A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis

Marley, Jennifer K.
Australian Bureau of Statistics, Data Access and Confidentiality Methodology Unit
45 Benjamin Way
Belconnen, ACT, 2617, Australia
E-mail: jennifer.marley@abs.gov.au

Leaver, Victoria L.
Australian Bureau of Statistics, Data Access and Confidentiality Methodology Unit
45 Benjamin Way
Belconnen, ACT, 2617, Australia
E-mail: victoria.leaver@abs.gov.au

Abstract

In August 2009, the Australian Bureau of Statistics (ABS) released Census TableBuilder, a flexible online tool that allows users to define and download tables of Census counts. To prevent disclosure through differencing attacks and repeated requests for identical tables, the ABS developed a method for automatically and consistently confidentialising tables of counts.

A key feature of the method is the perturbation look-up table, a fixed, two-dimensional array of perturbation values. The amount of perturbation applied to each cell in the requested table is determined by accessing this array using the cell count and a cell key calculated from random numeric keys that have been permanently assigned to all records in the underlying microdata.

In this paper, we examine some of the statistical properties of the method. We create three perturbation look-up tables that provide varying levels of protection and use them to perform a risk-utility analysis of the method. The measures of risk considered include the probability that a cell calculated from the difference of two tables will equal the true differenced value. We consider a range of utility measures, including the average absolute deviation.

Keywords: tabular confidentiality, disclosure risk measures, data utility measures.

1. Introduction

In Australia, as in many other countries, there is a strong demand from users of official statistics for flexible access to microdata to enable their social and economic research and analysis. However, like many national statistical institutes, the Australian Bureau of Statistics (ABS) is required by legislation, specifically the Australian Census and Statistics Act 1905, to ensure that the data it collects is not released in a manner that is likely to enable the identification of any individual person or organisation. One way to meet the user need for flexible access to data and still meet legislative obligations to protect respondent confidentiality is through flexible table-building products that incorporate dynamic confidentiality routines.

Specific confidentiality risks arise in an environment where users can obtain many tables of their own design. One risk is that users may be able to undo some of the confidentiality protections by making repeated requests for the same table. If the confidentiality method uses a random process to confidentialise the data, then a user could obtain different versions of the same table. Comparing the cell values across these different versions may reveal some information about the original, unconfidentialised table. The second main risk is "differencing", which can occur if users are able to request tables for similar sub-populations and then take the difference between the two tables to find the data for a much smaller sub-population.

The ABS has developed a method to address these risks, and has implemented it in Census TableBuilder, an online tool that allows users to create customised tables of data from the Australian Census of Population and Housing. The tables contain counts of people, families or dwellings within categories determined by the variables on the Census Output Record File, the underlying microdata feeding into Census TableBuilder. To build a table, users can select combinations of any of the variables on the file, including a wide range of geographic areas. The confidentiality method, originally proposed in Fraser and Wooton (2005), ensures all tables produced by users of Census TableBuilder are confidentialised. Leaver (2009) describes the implementation of the method in Census TableBuilder.

The confidentiality method applies perturbation to every cell in the table. For this reason, all cells in a table, not just the ones deemed sensitive, have a chance of being changed. Risk of identification is thus reduced, but the utility of the data is also reduced to a certain extent. This paper presents a summary of the Census TableBuilder method, examines some possible measures of risk and utility loss, and uses these measures to evaluate three variants of the method that provide different amounts of protection. These variants use different parameters from those implemented in Census TableBuilder.

2. Summary of the Census TableBuilder method

In the underlying microdata, a permanent numeric value known as a "record key" is randomly generated and assigned to each unit record. When a table is requested, the record keys of the units within each cell are combined using modular arithmetic to create a cell-level key. The cell-level key is used to determine the amount of perturbation applied to that cell, via a perturbation look-up table. A perturbation look-up table is a fixed, two-dimensional array of numeric values. The numeric value located in the row dictated by the cell-level key and the column dictated by the original cell value is selected as the cell perturbation amount and then added to the original cell value to produce the perturbed cell count for output.

The record keys and perturbation look-up table used in Census TableBuilder were designed to ensure that the following criteria hold:

1. the perturbations take integer values;
2. the mean of the perturbation values is zero;
3. the perturbations will not produce negative cell values or positive cell values below a specified threshold;
4. the perturbations have a fixed variance; and
5. the absolute value of any perturbation is less than a specified integer value.

Since the perturbations applied to the cells in a table (both internal and marginal cells) are determined independently, it is more than likely that the resulting perturbed table will no longer be additive. In the Census TableBuilder product, there is a subsequent algorithm that restores additivity to the table. However, this paper will not consider the effect of the additivity step on risk or utility.

There is more information about the online tool on the ABS website. See:

- http://www.abs.gov.au/TableBuilder and
- http://www.abs.gov.au/CDATAOnline.

3. Creating Perturbation Look-Up Tables

A perturbation look-up table is constructed by choosing a distribution of perturbation values which minimises disclosure risk subject to information loss constraints. More specifically, entropy (a measure of uncertainty) is maximised subject to bias and variance constraints.

Let Π_i denote the set of integer perturbation values available in perturbation look-up table *i*, where P_{iL} and P_{iU} are the lower and upper bounds of the perturbation values in perturbation look-up table *i*, respectively.

