Measuring Disclosure Risk and Information Loss in Population Based Frequency Tables

László Antal, Natalie Shlomo, Mark Elliot
University of Manchester
laszlo.antal@postgrad.manchester.ac.uk

8 September 2014

1 Introduction

Frequency tables disseminated by statistical agencies have always been of high interest. However, the agencies have to ensure that the risk of identifying individuals and disclosing individuals' attributes from the released data is low. Therefore they assess the risk of disclosure and apply statistical disclosure control (SDC) methods if necessary. The main objective of this work is to measure dislosure risk in population based frequency tables. The disclosure risk assessment of such tables is often based on the so-called threshold rule. A cell of the table is of high disclosure risk according to this rule if the cell value does not exceed a certain threshold, for example 2.

In this work we propose to measure the disclosure risk in an alternative way. Our approach takes the entire table (and rows/columns of the table) into consideration. We introduce a disclosure risk measure, which is based on information theoretical definitions, such as the entropy and the conditional entropy.

There are two main types of SDC methods. Pre-tabular methods, such as record swapping, alter the values of a variable (or more variables) for selected individuals

in the microdata. Tables might be generated from the perturbed microdata set. Post-tabular methods, such as random rounding, perturb tabular data. Once a table has been generated, its cell values are modified according to the perturbation method. In this work we deal with post-tabular perturbation methods for population based frequency tables.

The disclosure risk of a frequency table should be assessed not just before but also after an SDC method has been applied to the table. If the disclosure risk is high after perturbation, then the agency might reject the release of the table. The disclosure risk measure we introduce follows this approach.

The main objective of this work is to measure attribute disclosure. Attribute disclosure occurs if a respondent's (or more respondents') previously unknown attribute can be revealed from the disseminated data. Attribute disclosure might happen if a row/column of a frequency table includes one populated cell, the other cells are zeroes.

Our aim is to develop a disclosure risk measure around the following properties.

Property 1A If only one cell is populated in the table, then the disclosure risk is high.

Property 1B Uniformly distributed frequencies imply low risk.

Property 2 Small cell values (i.e. ones and twos) are more disclosive than higher values. In general, the greater the cells, the lower the disclosure risk.

Property 3 Assume that two tables are given and there is only one cell populated in each table. The frequencies of the non-zero cells are equal. In this case we deem the table that has more cells (and therefore more zeroes) to be of higher disclosure risk.

Property 4 We would like the disclosure risk measure to be bounded by 0 and 1.

The motivation behind the properties is as follows. The risk of attribute disclosure is normally high if the population is concentrated in one cell, see Shlomo (2007). It

explains Property 1A. On the other hand, attribute disclosure is unlikely to occur if the frequencies are uniformly distributed, which drives Property 1B. The ground of Property 2 is the fact that revealing new information about a respondent becomes more difficult as the cell frequencies increase. The rationale behind Property 3 is that a table may be a more detailed version of another table, for example the breakdown of a table-spanning variable might be different in two tables. For example, if we replace super output area with output area, then the table will contain more detailed information. An intruder may obtain more information from more detailed tables. Property 4 is driven by the desire of comparing the disclosure risk of different tables.

2 Notation

Throughout the text we use the following notations. The number of cells in the frequency table(s) is K. The cells are denoted $C = \{c_1, c_2, \ldots, c_K\}$. Since we put the emphasis on the comparison of the original and the perturbed tables, we need at least two different notations for the used objects, such as probability distributions, frequencies, frequency tables. The vectors of frequencies in the original and perturbed tables are $F = (F_1, F_2, \ldots, F_K)$ and $G = (G_1, G_2, \ldots, G_K)$ respectively. The sum of the frequencies are $N = \sum_{i=1}^{K} F_i$ and $M_G = M = \sum_{i=1}^{K} G_i$. It means implicitly that the population consists of N individuals. Although the M_G notation is more precise than M and it shows that there may be more potential outcomes of the perturbation, dropping the subscript typically does not lead to confusion. Therefore, we prefer M and use M_G if it is inevitable. The set of individuals will be denoted $I = \{a_1, a_2, \ldots, a_N\}$. Two probability distributions, $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$, will provide the probability that an individual falls in cell c_i in the original and in the perturbed table respectively. In few cases we will use the uniform distribution. If A is a (finite) set (for example C or I), then U_A is the uniform distribution on A. The power set of A is $\mathcal{P}(A) = \{B : B \subseteq A\}$. The cardinality of A is $|A|$.

