Data Swapping for Protecting Census Tables

Natalie Shlomo[1], Caroline Tudor[2], and Paul Groom[2]

[1] Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom
N.Shlomo@soton.ac.uk
[2] Office for National Statistics, Segensworth Road, Titchfield, Fareham, PO15 5RR, United Kingdom
{Caroline.Tudor,Paul.Groom}@ons.gsi.gov.uk

Abstract. The pre-tabular statistical disclosure control (SDC) method of data swapping is the preferred method for protecting Census tabular data in some National Statistical Institutes, including the United States and Great Britain. A pre-tabular SDC method has the advantage that it only needs to be carried out once on the microdata and all tables released (under the conditions of the output strategies, eg. fixed categories of variables, minimum cell size and population thresholds) are considered protected. In this paper, we propose a method for targeted data swapping. The method involves a probability proportional to size selection strategy of high risk households for data swapping. The selected households are then paired with other households having the same control variables. In addition, the distance between paired households is determined by the level of risk with respect to the geographical hierarchies. The strategy is compared to a random data swapping strategy in terms of the disclosure risk and data utility.

Keywords: Targeted Data Swap, Random Data Swap, Disclosure risk, Data Utility, R-U confidentiality map.

1 Introduction

Protecting tables containing Census counts is more difficult than protecting tabular data from a survey sample since sampling a priori introduces ambiguity into the frequency counts. More invasive statistical disclosure control (SDC) methods are needed to protect against disclosure risks in a Census context where tables include whole population counts and this impacts negatively on the utility of the data. It is well known that Census data have errors due to data processing, coverage adjustments, imputations for 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 these errors and the protection that is already inherent in the data. A quantitative measure of disclosure risk should consider for example the amount of imputation and adjust parameters of the SDC methods 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

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 41–51, 2010.

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 in the tables, 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 protecting Census tables that are typically implemented at National Statistical Institutes (NSI) include pre-tabular methods, post-tabular methods and combinations of both. In this paper we focus on a pre-tabular method which is implemented on the microdata prior to the tabulation of the data. The most commonly used method is data 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. Data 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 to adjust their statistical analysis. This method however has yet to be implemented for a large scale Census. In practice, NSIs prefer data 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 NSIs do not typically release parameters of the SDC methods, i.e. swapping rates or probability transition matrices, in order to minimize the risk of deciphering the perturbation process.

In this paper, we propose a data swapping strategy that is targeted to high risk households. In addition, the distance between paired households for carrying out the swap is determined by the geographical level that is most at risk as defined by unique cells on margins of key variables (Young, Martin and Skinner, 2009). The targeted data swapping strategy is compared to a random data swapping strategy through quantitative disclosure risk and data utility measures according to the disclosure risk–data utility framework as described in Willenborg and De Waal (2001), Duncan, Keller-McNulty, and Stokes (2001) and Shlomo (2007). The data utility is assessed by analyzing the impact of the data swapping strategies on chi-square tests for independence as well as measuring distortions to cell counts for specified Census tables. Disclosure risk is assessed by the proportion of unique cells that are not perturbed in the tables. The analysis will be demonstrated on a real data set extracted from the UK 2001 Census.

Section 2 outlines the data swapping methods that are assessed and Section 3 details the data and Census tables that are used in the analysis. Section 4 presents the results of the comparison between the swapping methods followed by a discussion and conclusions in Section 5.

2 Data Swapping Methods

The most common pre-tabular method of SDC for Census tables is data swapping on the microdata prior to tabulation where values of 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. Data swapping can be targeted to high risk households found in small cells of tables as described in Section 2.1 thereby ensuring that households that are most at risk for disclosure are more 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 be induced. In addition, because of the conditional independence assumption, swapping geography will not necessarily result in inconsistent and illogical records. In contrast, swapping a variable such as age would result in many inconsistencies with other Census 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 arising from differencing two tables with nested geographies since data swapping introduces ambiguity into the true cell counts and in particular the zero counts.