Let *u* denote the original (unconfidentialised) cell count, *c* denote the consistently perturbed (confidentialised) cell count and *p* denote the perturbation value as determined from the perturbation look-up table so that $c = u + p$.

In creating a perturbation look-up table, the aim is to determine an appropriate distribution for *p*, that is $P_i(p \mid u)$ for perturbation look-up table *i*. We do this by maximising entropy, which, for the *i*th perturbation look-up table, is defined as

$$-\sum_{p \in \Pi_i} P_i(p \mid u) \log\left[P_i(p \mid u)\right], \quad u \in \mathbb{N}, \qquad \ldots(1)$$

subject to a set of constraints, namely:

i. The set of available perturbation values, Π_i ;

ii. The basic requirements for a probability distribution, $\sum_{p \in \Pi_i} P_i(p \mid u) = 1$ where $P_i(p \mid u) \geq 0,\ \forall u \in \mathbb{N};$

iii. The confidentialised cell values cannot be negative, $c \geq 0$, nor less than some positive specified value *l*, $c \geq l$;

iv. The perturbation values cannot add bias to the cells in a generated table, $E\left(p \mid u\right)=0,\ \forall u \in \mathbb{N}$; and

v. The variance of the perturbation values cannot exceed a specified threshold v_p, $Var\left(p \mid u\right) \leq v_p,\ \forall u \in \mathbb{N}.$

The perturbation probability distributions can be found by applying the Lagrange multiplier method to the above maximisation problem and then numerically solving the resulting systems of non-linear equations. Approximations to the solutions have to be made to convert the perturbation distributions into look-up table form. That is, for each of the *C* columns (a column corresponds to a *u* value), the *R* rows need to be randomly populated with the appropriate *p* values so as to best approximate the distributions determined from our solution to the maximisation problem. For some integer $X < C$, if $u > C - X$, then the perturbation value *p* that will be added to *u* will come from column $C-(X-1)+\mathrm{mod}(u,X)$.

The amount of perturbation applied by the Census TableBuilder method can be adjusted by altering the perturbation look-up table.

4. Risk-Utility Analysis: Measuring Risk

To assess the risk of the Census TableBuilder confidentiality method, the following three measures of risk will be used:

1) Inverse of the variance of the confidentialised counts

In the Census TableBuilder online tool, it is possible for a cell to appear in many different tables. Given the nature of the confidentiality method, no matter what table this particular cell appears in, the unconfidentialised cell count *u* will always be perturbed by the same perturbation value *p* as determined by the cell-level key and the perturbation look-up table. Although our particular cell will be consistently confidentialised, the resulting confidentialised count *c* is random due to the properties of the method, and in particular, the random allocation of record keys to each unit record in the microdata.

Since our confidentialised count *c* is random, we can consider its variance, which can be expressed as

$$Var\left(c \mid u\right) = Var(u+p \mid u) = Var(p \mid u).$$

From a disclosure risk perspective, the greater the variance of *c*, the lower the risk of discovering *u*. Hence, a measure of risk could be the inverse of the variance of *c*, that is

$$R_1(u) = \left[Var\left(c \mid u\right)\right]^{-1} = \left[Var\left(p \mid u\right)\right]^{-1}. \qquad \ldots(2)$$

2) Percentage of cells that are unchanged

The percentage of cells that are unchanged is equivalent to the percentage of cells that are perturbed by $p = 0$. As this percentage increases, (a) the level of confidentiality protection decreases and (b) the risk of identification increases. This measure of risk can be expressed as

$$R_2(u) = P\left(p = 0 \mid u\right). \qquad \ldots(3)$$

3) Probability of an observed difference of 1 corresponding to a true difference of 1

As mentioned in Section 1, the Census TableBuilder confidentiality method was developed specifically to address the risk associated with differencing attacks. The risk of differencing was mitigated in the method by designing the perturbation look-up table to ensure that the difference between two perturbed cells has a certain variance. However, the method does not guarantee that the difference of two perturbed cells will be perturbed away from the difference of the original cells. One measure of risk could be the probability that an observed difference of 1 corresponds to a true difference of 1. If this probability is high, then the risk of identification due to differencing is high.

To perform a differencing attack, a user will request identical tables for two similar subpopulations U_b and U_s and then take the difference between the two tables to obtain the data for a much smaller subpopulation U_m where,

$$U_m = \begin{cases} U_b \cap U_s, & U_s \not\subset U_b, \\ U_b \setminus U_s, & U_s \subset U_b. \end{cases}$$

Let u_b, p_b and c_b be the original cell count, the perturbation value that will be added to this cell count, and the confidentialised cell count respectively, for a given cell in the table for U_b. Define u_s, p_s and c_s

similarly for U_s. Let $d_u = u_m = u_b - u_s$ denote the true difference for our given cell, that is, the original cell count for U_m. A user will not be able to observe d_u but will be able to observe $d_c = c_m = c_b - c_s$, which is the confidentialised difference for our given cell or, in other words, the consistently perturbed (confidentialised) cell count for U_m.