Although in this work we consider only post-tabular perturbation methods, i.e. the perturbation is carried out on the F frequency table, we need to define random variables on the set of individuals. According to the general definition of random variables, we consider them as measurable functions. The

$$X : (I, \mathcal{P}(I)) \to (C, \mathcal{P}(C)) \tag{2.1}$$

variable will indicate where the individuals fall in C. If we attach the U_I measure to the $(I, \mathcal{P}(I))$ measurable space, then the distribution of the X variable will be $\left(\frac{F_1}{N}, \frac{F_2}{N}, \ldots, \frac{F_K}{N}\right)$. This distribution is equal to P.

The generation of a frequency table can be referred to as the categorization of individuals into cells. (In the situation above the categorization is given by X.) The perturbed frequency table can be also imagined as the categorization of individuals into C. The number of individuals must be equal to M. We denote these imaginary individuals by $J = \{b_1, b_2, \ldots, b_M\}$. The categorization of the b_k, $k = 1, 2, \ldots, M$ individuals should be given by a similar random variable to (2.1).

$$(J, \mathcal{P}(J)) \to (C, \mathcal{P}(C))$$

We attach the U_J distribution to the $(J, \mathcal{P}(J))$ measurable space. The distribution of the random variable should be $\left(\frac{G_1}{M}, \frac{G_2}{M}, \ldots, \frac{G_K}{M}\right)$. However, there may be more than one random variables with this property; the image of a single individual and therefore the variable is not uniquely determined. We denote the set of variables with $\left(\frac{G_1}{M}, \frac{G_2}{M}, \ldots, \frac{G_K}{M}\right)$ distribution by Ω_G.

$$\Omega_G = \{Y : (J, \mathcal{P}(J)) \to (C, \mathcal{P}(C)) : \quad |\{b_k \in J : Y(b_k) = c_l\}| = G_l, l = 1, 2, \ldots, K\}$$

The elements of Ω_G are $Y_1, Y_2, \ldots, Y_{|\Omega_G|}$. If we need to select only one of these variables, then we will omit the subscript and refer to the variable as Y.

3 Disclosure Risk Measure

3.1 Information Theoretical Definitions

The field of information theory and related areas are covered comprehensively in Cover and Thomas (2006). Entropy is one of the most important formulae. It is defined for random variables and depends on their distribution. Although we use the X and Y variables as defined in Section 2, the definitions and properties in the present section can be equivalently rewritten for arbitrary random variables. The sum determining the entropy of X is:

$$H(X) = -\sum_{i=1}^{K} p_i \cdot \log p_i \tag{3.1}$$

If $p_i = 0$ for a certain i, the respective term in the sum will be considered 0, since $\lim_{x \to 0} x \cdot \log x = 0$.

The conditional entropy of two variables is defined as follows.

$$H(X|Y) = -\sum_{j=1}^{K} Pr(Y = c_j) \cdot \sum_{i=1}^{K} Pr(X = c_i|Y = c_j) \cdot \log Pr(X = c_i|Y = c_j)$$

If $Pr(X = c_i|Y = c_j) > 0$ for all i and j, then the conditional entropy can be rewritten using the $Pr(Y = c_j|X = c_i)$ probabilities, as it can be found in Willenborg and de Waal (2001).

$$H(X|Y) = -\sum_{i=1}^{K}\sum_{j=1}^{K} Pr(X = c_i) \cdot Pr(Y = c_j|X = c_i) \cdot \log \frac{Pr(X = c_i) \cdot Pr(Y = c_j|X = c_i)}{\sum_{k=1}^{K} Pr(X = c_k) \cdot Pr(Y = c_j|X = c_k)} \tag{3.2}$$

It is well-known that $H(X|Y) \leq H(X)$.