2.1 Targeted Data Swap Strategy

Targeted data swapping is based on an allocation of a $p\%$ sample of households where p is the swapping rate to be determined by the NSI. Typically the data swapping is carried out within blocks of large geographical areas, eg., Estimation Area or Census Delivery Group Area. Within these large areas are hierarchies of geographies. For the UK 2001 Census data, there are three layers of nested geographies: Local Authority (LA), Wards and Output Areas (OA).

Census tables contain counts of individuals so to identify high risk households we need to first identify high risk individuals. High risk individuals are defined on the basis of frequency counts of univariate distributions on a set of key variables that are typically used to span Census tables at different levels of geography. A cell of size one on the univariate distribution means that there will be a unique individual on one of the margins of the table. Uniques on the margin of a table increase the risk of attribute disclosure since individuals can be identified on the basis of some of the variables spanning the table, and once identified, a new attribute can be learnt. High risk is defined through a score that is calculated for each individual as follows:

- Calculate frequency counts for M selected key variables each having k_m, $(m = 1,...,M)$ categories at the geographical level g: $N^g_{k_m}$ (not including individuals that have been imputed to account for the Census under-coverage).
- For every individual with values of categories $k = (k_1, k_2, ..., k_M)$, calculate a score at each level of geography g by taking the average of the reciprocals of the counts: $HR^g_k = (\sum_{m=1}^{M} 1 / N^g_{k_m}) / M$.
- A threshold is set for each level of geography and those scores above the thresholds determine high risk individuals.
- High risk households are defined as any household having at least one high risk individual.

The overall sample size in a Delivery Group area is calculated by multiplying the swap rate by the number of non-imputed households. This sample is then allocated across the lower level geographies (eg., OAs). We propose using two proportional allocations according to:

(1) the inverse number of (non-imputed) households in the OA, i.e. the larger the OA the less swapping required,
(2) the percentage of high risk households in the OA.

The final sample size for each OA is taken as the average sample size across the two proportional allocations provided that the final sample size is not over 20% of the number of (non-imputed) households in the OA. The random sample of households is drawn within each OA using a probability proportional to size (pps) design according to the above allocation. The size variable for the pps sampling is calculated so that a high value is given to high risk households and a low value is given to low risk households. This ensures that a disproportionate number of households at high risk will be selected in the sample and at the same time, guarantees that households of low risk will have a small but positive chance of being selected in the sample in order to introduce some randomness into the data swapping.

After the sample is selected, each of the households must be paired with another household in order to swap the geographical variables. The paired household must match on a set of control variables, eg. household size, age group and sex distribution, ethnicity indicator, 'hard to count' index. For the targeted swapping strategy we introduce the notion of distance swapping as defined in Young, Martin and Skinner, 2009. The idea is to pair households for swapping at a distance that is consistent with the geographical level of disclosure risk. Similar to the method carried out for defining high risk individuals, we first calculate the univariate distribution frequencies of individuals for the key variables at each geographical level. If there is a unique individual on any of the categories of the key variables at a geographical level g, the individual is flagged for that geographical level. The household geographical disclosure risk level is then defined as the highest geographical risk level from among all individuals in the household. For example, if there is an individual in a household that is flagged as being unique on one of the categories of the key variables at the ward level and another individual in the same household that is flagged at the LA level, the entire household is flagged at the LA

level of disclosure risk. The geographical level of disclosure risk is used in the algorithm for swapping as described below:

For the selected household to be swapped, we first check the level of geographical disclosure risk and choose a paired household at the appropriate geography having the same control variables. For example, if the level of geographical disclosure risk is flagged at LA, then the household must be swapped with a similar household having the same control variables in a different LA but within the large Delivery Group Area. If the level of geographical disclosure risk is flagged at ward, then the household must be swapped with another household having the same control variables in a different ward but within the same LA. If the level of geographical disclosure risk is flagged at OA, then the household must be swapped with another household having the same control variables in a different OA but within the same ward. Therefore, selected sampled households are swapped with other households having the same control variables but only at a distance that is appropriate to the geographical level of risk with respect to the uniqueness on marginal distributions of the key variables. The advantage of 'localized' data swapping is that we minimize the distance between pairs of households depending on the geographical level of risk and therefore at higher aggregations of geography we expect less distortion.