If an observed difference of 1 corresponds to a true difference of 1, then both cells must have received the same amount of perturbation. That is:

$$P\left(d_u = 1 \mid d_c = 1\right) = P\left(p_b = p_s \mid d_c = 1\right). \qquad \ldots(4)$$

It can be shown that (4) can be expressed in terms of the perturbation probability distributions of the perturbation look-up table. For a general perturbation look-up table where P_L and P_U denote the lower and upper bounds of the perturbation values, respectively, the third measure of risk can be expressed as:

$$R_3(c_b) = P\left(d_u = 1 \mid d_c = 1\right) = \sum_{p=P_L}^{P_U} \left\{ \frac{P\left(p_b = p \mid u_b = c_b - p\right)}{\sum_{p'=P_L}^{P_U} P\left(p_b = p' \mid u_b = c_b - p'\right)} \times \frac{P\left(p_s = p \mid u_s = c_s - p\right)}{\sum_{p'=P_L}^{P_U} P\left(p_s = p' \mid u_s = c_s - p'\right)} \right\}. \qquad \ldots(5)$$

The probability of an observed difference of 1 corresponding to a true difference of 1 may not be the only event considered to be risky. For example, it may be considered that an observed difference of, say, less than or equal to 5 which corresponds to a true difference of 1 or 2 is a risky event. The expression in (5) can be generalised to accommodate such a situation. Let $R_g\left(c_b\right)$ be the general form of $R_3(c_b)$ and

$$R_g\left(c_b\right) = P\left(d_u \in D_u \mid d_c \in D_c\right),$$

where D_u is the set of true difference that are of interest and D_c is the set of observed differences that are of interest. In the given example, $D_u = \{1, 2\}$ and $D_c = \{0, 1, 2, 3, 4, 5\}$.

5. Risk-Utility Analysis: Measuring Utility

The confidentiality method described in this paper introduces perturbation to protect respondents. This perturbation will alter the data and introduce information loss. The impact of the information loss on the utility of the data depends on the type of analysis the user wishes to conduct. In this section, we explore some ways of measuring information and utility loss for tables.

Many information loss measures for tables have been proposed in the literature. The following four measures have been selected for use in this paper:

1) Average percentage change

The Census TableBuilder confidentiality method perturbs cell values independently of each other, where the probability distribution of the perturbation value to be added is dependent on the unconfidentialised cell count. At the cell level, we are interested in the expected or average percentage change of the unconfidentialised cell count. That is,

$$U_1\left(u\right) = APC = \sum_{p'=P_L}^{P_U} \frac{|p'|}{u} P\left(p = p' \mid u\right) \times 100\% \qquad \ldots(6)$$

2) Average absolute difference

The Absolute Average Difference (AAD) is the mean, across the table cells, of the absolute value of the difference between the original and the confidentialised cell values (Shlomo and Young 2005):

$$U_2 = AAD = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} | u_{ij} - c_{ij} | \qquad \ldots(7)$$

where *I* is the number of rows, *J* is the number of columns, u_{ij} is the original cell value in the i^{th} row and j^{th} column, and c_{ij} is the corresponding confidentialised cell value.

3) Mean variation

The Mean Variation (MV) measure is described in Yancey, Winkler and Creecy (2002). It was proposed as an information loss measure for confidentialised microdata, but it can be adapted to measure information loss in confidentialised tables.

$$U_3 = MV = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{|u_{ij} - c_{ij}|}{\sqrt{2}S_{c_{ij}}} \qquad \ldots(8)$$

In (8), $\sqrt{2}S_{c_{ij}}$ is used as a scaling factor. Tables with larger cell values are likely to have larger standard deviations. This scaling factor means that small perturbations applied to tables with large cell values will contribute less to the information loss measure than small perturbations applied to tables with small cell values.

4) Relative difference in Cramer's V

The Relative difference in Cramer's V (RCV) is a measure of the impact of the confidentialisation on the association between the variables in the table (Shlomo and Young 2005). Cramer‟s V is given by:

$$CV = \sqrt{\frac{\chi^2 / n}{\min(I-1, J-1)}} \qquad \ldots(9)$$

where χ^2 is the standard Pearson Chi-squared statistic for the table and $n = \sum_{i=1}^{I}\sum_{j=1}^{J} u_{ij}$. The relative difference in Cramer‟s V is given by:

$$U_4 = RCV = \frac{CV_c - CV_u}{CV_c} \times 100\%, \qquad \ldots(10)$$

where CV_u is the Cramer‟s V for the original, unconfidentialised table and CV_c is the Cramer‟s V for the confidentialised table.

6. Risk-Utility Analysis: Theoretical and Empirical Results

For the investigations in this paper, we have created three new look-up tables, which should provide varying amounts of protection. They were created as outlined above in Section 3, where the specific constraints for maximising (1) for each look-up table are as follows:

❖ *Perturbation Look-Up Table 1*

i. $\Pi_1 = \{-1, 0, 1\}$;

ii. $\sum_{p\in\Pi_1} P_1(p \mid u) = 1$ where $P_1(p \mid u) \geq 0,\ \forall u \in \mathbb{N}$;

iii. $c \geq 0$;

iv. $E(p \mid u) = 0,\ \forall u \in \mathbb{N}$; and

v. $Var(p \mid u) = 0.665,\ \forall u \in \mathbb{N}$.