Another important information theoretical formula is the f-divergence of two distributions. The concept of f-divergence can be found in Csiszár (1967) and Csiszár and Shields (2004). To define an f-divergence we need two probability distributions (P and Q) and an $f : \mathbb{R}^+ \to \mathbb{R}$ convex function. For the sake of convenience we

assume that $f(1) = 0$. The divergence between the two distributions determined by f is

$$D_f(P\|Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) \tag{3.3}$$

f-divergences are non-negative, $D_f(P\|Q) \geq 0$.

Two important f-divergences are the relative entropy and the (square of) Hellinger distance. We obtain the relative entropy (or Kullback-Leibler divergence) if $f(x) = x \cdot \log x$.

$$D(P\|Q) = D_{x\cdot\log x}(P\|Q) = \sum_{i=1}^{K} p_i \cdot \log\frac{p_i}{q_i}$$

It is important to note that

$$D(P \parallel U_C) = \sum_{i=1}^{K} p_i \cdot \log\left(\frac{p_i}{1/K}\right) = \sum_{i=1}^{K} p_i \cdot \log p_i + \sum_{i=1}^{K} p_i \cdot \log K = \log K - H(P)$$

The Hellinger distance is the L_2-norm of a vector by definition. Introduce the $\sqrt{P} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K})$ and $\sqrt{Q} = (\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_K})$ notation. The definition of the Hellinger distance is

$$HD(P,Q) = \frac{1}{\sqrt{2}} \cdot \|\sqrt{P} - \sqrt{Q}\|_2 = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2}$$

If we expand the latter expression, we will obtain $HD(P,Q) = \sqrt{1 - \sum_{i=1}^{K} \sqrt{p_i q_i}}$. Consider the f-divergence given by $f(x) = 1 - \sqrt{x}$. (3.3) provides

$$D_f(P\|Q) = D_{1-\sqrt{x}}(P\|Q) = 1 - \sum_{i=1}^{K} \sqrt{p_i q_i} = HD^2(P,Q)\ .$$

3.2 Disclosure Risk Measure Before Perturbation

We will denote the set of zeroes in the F table by D.

$$D = \{c_i \in C : F_i = 0\}$$

The disclosure risk measure is the convex combination of the proportion of zeroes, an entropy based term and a term depending on the N population size. The proportion of zeroes accounts for Properties 1A and 3, the entropy based term for Properties 1A and 1B, while the third term for Property 2. Each term will be bounded by 0 and 1, therefore Property 4 also satisfies. $\mathbf{w} = (w_1, w_2, w_3)$ is the vector of weights in the disclosure risk measure below, $w_1 + w_2 + w_3 = 1$, $w_1, w_2, w_3 \geq 0$.

$$R_1(F, \mathbf{w}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}}$$

Here e is the base of the natural logarithm.

3.3 Disclosure Risk Measure After Perturbation

Denote the set of zeroes in G by E.

$$E = \{c_i \in C : G_i = 0\}$$

The disclosure risk after perturbation is as follows.

$$R_2(F, G, \mathbf{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \frac{H(X|Y)}{H(X)} - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}}$$

The first and second terms are never higher than the respective terms of $R_1(F, \mathbf{w})$. The third term is unaltered compared to $R_1(F, \mathbf{w})$. Consequently, the $R_2(F, G, \mathbf{w}) \leq R_1(F, \mathbf{w})$ always satisfies. It coincides with the expectation that the disclosure risk after perturbation should not exceed that before perturbation. In order to make the above expression exact, we discuss the properties of the $H(X|Y)$ conditional entropy below.

3.3.1 The conditional entropy

In order to reduce the entropy-based term of the risk measure, we use the conditional entropy. The $H(X|Y) \leq H(X)$ relationship helps to decrease the second term of the disclosure risk measure.

To define the $H(X|Y)$ conditional entropy, we need to ensure that X and Y are defined on the same probability space. Therefore, we assume temporarily that $I = J$. It implies that $N = M$.