The search for a paired household in the swapping algorithm is carried out through several iterations. In the first attempt, the sampled household must match the paired household on a full set of control variables, eg. 'hard to count' index, household size, ethnicity indicator, sex and broad age distribution in the household. In subsequent attempts to search for a paired household, control variables undergo gradual collapsing to allow a better chance of finding a pair for the sampled household. Note that no household can be paired twice. Once a household is selected, all geographical variables are swapped between the two households.

2.2 Random Data Swap Strategy

We compare the targeted data swapping strategy in Section 2.1 with a random data swapping strategy. The same swapping rates are used for both strategies. The difference between the strategies is that households are selected for swapping using a simple random sample without replacement design in each OA, i.e. all households have equal chance of being selected for swapping. The sampled household is then paired with another household having the same control variables using the iterative procedure described above but no attempt is made to target high risk households or control the distance between swapped households.

3 Data

For this analysis, targeted and random data swapping strategies described in Section 2 were carried out on households from an extract of the 2001 UK Census containing two LAs at the following swapping rates: 2% and 5%. The extract included 327,718 individuals in 124,979 households. In the two LAs there were 35 wards and 1,111 OAs.

We define the following Census tables of individuals at the lower level of geography OA, where the number of categories are given in parenthesis:

(1) Religion(9) × Age-Sex(6) × OA

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

(3) Ethnicity (17) × Sex(2) × OA

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

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

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

The characteristics of the five tables are presented in Table 1. As can be seen, the tables have different average cell sizes and distributions of small cells. We also produce the same Census tables (1) to (6) at the ward level geography.

Table 1. Characteristics of Census tables at the OA geography

	Table 1	Table 2	Table 3	Table 4	Table 5	Table 6
Number of Individuals	327718	240,797	327,718	327,718	238,727	325,594
Number of internal cells	59,994	159,984	37,774	37,774	39,996	77,770
Average cell size	5.46	1.51	8.68	8.68	5.97	4.19
Number of zeros	34,546 (57.6%)	103,361 (64.6%)	23,939 (63.4%)	19,723 (52.2%)	12,697 (31.7%)	40,363 (51.9%)
Number of 1s	5,298 (8.8%)	20,793 (13.0%)	5,468 (14.5%)	7,329 (19.4%)	6,634 (16.6%)	11,260 (14.5%)
Number of 2s	2,771 (4.6%)	10,304 (6.4%)	2,607 (6.9%)	3,767 (10.0%)	4,511 (11.3%)	6,183 (8.0%)

4 Analysis

To compare the two data swapping strategies described in Section 2, we assess disclosure risk in terms of the proportion of unswapped unique cells in the Census tables described in Section 3, and data utility in terms of distortions to distributions and statistical inference.

4.1 Disclosure Risk

Disclosure risk arises from small cells in tables (or small cells appearing in potential slithers of differenced tables). In addition, the number and placement of zero cells can lead to identification and attribute disclosure when many tables are disseminated from one database.

Pre-tabular methods of disclosure control, and in particular data swapping, will not inhibit small cells from appearing in tables and therefore a quantitative disclosure risk measure is needed which reflects whether the small cells in tables are true values.