❖ *Perturbation Look-Up Table 2*

i. $\Pi_2 = \{p \in \mathbb{Z} : -5 \leq p \leq 5\}$;

ii. $\sum_{p\in\Pi_2} P_2(p \mid u) = 1$ where $P_2(p \mid u) \geq 0,\ \forall u \in \mathbb{N}$;

iii. $c \geq 0$ and $c \notin \{1, 2, 3, 4\}$;

iv. $E(p \mid u) = 0,\ \forall u \in \mathbb{N}$; and;

v. $Var(p \mid u) \leq 8,\ \forall u \in \mathbb{N}$.

❖ *Perturbation Look-Up Table 3*

i. $\Pi_3 = \{p \in \mathbb{Z} : -20 \leq p \leq -1 \text{ and } 5 \leq p \leq 20\}$;

ii. $\sum_{p\in\Pi_3} P_3(p \mid u) = 1$ where $P_3(p \mid u) \geq 0,\ \forall u \in \mathbb{N}$, and

- $P_3(p = -u \mid u < 10) > 0$,
- $P_3(p = -u \mid u \geq 10) = 0$, and
- $P_3(p \mid u) = 0$ for p where c not $\equiv 0 (\text{mod} 5)$;

iii. $c \geq 0$;

iv. $E(p \mid u) = 0,\ \forall u \in \mathbb{N}$; and

v. $Var(p \mid u) \leq 100,\ \forall u \in \mathbb{N}$.

The first look-up table assigns a maximum perturbation of ±1 to the unconfidentialised cell counts. The second look-up table assigns a maximum perturbation of ±5 to the unconfidentialised cell counts, and ensures that there are no cells with confidentialised values between 1 and 4 inclusive. The third look-up table assigns a maximum perturbation of ±20, and is designed to ensure that no cell values are unchanged, and that all confidentialised cell values are multiples of 5. For perturbation look-up tables 2 and 3, the specific variance constraint may vary across different values of *u*. All three perturbation look-up tables have $C = 30$ columns, where the last $X = 10$ columns are cycled through for $u > 20$.

Clearly, we would expect perturbation look-up table 1 to cause the least amount of information loss but not provide very much protection. Perturbation look-up table 3, on the other hand, provides a great deal of confidentiality protection but will cause quite a substantial amount of information loss since most values are perturbed by at least 5 and at most 20 to a value that is a multiple of 5. We would expect perturbation look-up table 2 to fall somewhere between the other two in terms of the confidentiality protection it provides and the information loss it causes.

For the three look-up tables, theoretical values for all three risk measures and the utility loss measure $U_1(u)$ were simply calculated using the formulae given in (2), (3), (5) and (6), respectively. Notice that all of these measures are calculated at the cell level. The remaining utility loss measures are calculated at the table level and are dependent on the table; theoretical values were not calculated for these measures.

Empirical values were calculated for all three risk measures and all four utility loss measures. For $R_1(u)$ and $R_2(u)$, the empirical values were calculated from 1000 confidentialised cells for each original cell count $u \in \{1,2,...,30\}$ which were simulated by randomly generating record keys and following the confidentiality method outlined in Section 2. For $R_3(c_b)$ and the four utility loss measures, the empirical values were calculated from a set of ten test tables, some with a broad level of detail and some with a fine level of detail, which were created from a synthetic dataset.

Measures of risk

- *Inverse of the variance of the confidentialised counts*

The graphs in Figure 1 show the theoretical and empirical values of $R_1(u)$ for unconfidentialised cell values $u \in \{1,2,...,30\}$ for each perturbation look-up table. The open circles are the theoretical values while the filled-in circles are the empirical values based on the 1000 simulated cells.

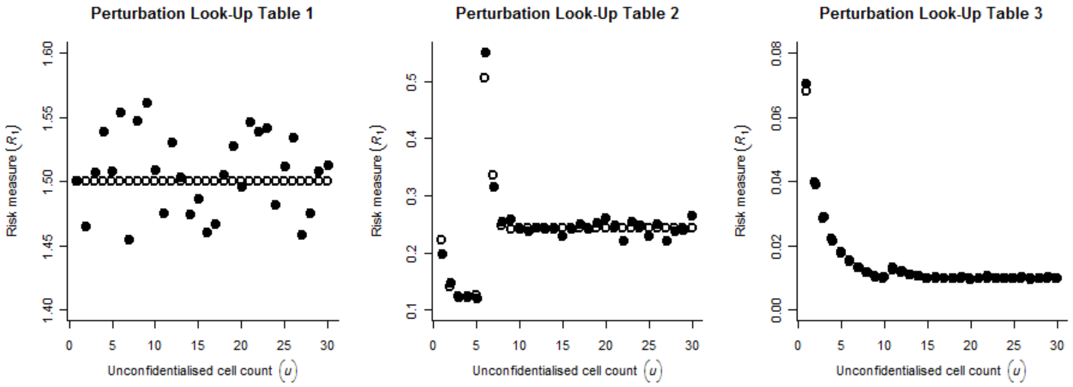

Figure 1 – Theoretical and empirical values of $R_1(u)$ for each perturbation look-up table.