According to (3.2) we need to express the $Pr(Y = c_j|X = c_i)$ conditional probabilities in order to determine the conditional entropy. (The $Pr(X = c_i)$ probabilities can be estimated by F_i/N.) $Pr(Y = c_j|X = c_i)$ varies if we select different elements of Ω_G. The choice of the $Y \in \Omega_G$ variable is arbitrary. We may assume that an R_G probability distribution is given on Ω_G and we select the Y variable according to this distribution. $R_G(Y)$ is the probability that Y is selected from Ω_G. Another option is as follows. Once the R_G distribution is given, the expectation of the $Pr(Y = c_j|X = c_i)$ probabilities can be calculated for every fixed i and j. For every i and j we define a

$$Z_{ij}^G = Z_{ij} : \Omega_G \to [0, 1]$$

variable. By definition, $Z_{ij}(Y_l) = Pr(Y_l = c_j|X = c_i)$. The expectation is

$$E(Z_{ij}) = \sum_{l=1}^{|\Omega_G|} R_G(Y_l) \cdot Pr(Y_l = c_j|X = c_i) \tag{3.4}$$

The following theorem can be formulated if R_G is the uniform distribution.

Theorem 1. *If $R_G = U_{\Omega_G}$, then*

$$E(Z_{ij}) = \begin{cases} G_j/N & \text{if } F_i > 0 \\ 0 & \text{if } F_i = 0 \end{cases}$$

Proof. The proof can be found in the Appendix. □

Note that in Theorem 1 the value of $E(Z_{ij})$ does not depend on i. In this case (3.2), calculated on $E(Z_{ij})$, is

$$H(X|Y) = -\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{F_i}{N} \cdot \frac{G_j}{N} \cdot \log \frac{\frac{F_i}{N} \cdot \frac{G_j}{N}}{\sum_{k=1}^{K} \frac{F_k}{N} \cdot \frac{G_j}{N}} =$$
$$-\sum_{j=1}^{K} \frac{G_j}{N} \sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} = H(X)$$

This equation proves that the entropy cannot be lowered by the conditional entropy if $R_G = U_{\Omega_G}$. However, the risk measure should show difference between the original disclosure risk and the disclosure risk after perturbation. Therefore, we select a new R_G distribution and calculate the conditional entropy accordingly.

The new R_G assigns positive probability to a Y_l variable if the number of $a \in I$ individuals with $X(a) = Y_l(a)$ is maximal. The criterion means that the highest number of individuals remain in the same cell after perturbation. If a Y_l variable does not satisfy this criterion, then its probability is zero, $R_G(Y_l) = 0$. The positive probabilities are distributed uniformly.

For a fixed j consider the $X^{-1}(c_j) = \{a \in I : X(a) = c_j\}$ set of individuals. Obviously, $|X^{-1}(c_j)| = F_j$. The highest number of individuals in $X^{-1}(c_j)$ that can remain in the same c_j cell after perturbation is $\min(F_j, G_j)$. Therefore, the highest number of individuals in the population that can remain in the same cell is $\sum_{j=1}^{K} \min(F_j, G_j)$. Consequently, the new R_G distribution assigns positive probabilities to the variables in the following set, denoted by Ω_G^*.

$$\Omega_G^* = \left\{ Y_l \in \Omega_G : |\{a \in I : X(a) = Y_l(a)\}| = \sum_{j=1}^{K} \min(F_j, G_j) \right\}$$

Our aim is to determine the (3.4) average using the new R_G distribution.

Theorem 2. *Assume that $F \neq G$ and*

$$R_G(Y_l) = \begin{cases} 1/|\Omega_G^*| & \text{if } Y_l \in \Omega_G^* \\ 0 & \text{if } Y_l \notin \Omega_G^* \end{cases}$$

Then

$$E(Z_{ij}) = \sum_{Y_l \in \Omega_G^*} R_G(Y_l) \cdot Pr(Y_l = c_j | X = c_i) =$$

$$\begin{cases} \dfrac{\min(F_i, G_i)}{F_i} & \text{if } i = j \text{ and } F_i > 0 \\ \dfrac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} & \text{if } i \neq j \text{ and } F_i > 0 \\ 0 & \text{if } F_i = 0 \end{cases}$$

Proof. The proof can be found in the Appendix. □

We can calculate the conditional entropy again with the new $E(Z_{ij})$ values. The formula is given below. The proof can be found in the Appendix.

$$H(X|Y) = -\sum_{i=1}^{K} \frac{\min(F_i, G_i)}{N} \cdot \log \frac{\min(F_i, G_i)}{G_i} - \sum_{i=1}^{K} \frac{F_i - \min(F_i, G_i)}{N} \cdot \log \frac{F_i - \min(F_i, G_i)}{N - \sum_{k=1}^{K} \min(F_k, G_k)} - \sum_{j=1}^{K} \frac{G_j - \min(F_j, G_j)}{N} \cdot \log \frac{G_j - \min(F_j, G_j)}{G_j} \tag{3.5}$$

3.3.2 Uneven sums of frequencies

In Section 3.3.1 we assumed that $I = J$ and $N = M$. In numerous cases the sum of the original frequencies is not equal to the sum of the perturbed frequencies, that is, $N \neq M$. In this section we extend our results to this situation.