The quantitative disclosure risk measure for assessing the impact of data swapping is the proportion of cells of size one that have not been perturbed. This is calculated by counting the number of cells that have both an original and perturbed count of one divided by the number of cells with an original count of one. Let T^O represent the original table and let $T^O(c)$ be the cell frequency c in table T^O. Similarly, T^P represents the swapped table. The risk measure is defined as:

$$DR = \frac{\sum_c I(T^O(c) = 1, T^P(c) = 1)}{\sum_c I(T^O(c) = 1)},$$

where I is the indicator function receiving a value of 1 if it is true and 0 otherwise. Note that we ignore those individuals that have been imputed to adjust for the Census under-coverage since these are not considered at risk. In Table 2 we present the *DR* proportions for the Census tables described in Section 3. We also present the *DR* in Table 3 for smaller Census tables defined by the main marginal variable crossed with the OA geography.

In Tables 2 and 3, we see that the higher swapping rate protects more unique cells than the lower swapping rate. The random swapping has higher proportions of unique cells that are unperturbed than the targeted swapping. These results are as expected. The overall disclosure risk in some Census tables is high, even for the 5% data swapping, with over 80% of unique cells unperturbed. In addition, there are two clear patterns in Tables 2 and 3. All Census tables have the highest disclosure risk at the 2% random swap and the lowest disclosure risk at the 5% targeted swap. For some tables, the 2% targeted swap provides lower disclosure risk than the 5% random swap, for example in Tables (1), (3) and (4). The reason for this pattern is that the marginal

Table 2. Proportion of unperturbed unique cells (*DR)* in the Census tables

Table	Random 2%	Random 5%	Target 2%	Target 5%
(1)	0.939	0.853	0.749	0.650
(2)	0.932	0.837	0.912	0.822
(3)	0.944	0.848	0.549	0.457
(4)	0.924	0.831	0.723	0.629
(5)	0.910	0.805	0.894	0.779
(6)	0.929	0.828	0.925	0.819

Table 3. Proportion of unperturbed unique cells *(DR)* on the margins of the Census tables

Table	Random 2%	Random 5%	Target 2%	Target 5%
Religion×OA	0.954	0.851	0.595	0.495
Travel to work ×OA	0.899	0.826	0.912	0.800
Ethnicity ×OA	0.934	0.832	0.421	0.347
Country of birth ×OA	0.916	0.803	0.590	0.518
Economic activity×OA	0.823	0.668	0.816	0.584
Health status ×OA	0.811	0.684	0.821	0.737

distributions of religion, ethnicity and country of birth (as well as age-sex distribution) were used to define high risk households which according to the pps sampling had more chance of being selected into the sample for swapping. It is clear that the disclosure risk based on variables that are used to define high risk households would be reduced considerably under the targeted data swapping approach.

4.2 Data Utility

Data utility measures used in this analysis are based on a distance metric to measure the distortion to distributions and the impact on a measure of association based on a statistical test for independence between categorical variables.

Some useful distance metrics were presented in Gomatam and Karr (2003). The distance metric that is used in this analysis is the average absolute distance per cell of a Census table calculated as: $AD(T^O, T^P) = \sum_c |T^P(c) - T^O(c)| / n_T$ where n_T is the number of cells in the Census table.

Table 4 presents results of the distance metric *AD* for the Census tables defined in Section 3 at the OA geography and Table 5 the same distance metric *AD* for the Census tables defined at the ward geography. The aim is to show that at higher aggregations of geography, the targeted data swapping strategy obtains less distortion, i.e. smaller distance metrics, compared to the random data swapping at a given swapping rate because of the 'localized' search for the household pair.

In Table 4 the highest *AD* metric representing the most distortion in cell counts according to the OA geographical level is obtained under the targeted 5% swap and the lowest *AD* metric is obtained under the random 2% swap. The random swapping strategy has lower *AD* metrics than the targeted swapping strategy which means that more bias is introduced into the Census tables at the OA geography due to the

Table 4. Average absolute distance per cell *(AD)* for Census tables with OA geography

Tables (OA)	Random 2%	Random 5%	Target 2%	Target 5%
(1)	0.391	0.732	0.499	0.841
(2)	0.208	0.427	0.238	0.455
(3)	0.266	0.523	0.665	0.916
(4)	0.261	0.496	0.486	0.713
(5)	0.294	0.577	0.338	0.621
(6)	0.272	0.526	0.290	0.555