For perturbation look-up table 1, we expect the inverse of the variance of the confidentialised cell values to be constant over all values of *u* but the empirical values appear to be rather variable and appear to deviate more from the theoretical values than those for perturbation look-up tables 2 and 3. The empirical values for perturbation look-up table 3 appear to follow the pattern of the theoretical values very closely. Relative to the theoretical values, however, the empirical values of $R_1(u)$ are within 5%, 12% and 7% of the theoretical values for perturbation table 1, 2 and 3 respectively. As expected, the values of $R_1(u)$ for perturbation look-up table 3 are clearly less than the equivalent values for perturbation look-up table 2 which are clearly less than the equivalent values for perturbation look-up table 1, indicating their order in terms of the amount of confidentiality protection they provide.

- *Percentage of cells that are unchanged*

The graphs in Figure 2 show the theoretical and empirical values of $R_2(u)$ for unconfidentialised cell values $u \in \{1,2,...,30\}$ for perturbation look-up tables 1 and 2. Perturbation look-up table 3 was designed to perturb all cell values away from the original cell value and thus contains no zeros. Consequently, a graph for perturbation look-up table 3 was not included since the probability of any cell remaining unchanged after the perturbation process is zero. Again, the open circles are the theoretical values while the filled-in circles are the empirical values based on the 1000 simulated cells.

Since the probability distribution of the perturbation values is the same for all values of *u* in perturbation look-up table 1, we expect the proportion of cells to be unchanged after the perturbation process to be constant at 33.6% for all values of *u*. The empirical values in the first graph in Figure 2 follow this constant pattern closely. The empirical values for perturbation look-up table 2 also follow closely the pattern of the corresponding theoretical values in the second graph. Note that original cell values of 1, 2, 3 and 4 are never unchanged after the confidentiality method has been applied; this property was designed for in the creation of perturbation look-up table 2. Again, as expected, the values of this risk measure are always greater for perturbation look-up table 1 than perturbation look-up table 2 for all values of *u*, indicating that the former provides less confidentiality protection than the latter.

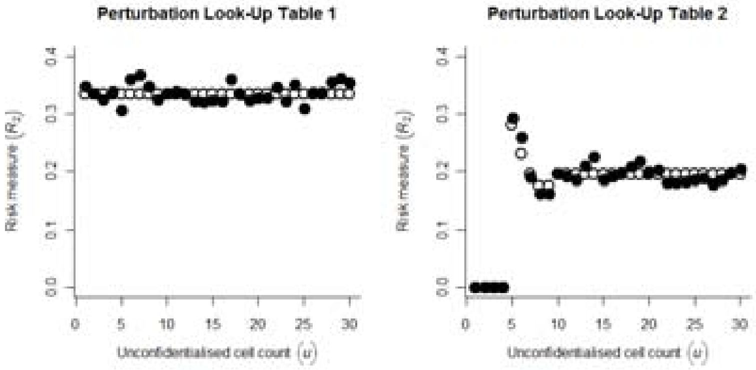

Figure 2 – Theoretical and empirical values of $R_2(u)$ for perturbation look-up tables 1 and 2 .

- *Probability of an observed difference of 1 corresponding to a true difference of 1*

The graphs in Figure 3 show the theoretical and empirical values of $R_3(c_b)$ for confidentialised cell values $c_b \in \{1,2,...,30\}$ for perturbation look-up tables 1 and 2. For perturbation look-up table 3, the value of $R_3(c_b)$ for all values of c_b is zero since it was designed to ensure that all confidentialised cell values are multiples of 5, meaning the smallest observable non-zero difference between two cells is 5. Therefore, the graph in Figure 3 corresponding to perturbation look-up table 3 shows the theoretical and empirical values of a variant of $R_3(c_b)$, namely,

$$R_4(c_b) = P\left(d_u = 1 \mid d_c = 5\right),$$

for confidentialised cell values $c_b \in \{1,2,...,30\}$. The open circles are the theoretical values while the filled-in circles are the empirical values based on the cells in the test tables that had a confidentialised value of c_b.

All of the empirical values in Figure 3 were calculated from far fewer cells than those calculated for the two previous risk measures. For most values of c_b, the number of cells from which the empirical values of $R_3(c_b)$ were calculated was ≤ 100, which explains some of the deviation from the theoretical values that we see in the graphs. In particular, the empirical values for $c_b \geq 18$ for perturbation look-up table 1 and $c_b \geq 20$ for perturbation look-up table 2 were calculated from ≤ 40 cells. For corresponding values of c_b, both the theoretical and empirical values of $R_3(c_b)$ for perturbation look-up table 1 are