We define first a new set of individuals, denoted by I'. This set will consist of $N \cdot M$ (imaginary) individuals. If $a \in I$, then I' will contain M 'copies' of a. We define the J' set similarly. If $b \in J$, the J' set of (imaginary) individuals will contain N individuals identical to b. It implies that the new frequency tables are $M \cdot F = (M \cdot F_1, M \cdot F_2, \ldots, M \cdot F_K)$ and $N \cdot G = (N \cdot G_1, N \cdot G_2, \ldots, N \cdot G_K)$. Note that $\sum_{i=1}^{K} M \cdot F_i = \sum_{j=1}^{K} N \cdot G_j = N \cdot M$ and the entropies of the new frequency tables are equal to those of the initial tables. It means that we can assume that $I' = J'$ and calculate the conditional entropy as it is described in Section 3.3.1. Although the X, Y_l and Z_{ij} variables are different from those used in Section 3.3.1, we do not change the notation.

Theorems 1 and 2 can be rewritten as follows.

Theorem 3. *If $R_{N\cdot G} = U_{\Omega_{N\cdot G}}$, then*

$$E(Z_{ij}) = \sum_{l=1}^{|\Omega_{N\cdot G}|} R_{N\cdot G}(Y_l) \cdot Pr(Y_l = c_j | X = c_i) = \begin{cases} \frac{G_j}{M} & \text{if } F_i > 0 \\ 0 & \text{if } F_i = 0 \end{cases}$$

Proof. The proof is the same as that of Theorem 1. □

The conditional entropy takes the following form.

$$H(X|Y) = -\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{M\cdot F_i}{N\cdot M} \cdot \frac{N\cdot G_j}{N\cdot M} \cdot \log \frac{\frac{M\cdot F_i}{N\cdot M}\cdot\frac{N\cdot G_j}{N\cdot M}}{\sum_{k=1}^{K}\frac{M\cdot F_k}{N\cdot M}\cdot\frac{N\cdot G_j}{N\cdot M}} = -\sum_{j=1}^{K}\frac{G_j}{M}\cdot\sum_{i=1}^{K}\frac{F_i}{N}\cdot\log\frac{F_i}{N} = H(X)$$

It yields again that the conditional entropy (calculated on this choice of $R_{N\cdot G}$) does not lower the disclosure risk.

Theorem 4. *Assume that $F \neq G$ and*

$$R_{N\cdot G}(Y_l) = \begin{cases} 1/|\Omega^*_{N\cdot G}| & \text{if } Y_l \in \Omega^*_{N\cdot G} \\ 0 & \text{if } Y_l \notin \Omega^*_{N\cdot G} \end{cases}$$

Then

$$E(Z_{ij}) = \sum_{Y_l \in \Omega_{N\cdot G^*}} R_{N\cdot G}(Y_l)\cdot Pr(Y_l = c_j | X = c_i) =$$

$$\begin{cases} \dfrac{\min(M\cdot F_i, N\cdot G_i)}{M\cdot F_i} & \text{if } i = j \text{ and } F_i > 0 \\ \dfrac{(M\cdot F_i - \min(M\cdot F_i, N\cdot G_i))\cdot(N\cdot G_j - \min(M\cdot F_j, N\cdot G_j))}{M\cdot F_i\cdot(N\cdot M - \sum_{k=1}^{K}\min(M\cdot F_k, N\cdot G_k))} & \text{if } i \neq j \text{ and } F_i > 0 \\ 0 & \text{if } F_i = 0 \end{cases}$$

Proof. The proof is the same as that of Theorem 2. □

We can calculate the conditional entropy on $E(Z_{ij})$ again. The proof of the folowing formula can be found in the Appendix.