Table 5. Average absolute distance per cell *(AD)* for Census tables with ward geography

Tables (wards)	Random 2%	Random 5%	Target 2%	Target 5%
(1)	1.627	2.754	1.227	1.547
(2)	1.141	2.034	0.559	0.678
(3)	1.273	2.260	1.708	2.219
(4)	1.567	2.677	1.133	1.528
(5)	2.055	3.468	1.081	1.366
(6)	1.664	2.781	0.822	1.035

targeted selection of households to swap. For most of the Census tables, the targeted 2% swap has less distortion to the cell counts compared to the random 5% swap, with the exception of Census table (3) involving the variable ethnicity. Ethnicity in particular was used for the targeted data swapping strategy for defining high risk and also as an indicator in the control variables for selecting paired households. This likely induced more bias into the table.

Table 5, however, presents a different picture for the Census tables at the aggregated ward geography level. Obviously, the disclosure risk is considerably less when aggregating to the ward level with less possibility of unique cells. The random data swapping shows more distortions per cell than the targeted data swapping for each of the swapping rates. This clearly demonstrates that taking into account the geographical level of risk when pairing households for swapping as implemented in the targeted data swapping strategy ensures much less bias at higher aggregations of geographies.

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: $\chi^2 = \sum_i \sum_j (o_{ij} - e_{ij})^2 / e_{ij}$ where under the null hypothesis of independence: $e_{ij} = (n_{i.} \times n_{.j}) / n$, $n_{i.}$ 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: $RCV\ (T^O, T^P) = 100 \times \frac{CV\ (T^P) - CV\ (T^O)}{CV\ (T^O)}$

Table 6 presents results of the percent relative difference in the Cramer's V Statistic *(RCV)* based on the different data swapping strategies and swapping rates for each of the Census tables in Section 3 according to the OA geography.

Table 6. Percent difference in Cramer's V *(RCV)* for Census tables with OA geography

Tables	Random 2%	Random 5%	Target 2%	Target 5%
(1)	-0.660	-1.228	-1.090	-2.345
(2)	-0.641	-1.525	-0.689	-1.213
(3)	-0.927	-1.614	-1.710	-2.567
(4)	-0.601	-1.380	0.347	-0.884
(5)	-0.920	-1.635	-1.046	-2.065
(6)	-0.573	-0.986	-0.479	-0.941

The values in Table 6 are all generally negative which means that the swapped tables provide a measure of association that is always smaller than the measure of association based on the original table. This implies that data swapping of geographical variables attenuates the distributions in the tables and they lean more towards independence. Table 6 shows mixed results between the random and targeted swapping strategies at the same level of swapping rate. Tables (1), (3), (5) have higher *RCV* under the targeted data swapping while the other tables have lower *RCV*. Again, the reason for this pattern is due to the use of key variables to define high risk households which had a disproportionate chance of being selected for swapping. For those variables, the targeted swapping strategy induces more bias. In general, the 2% targeted swapping have lower values of *RCV* compared to the 5% random swapping, although this is not the case for Census table (3) where the 2% targeted swap has a higher *RCV* than the 5% random swap. Similar results are obtained at the ward level geography.

4.3 R-U Confidentiality Map

In this section, an R-U Confidentiality Map (Duncan, et al., 2001) is presented for the different data swapping strategies on each of the Census tables at the OA geography defined in Section 3. Figure 1 presents the empirical R-U confidentiality map based on the disclosure risk measure *DR* on the y-axis and the distance metric *AD* on the x-axis (note that the x-axis is reversed since a high *AD* represents low utility).