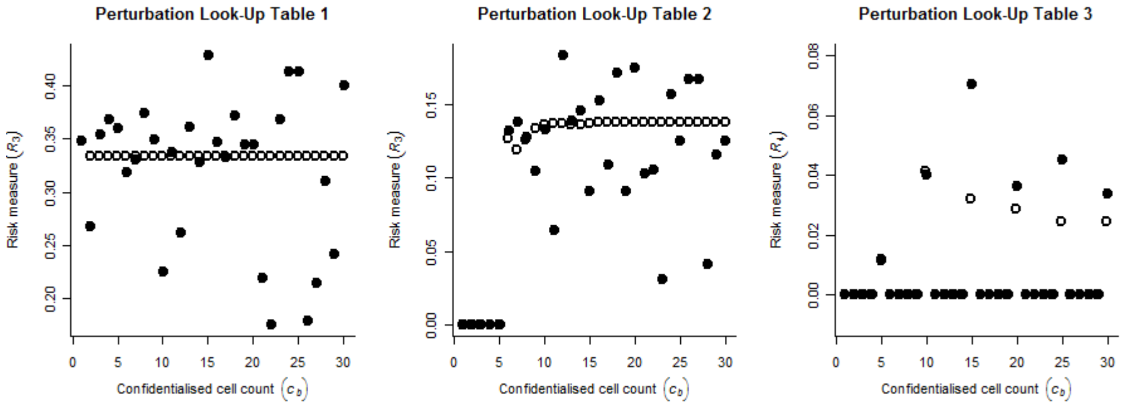

Figure 3 – Theoretical and empirical values of $R_3(c_b)$ for each perturbation look-up table.

greater than those for perturbation look-up table 2. Although the risk measure, $R_4(c_b)$, used for perturbation look-up table 3 is not the same as the risk measure $R_3(c_b)$, used for perturbation look-up tables 1 and 2, they are still comparable in the sense that they are both the probability of the smallest observable positive difference corresponding to a small true difference. As expected, for corresponding values of c_b, the values of $R_4(c_b)$ for perturbation look-up table 3 are less than the values of $R_3(c_b)$ for perturbation look-up tables 1 and 2.

Measures of utility

- *Average percentage change*

The graphs in Figure 4 show the theoretical and empirical values of $U_1(u)$ for unconfidentialised cell values $u \in \{1,2,...,30\}$ for each perturbation look-up table. The open circles are the theoretical values while the filled-in circles are the empirical values based on the cells in the 10 test tables that had an unconfidentialised value of *u*.

For some values of *u*, the number of cells from which the empirical value of $U_1(u)$ was calculated was ≤ 40 but this hasn‟t caused any major deviations between the empirical and theoretical values. All three graphs in Figure 4 follow the same pattern of the percentage change due to perturbation decreasing as the unconfidentialised cell count increases. We expect perturbation look-up table 3 to do the most damage to the utility of the data and this is reflected in the values of $U_1(u)$ corresponding to look-up table 3 being greater than those for perturbation look-up tables 1 and 2.

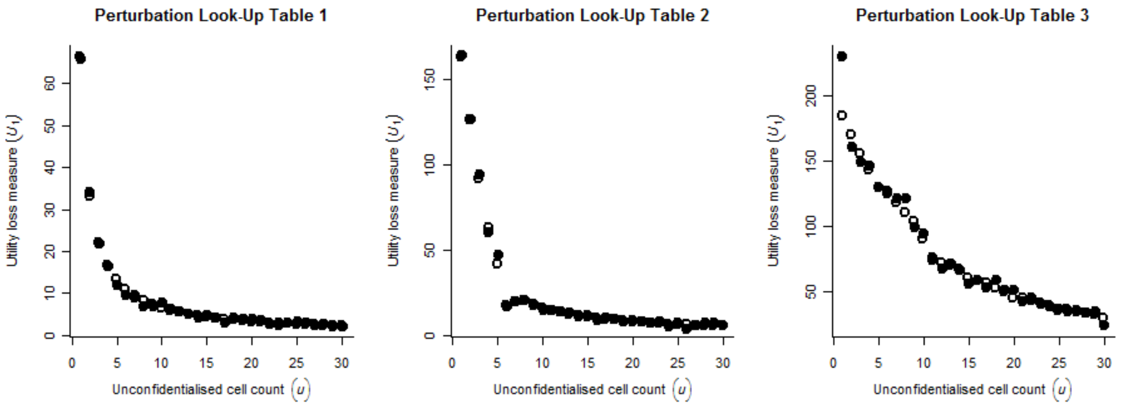

Figure 4 – Theoretical and empirical values of $U_1(u)$ for each perturbation look-up table.

- *Average Absolute Difference*

The AAD provides a simple measure of the amount of change introduced into a table by the perturbation. It is a useful measure for this situation, because it can be applied to tables with more than one dimension, and to tables containing cells with an original count of zero. This measure could be applied to all of the test tables. The graphs in Figure 5 show the AAD measure against the number of cells in the table, for each test table and each perturbation look-up table.