$$H(X|Y) = -\sum_{i=1}^{K} \frac{\min(M \cdot F_i, N \cdot G_i)}{N \cdot M} \cdot \log \frac{\min(M \cdot F_i, N \cdot G_i)}{N \cdot G_i} -$$
$$\sum_{j=1}^{K} \frac{N \cdot G_j - \min(M \cdot F_j, N \cdot G_j)}{N \cdot M} \cdot$$
$$\sum_{i=1}^{K} \frac{M \cdot F_i - \min(M \cdot F_i, N \cdot G_i)}{N \cdot M - \sum_{k=1}^{K} \min(M \cdot F_k, N \cdot G_k)} \cdot \log \frac{M \cdot F_i - \min(M \cdot F_i, N \cdot G_i)}{N \cdot M - \sum_{k=1}^{K} \min(M \cdot F_k, N \cdot G_k)} -$$
$$\sum_{i=1}^{K} \frac{M \cdot F_i - \min(M \cdot F_i, N \cdot G_i)}{N \cdot M - \sum_{k=1}^{K} \min(M \cdot F_k, N \cdot G_k)} \cdot$$
$$\sum_{j=1}^{K} \frac{N \cdot G_j - \min(M \cdot F_j, N \cdot G_j)}{N \cdot M} \cdot \log \frac{N \cdot G_j - \min(M \cdot F_j, N \cdot G_j)}{N \cdot G_j} \tag{3.6}$$

This formula is the generalisation of (3.5).

3.3.3 The perturbation method

So far we have dealt with a fixed G perturbed frequency vector without taking the perturbation method into account. A post-tabular perturbation method assigns probabilities to potential perturbed tables given the F original table. In order to maintain the generality, at this point we do not select a perturbation method, therefore the set of potential perturbed tables consists of every integer vector of length K. We denote the set of potential perturbed tables by PG.

$$PG = \{G : G = (G_1, G_2, \ldots, G_K) \in \mathbb{Z}^K\} = \mathbb{Z}^K$$

However, we assume that the perturbation method assigns non-zero probability to finite number of perturbed vectors. The probability assigned to the G table will be denoted $T(G)$. It implies that $T(\cdot)$ provides a probability distribution on PG.

We have defined the Ω_G set for an arbitrarily chosen G perturbed table. Denote

the union of these sets by Ω.

$$\Omega = \bigcup_{G \in PG} \Omega_G$$

Assume that we have defined an $R_G : \Omega_G \to \mathbb{R}$ probability distribution for all G. It provides an $R : \Omega \to \mathbb{R}$ probability distribution as follows. If $Y \in \Omega$, then Y is an element of an Ω_G. By definition,

$$R(Y) = T(G) \cdot R_G(Y) \quad \text{if } Y \in \Omega_G$$

R is a probability distribution on Ω, since

$$\sum_{Y \in \Omega} R(Y) = \sum_{G \in PG} \sum_{Y \in \Omega_G} T(G) \cdot R_G(Y) = \sum_{G \in PG} T(G) \cdot \sum_{Y \in \Omega_G} R_G(Y) = \sum_{G \in PG} T(G) = 1$$

The Ω^* set can be defined similarly.

$$\Omega^* = \bigcup_{G \in PG} \Omega^*_G$$

At this point it is inevitable to replace the M and Z_{ij} notations with M_G and Z^G_{ij}. Theorems 3 and 4 can be extended as follows.

Theorem 5. *If $R_{N \cdot G} = U_{\Omega_{N \cdot G}}$ for all G, then*

$$\sum_{G \in PG} T(G) \cdot E(Z^G_{ij}) = \begin{cases} \sum_{G \in PG} T(G) \cdot \dfrac{G_j}{M_G} & \text{if } F_i > 0 \\ 0 & \text{if } F_i = 0 \end{cases}$$