The lower left hand quadrant in Figure 1 represents low utility-low disclosure risk and the upper right hand quadrant high utility-high disclosure risk. In many cases, the 2% targeted data swapping has a lower disclosure risk than the 5% random data swapping and in general higher utility, i.e. smaller distance metrics *AD*. A line on the frontier of the data points is drawn in Figure 1 representing the points with the highest utility at each given disclosure risk. Three of the points are based on the 2% targeted

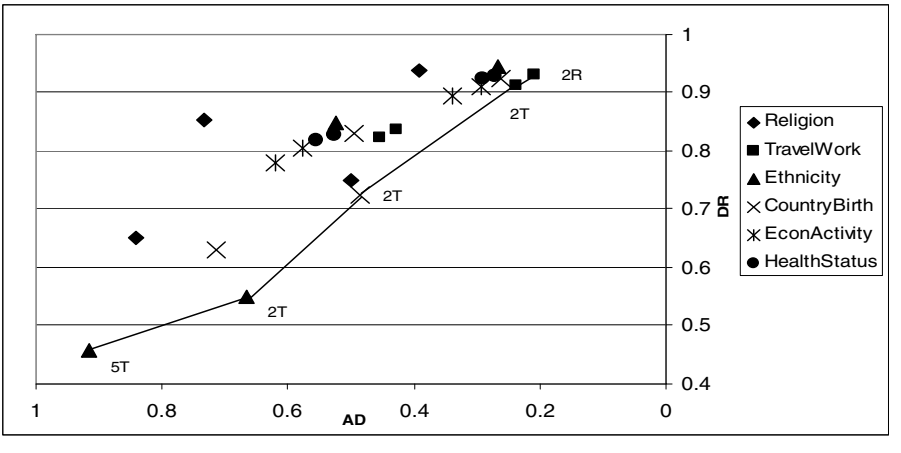

Fig. 1. R-U confidentiality map with *DR* (proportion of unperturbed unique cells) on the y-axis and the *AD* (distance metric) on the x-axis for all Census tables

swap and this would be the preferred option for this analysis based on the swapping rates and swapping strategies studied.

5 Discussion

In general, data swapping as a sole SDC method for protecting Census tables results in high probabilities that small cells in tables are true values. The method should be used in combination with other SDC methods, for example implementing a comprehensive and strict output design strategy with fixed categories of variables, population thresholds, etc. or some small cell masking.

We propose a targeted data swapping strategy which lowers the disclosure risk for a given swapping rate compared to random data swapping, especially for Census tables involving key variables that are used to define high risk households that are targeted for swapping. Higher swapping rates raise the level of protection but also cause more loss of utility. The results from the analysis show that the proposed targeted data swapping lowers disclosure risk approximately equal to that of a random data swapping at double the swapping rate whilst having generally higher utility. The analysis also showed that there are considerable gains using the targeted data swapping strategy compared to a random data swapping strategy, especially when aggregating lower levels of geography.

In any perturbative SDC method that is used to protect statistical data there are hidden non-transparent effects to the data which impacts on the ability to carry out statistical analysis. While the Census tables have the advantage that they are consistent and additive, this is undermined by the inability to obtain confidence intervals that take into account the perturbation. NSIs need to provide information and guidance to users in order to inform them of the impact of SDC methods and how to analyze disclosure controlled statistical data. Quality measures should be disseminated with the release of the Census tables to allow users to try and correct inferences using measurement error models.

References

1. Duncan, G., Keller-McNulty, S., Stokes, S.: Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report LA-UR-01-6428. Statistical Sciences Group. Los Alamos National Laboratory, Los Alamos (2001)
2. Gomatam, S., Karr, A.: Distortion Measures for Categorical Data Swapping. Technical Report Number 131, National Institute of Statistical Sciences (2003)
3. Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., De Wolf, P.P.: Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics 14, 463–478 (1998)
4. Shlomo, N.: Statistical Disclosure Control Methods for Census Frequency Tables. International Statistical Review 75(2), 199–217 (2007)
5. Young, C., Martin, D., Skinner, C.J.: Geographical Intelligent Disclosure Control for Flexible Aggregation of Census Data. International Journal of Geographical Information Science 23(4), 457–482 (2009)
6. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics, vol. 155. Springer, New York (2001)