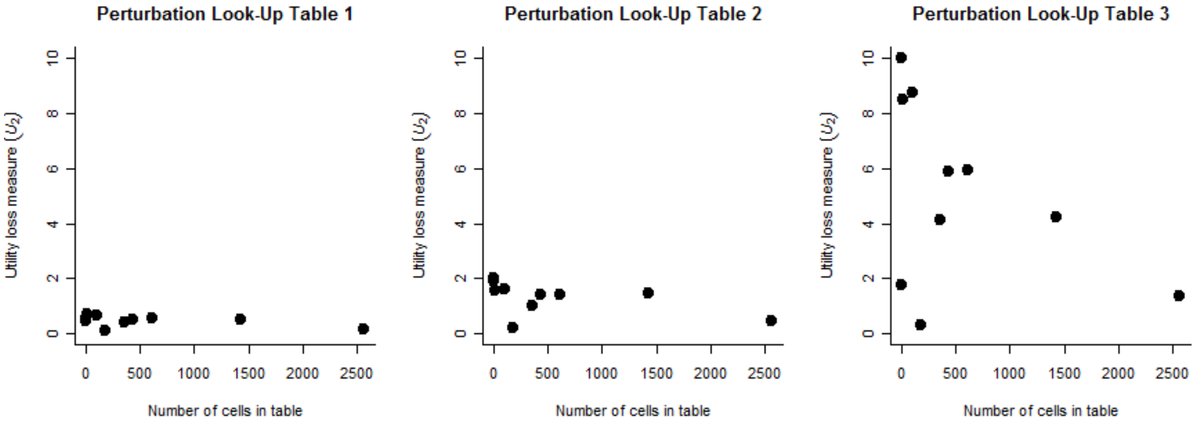

Figure 5 – Empirical values of U_2 for each perturbation look-up table.

The first look-up table would be expected to introduce the smallest amount of perturbation, and the third look-up table would be expected to introduce the largest amount of perturbation. The AAD measure reflects this. For all of the 10 test tables, the perturbation given by the first look-up table has the smallest value for the AAD, and the perturbation given by the third look-up table has the largest value for the AAD.

The AAD tends to be larger for tables that contain larger cell values. The next information loss measure introduces a scaling factor to adjust for the size of the cell values.

- *Mean Variation*

The graphs in Figure 6 show the MV measure against the number of cells in the table, for the set of test tables and for each perturbation look-up table.

As with the AAD, the MV measure reflects the amount of perturbation applied by the different look-up tables. Across all the test tables, the perturbation given by the first look-up table takes the smallest value for the MV, and the perturbation given by the third look-up table takes the largest value for the MV. However, the MV takes lower values for the tables with the large cell values. This measure is less affected by the size of the cell values.

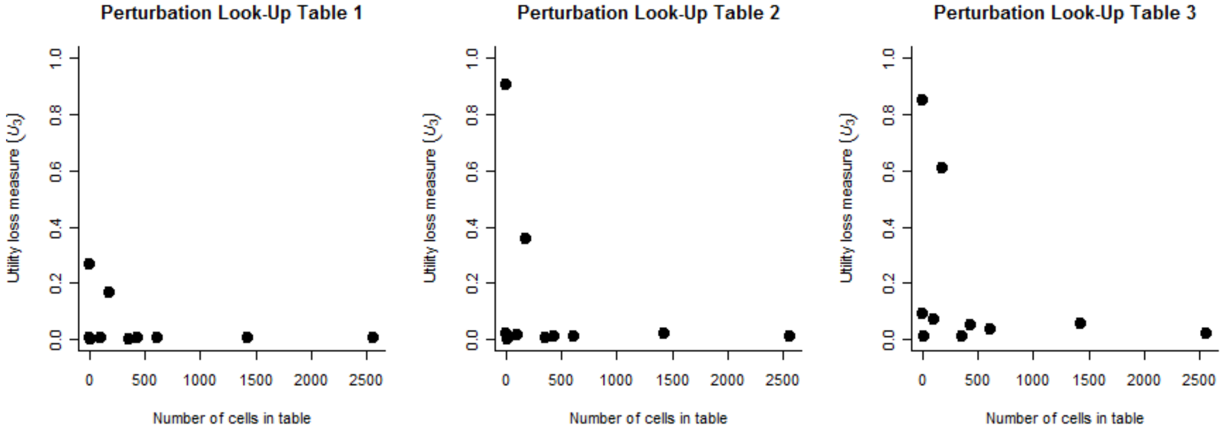

Figure 6 – Empirical values of U_3 for each perturbation look-up table.

- *Relative Difference in Cramer's V*

The RCV measure is only meaningful for two-dimensional tables. For our test tables, the RCV was applied to sub-tables created by taking the first two dimensions of each table. Some tables were excluded from this analysis because their structure was not appropriate for this test. The graphs in Figure 7 show the RCV measure against the number of cells in the table, for the test tables where this measure was appropriate.

For most of the tables, the RCV is largest for the perturbation given by the third look-up table. Generally, the added perturbation increases the association in the tables. However, this may have occurred because the test tables were generated from a synthetic data file, and the original tables did not contain strong associations.