Proof. This Theorem is the straightforward consequence of Theorem 3. □

Theorem 6. *Assume that*

$$R_{N \cdot G}(Y_l) = \begin{cases} 1/|\Omega^*_{N \cdot G}| & \text{if } Y_l \in \Omega^*_{N \cdot G} \\ 0 & \text{if } Y_l \notin \Omega^*_{N \cdot G} \end{cases}$$

for all $G \in PG$. Then

$$\sum_{G \in PG} T(G) \cdot E(Z_{ij}^G) =$$

$$\begin{cases} T(F) + \sum_{G \in PG \setminus \{F\}} T(G) \cdot \dfrac{\min(M_G \cdot F_i, N \cdot G_i)}{M_G \cdot F_i} & \text{if } i = j \text{ and } F_i > 0 \\ \sum_{G \in PG \setminus \{F\}} T(G) \cdot \dfrac{(M_G \cdot F_i - \min(M_G \cdot F_i, N \cdot G_i)) \cdot (N \cdot G_j - \min(M_G \cdot F_j, N \cdot G_j))}{M_G \cdot F_i \cdot (N \cdot M_G - \sum_{k=1}^{K} \min(M_G \cdot F_k, N \cdot G_k))} & \text{if } i \neq j \text{ and } F_i > 0 \\ 0 & \text{if } F_i = 0 \end{cases}$$

Proof. It can be seen easily that $E(Z_{ii}^F) = 1$ and $E(Z_{ij}^F) = 0$ if $i \neq j$. Otherwise the proof can be derived from Theorem 4. □

4 Application

We apply the disclosure risk measure to tables generated from the 2001 UK census data. The table-spanning variables for various tables include age, sex, output area, country of birth, mode of travel, religion. In this paper only two-dimensional tables are considered.

We investigate the output area × country of birth, output area × mode of travel, output area × sex and output area × religion tables, where only 10 output areas are taken into account. The population size is $N = 2449$. In case of the output area × mode of travel table the population is restricted to individuals between 16 and 74 years of age.

The perturbation method we apply to the frequency tables is random rounding to base 3. Random rounding moves the frequencies to one of the multiples of 3 with certain probability structure. If a cell value is a multiple of 3, it remains unaltered. If the remainder is 1 or 2 when dividing the cell value by 3, then we round it to the closest or second closest multiple of 3 with probability 2/3 or 1/3 respectively. Different cells in the table, including marginal cells, are rounded independently. Random rounding may not result in additive tables, that is, the internal cells may not add up to the marginal total. In this work we deal with internal cells only.

The entropy-based term is the core of the diclosure risk measure, therefore we assign high weight to that term in $R_1(F, \mathbf{w})$. We use $\mathbf{w} = (w_1, w_2, w_3) = (0.1, 0.8, 0.1)$.

In Section 3.3.3 we discussed how the perturbation method may be taken into consideration. In the numerical study we do not calculate the $T(G)$ probabilities. However, we carry out the perturbation 1,000 times and the final disclosure risk measure is the average of the disclosure risk measures of the iterations. The perturbation can result in the same frequency table for distinct iterations. Therefore the $T(G)$ probability of a specific G perturbed table is approximately reflected by the number of iterations resulting in G.

We evaluate the information loss random rounding causes by the Hellinger distance. The distance is calculated simultaneously with the disclosure risk measures for the 1,000 iterations.

The results can be seen in Tables 1, 2, 3 and 4 in the Appendix.

Appendix

Proof of Theorem 1. If $F_i = 0$, then $Pr(Y_l = c_j | X = c_i) = 0$, therefore $E(Z_{ij}) = 0$.

Assume now that $F_i > 0$.

$$E(Z_{ij}) = \sum_{l=1}^{|\Omega_G|} R_G(Y_l) \cdot Pr(Y_l = c_j | X = c_i) = \frac{1}{|\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} Pr(Y_l = c_j | X = c_i) =$$

$$\frac{1}{|\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} \frac{Pr(X = c_i, Y_l = c_j)}{Pr(X = c_i)} = \frac{N}{F_i \cdot |\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} Pr(X = c_i, Y_l = c_j) =$$

$$\frac{N}{F_i \cdot |\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} \frac{|\{a \in I : X(a) = c_i, Y_l(a) = c_j\}|}{N} =$$

$$\frac{1}{F_i \cdot |\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} |\{a \in I : X(a) = c_i, Y_l(a) = c_j\}|$$

