we reject the null hypothesis and the random rounding procedure was biased. Based on the results, we see a slight bias in the Tenure Table with respect to the small cell adjustments.

	Test for Ones		Tests for Twos	
	Proportion	**p-value**	**Proportion**	**p-value**
Tenure				
SCA	0.707	0.0403	0.628	0.0758
RaRo	0.663	0.3756	0.655	0.1756
Employment				
SCA	0.705	0.1851	0.673	0.4449
RaRo	0.685	0.1535	0.655	0.2596
Ethnicity				
SCA	0.677	0.4306	0.692	0.3670
RaRo	0.701	0.0997	0.656	0.3501

Table 6: Exact Binomial Test for Random Rounding Procedures

For the other SDC methods, we can use a Wilcoxen Signed Rank Test to check whether the location of the empirical distribution has changed. The null hypothesis for the test is no change. The standardized statistic is based on ranking the cells in the table and testing whether the sum of the ranking scores for the original cells deviates from the expected average under the null hypothesis of equal location. If there is a

large deviation (small p-value), then one can say that the location of the distribution has been shifted. Table 7 presents p-values for the Wilcoxon Signed Rank Test.

	Wilcoxen Sign Rank Test p- values		
	S-A	**S-WA**	**CR(3)**
Tenure	<0.001	0.0221	0.0017
Ethnicity	0.2166	0.3888	0.9383
Employment	0.0184	0.9559	0.9883

Table 7: p-Values for Wilcoxon Signed Rank Test for Same Location

The Tenure Table is showing significant p-values and we reject the null hypothesis of same location. Since the table is more uniform, it appears that the SDC methods have a larger impact on the distribution of the cell counts. The other tables are not significant except for the cell suppression with imputed simple averages on the Employment Table.

4.5 Impact on statistical analysis

We previously examined the impact of the perturbation schemes on regression analysis through the "between" variance. Another statistical tool for inferences is the 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 sets of data from the highest to the lowest. Therefore, one important assessment of the impact of the perturbation of statistical data is whether we are distorting the rankings of the variables. In the following example, we take target variables that are particularly sparse and therefore are subject to much perturbation: Male and Female students with long term illness (N=544 and N=380, respectively). We sort the original cell counts across wards according to their size and define deciles (10 equal groupings) v^{orig}. This is repeated for the perturbed cell counts which are sorted across wards according to their size and the original order to maintain consistency for the tied variables. Deciles v^{pert} are then defined for the perturbed variable after the sort. The information loss measure is the percent of wards that have changed deciles:

$$RC = \frac{100 \times \sum_{k=1}^{|W|} I(v_k^{orig} \neq v_k^{pert})}{|W|}$$ where I is the indicator function and is 1 if the statement is true and 0 otherwise, and $|W|$ is the number of wards.

Table 8 presents results of the percentage of deciles that have changed due to the perturbation method. Because of the sparseness of the target variables (70% of the cells in the tables take values less than 4), many cells were suppressed or small cells rounded which distorted the rankings of the cell counts. The imputation methods for the cell suppressions (S-A and S-WA) in particular distorted the rankings. An interesting result shown for these target variables is that the controlled rounding causes less distortion to the rankings than the other methods, including random rounding.

	SCA	RaRo	CR(3)	S-A	S-WA
Male Students with LTI	10.0%	25.7%	0%	35.7%	20.0%
Female Students with LTI	10.0%	5.7%	2.9%	20.0%	18.6%

Table 8: Percent Changes in Deciles for Male and Female Students with Long Term Illness (N=544)

Another statistical analysis frequently carried out on contingency tables is log linear modeling. For a 2-way table this narrows down to a test for independence and the Cramer's V statistic. For more variables in a contingency table, one can examine conditional dependencies and calculate expected cell frequencies based on the theory of log-linear modeling. The goodness of fit test for assessing the best fitting parsimonious model is the deviance or likelihood ratio L^2. This is the statistic that is minimized when calculating the maximum likelihood estimates of the parameters of the model. The information loss measure will be based on the ratio of the deviance between the perturbed table and the original table for a given model: $LR = \frac{L^2_{pert}}{L^2_{orig}}$.

Table 9 presents the *LR* measure for the table: Economic Activity (9) * Sex (2) * Long-Term Illness (2) * Ward. The model that is compared is:

$$\log(N_{ijkl}) = m + l_i^{Econ} + l_j^{Sex} + l_k^{LTI} + l_l^{Ward} + l_{ij}^{Econ*Sex} + l_{ik}^{Econ*LTI} + l_{il}^{Econ*Ward}$$

LR	Original	SCA	RaRo	CR(3)	S-A	S-WA
Deviance	4,486	5,283	5,316	5,214	6,404	4,744
Ratio	1.00	1.09	1.10	1.08	1.32	0.98

Table 9: Ratio of L^2 Statistic Between Perturbed and Original Table of Economic Activity* Sex*Long Term Illness*Ward

From Table 9, we see that the S-A method increased the deviance by 32%. It is likely that a different model would have been chosen based on the perturbed table as compared to the original table. Future work for this information loss measure will take a more in depth analysis on the impact of choosing different minimal sufficient statistics for original and perturbed tables.

5. Discussion

From this analysis, it is clear that the impact of the SDC methods with respect to information loss depends on the type of table and some general guidelines have emerged:

- A table that has only one or two columns of small values and the remaining columns with large values should not be suppressed since inevitably the secondary suppressions will involve some of the larger cells. A rounding procedure would be preferred.
- A table that is uniform has less information loss regardless of the SDC method so choose a method that causes the least changes to the table.
- A sparse table must have controlled totals so control round if possible, or apply semi-controlled random rounding.

Besides the characteristics of the table, the information loss measures perform differently depending on the outcome of the stochastic processes of the SDC methods. A more robust approach needs to be developed for assessing information loss.

The SDC methods should be tailored to the specific type of table. However, in a large Census context, one (or a combination) of SDC methods are usually applied across all tabular outputs regardless of the tables and the needs of the users. For example, small cell adjustments were implemented for all 2001 Census tabular outputs for England and Wales. This method had little impact on standard tables but had a large negative impact on the very large and sparse origin – destination tables. For the NeSS website which has localized (and less linked) tables, each type of tabular output should have an appropriate SDC method which minimizes the information loss of the data. In the future, we can envision on-line SDC methods tailored according to the users input as to the type of analysis that will be carried out and the variables of interest.

Future work will examine more closely the relationship between the information loss measures and the characteristics of the table and developing a set of guidelines on best practices for designing and protecting frequency tables. We aim to deliver a software tool for suppliers of NeSS tables in order to assess information loss prior to disseminating the tables over the internet. In addition, we need to develop information loss measures with respect to the users of the data and provide more guidance on analyzing perturbed tables, such as the effects of the SDC methods on statistical inferences, how to cope with tables having the same population base with different totals and sub-totals, and how to take into account suppressed cells.

6. Acknowledgements

We wish to thank Stephen Bond of the ONS who initiated this project and developed ideas for the imputation methods for suppressed cells and other related work.

References

Geissing, S. (2003), Coordination of Cell Suppressions: Strategies for Use of GHMITER, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, April 2003
www.unece.org/stats/documents/ 2003/04/confidentiality/wp.36.e.pdf

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

Hundepool, A., et. al. (2003) Argus Version 3.1 User's Manual, http://neon.vb.cbs.nl/casc/