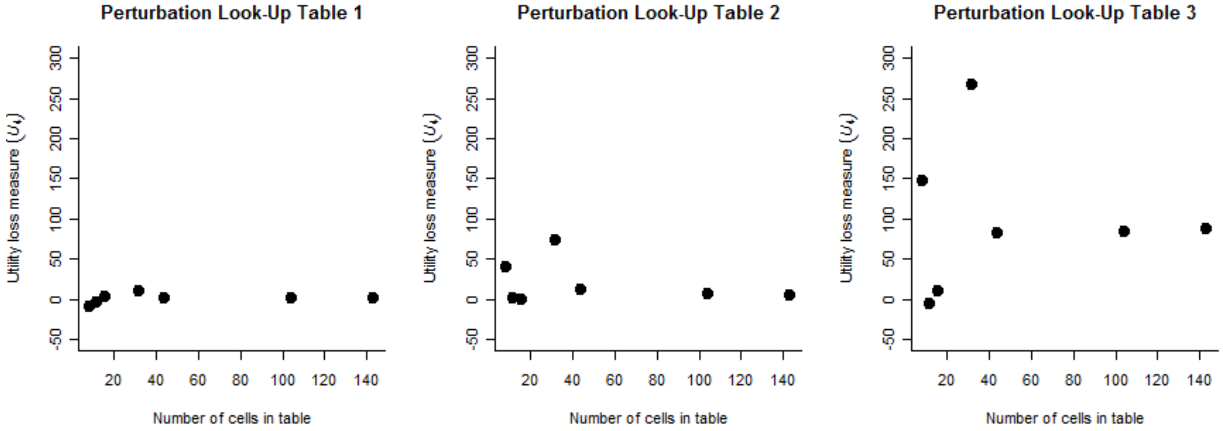

Figure 7 – Empirical values of U_4 for each perturbation look-up table.

Risk-Utility Curve

Some of these risk and utility measures can be combined to create a "risk-utility curve". For example, the first graph in Figure 8 shows the empirical values of $U_1(u)$ against the empirical values of $R_1(u)$ for all values of $u \in \{1, 2, ..., 30\}$, for the three look-up tables. The three look-up tables are clearly divided, with look-up table 1 in the area of the graph indicating little utility loss but a great amount of identification risk, look-up table 3 in the area of the graph indicating a great amount of utility loss but little identification risk, and look-up table 2 in the area between look-up tables 1 and 2. It is interesting to see that for most values of *u* (i.e. $u \geq 6$), the amount of utility loss that perturbation look-up table 2 causes is similar to that caused by perturbation look-up table 1 but identification risk is considerably decreased when look-up table 2 is used rather than look-up table 1. The second graph in Figure 8 shows, for the ten test tables, the MV measure against $R_2(u)$ (the number of unchanged non-zero cells, as a proportion of the number of cells whose original value was non-zero) for perturbation look-up tables 1 and 2. There is, again, a clear division between the two look-up tables: the points on the graph corresponding to perturbation look-up table 2 are in the area of the graph indicating less risk of identification but more utility loss.

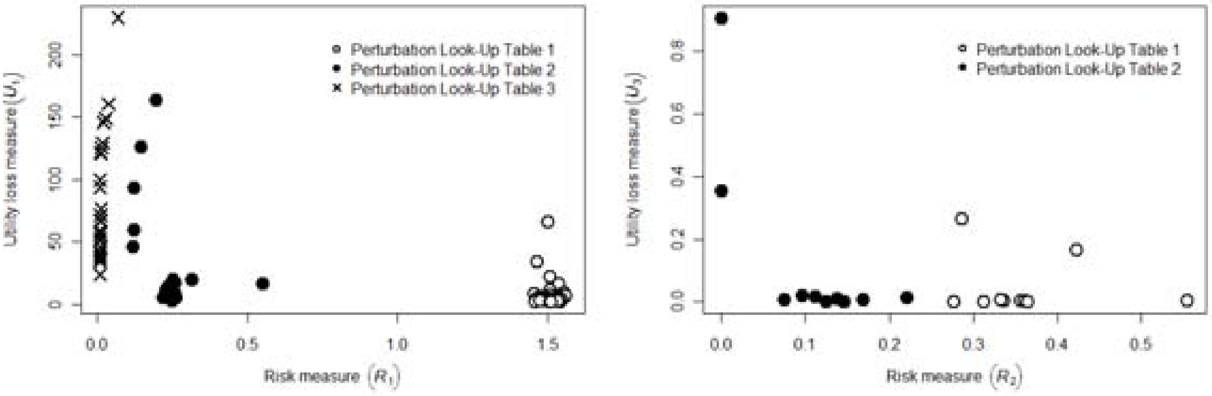

Figure 8 – Empirical risk-utility curves for the perturbation look-up tables.

7. Summary

The Census TableBuilder method can be adapted to provide different levels of perturbation, by altering the perturbation look-up table. The effect of varying the look-up table can be assessed using various measures of risk and utility. Although these measures tend to capture only certain aspects of risk and utility, their results are generally consistent given the properties of the different look-up tables. The measures may be useful when determining suitable parameters for the look-up table in different situations.

Acknowledgements

We"d like to thank Melissa Gare and Christine O"Keefe warmly for their very helpful comments during the revision of this paper.

REFERENCES

Fraser, B. and Wooton, J. 2005. „A proposed method for confidentialising tabular output to protect against differencing". *Presented at UNECE Work Session on Statistical Data Confidentiality*, November, 2005.

Leaver, V. 2009. „Implementing a method for automatically protecting user-defined Census tables". *Presented at Joint UNECE/Eurostat work session on statistical data confidentiality*, December , 2009.

Shlomo, N., and Young, C. 2005. „Information Loss Measures for Frequency Tables". *Presented at UNECE Work Session on Statistical Data Confidentiality*, November, 2005.

Yancey, W., Winkler, W., and Creecy, R. 2002. „Disclosure Risk Assessment in Perturbative Microdata Protection". *Inference Control in Statistical Databases: From Theory to Practice*. Springer: Berlin.