Note that

$$\sum_{l=1}^{|\Omega_G|}|\{a \in I : X(a) = c_i, Y_l(a) = c_j\}|= |\{(a, Y_l) \in I \times \Omega_G : X(a) = c_i, Y_l(a) = c_j\}|$$

It can be seen easily that

$$|\Omega_G|= \frac{N!}{G_1! \cdot G_2! \cdot \ldots \cdot G_K!} \tag{4.1}$$

Since $F_i > 0$, we can choose an $a_0 \in I$ individual such that $X(a_0) = c_i$. The number of $Y \in \Omega_G$ variables with $Y(a_0) = c_j$ is

$$|\{Y \in \Omega_G : Y(a_0) = c_j\}|= \frac{(N-1)!}{G_1! \cdot \ldots \cdot G_{j-1}! \cdot (G_j - 1)! \cdot G_{j+1}! \cdot \ldots \cdot G_K!} = \frac{G_j}{N} \cdot |\Omega_G|$$

There are F_i choices to select the a_0 individual, therefore

$$|\{(a, Y_l) \in I \times \Omega_G : X(a) = c_i, Y_l(a) = c_j\}|= F_i \cdot \frac{G_j}{N} \cdot |\Omega_G|$$

This equation completes the theorem. □

Proof of Theorem 2. Similarly to the proof of Theorem 1, we obtain again that

$$E(Z_{ij}) = \frac{1}{F_i \cdot |\Omega_G^*|} \cdot \sum_{Y_l \in \Omega_G^*} |\{a \in I : X(a) = c_i, Y_l(a) = c_j\}|= \frac{1}{F_i \cdot |\Omega_G^*|} \cdot |\{(a, Y_l) \in I \times \Omega_G^* : X(a) = c_i, Y_l(a) = c_j\}| \tag{4.2}$$

Our next aim is to determine $|\Omega_G^*|$. First we need to choose $\sum_{k=1}^{K} \min(F_k, G_k)$ individuals that remain in the same cell. The number of choices is

$$\binom{F_1}{\min(F_1, G_1)} \cdot \binom{F_2}{\min(F_2, G_2)} \cdot \ldots \cdot \binom{F_K}{\min(F_K, G_K)}$$

Assume that we have selected $\sum_{k=1}^{K} \min(F_k, G_k)$ individuals that remain in their respective category after perturbation. However, the categories where the not selected individuals fall after perturbation are not given. Similarly to (4.1), there are

$$\frac{(N - \sum_{k=1}^{K} \min(F_k, G_k))!}{\prod_{j=1}^{K}(G_j - \min(F_j, G_j))!}$$

possible variables on the not selected individuals. Therefore the cardinality of Ω_G^* is

$$|\Omega_G^*|= \frac{(N-\sum_{k=1}^K \min(F_k,G_k))!}{\prod_{j=1}^K (G_j-\min(F_j,G_j))!}\cdot \prod_{i=1}^K \binom{F_i}{\min(F_i,G_i)} \tag{4.3}$$

If $F_i=0$, then $E(Z_{ij})=0$. Assume that $F_i>0$ and select an $a_0\in I$ individual with $X(a_0)=c_i$.

If $i=j$, that is, $Y_l(a_0)=c_i$, then

$$|\{Y_l\in\Omega_G^* : Y_l(a_0)=c_i\}|= \frac{\left(N-1-\min(F_i-1,G_i-1)-\sum_{k\neq i}\min(F_k,G_k))\right)!}{(G_i-1-\min(F_i-1,G_i-1))!\cdot\left(\prod_{k\neq i}G_k-\min(F_k,G_k)\right)!}\cdot$$
$$\binom{F_i-1}{\min(F_i-1,G_i-1)}\cdot\prod_{k\neq i}\binom{F_k}{\min(F_k,G_k)}=$$
$$\frac{\left(N-\sum_{k=1}^K\min(F_k,G_k)\right)!}{\prod_{k=1}^K(G_k-\min(F_k,G_k))!}\cdot\binom{F_i-1}{\min(F_i,G_i)-1}\cdot\prod_{k\neq i}\binom{F_k}{\min(F_k,G_k)}$$

There are F_i choices to fix the a_0 individual. Therefore

$$|\{(a,Y_l)\in I\times\Omega_G^* : X(a)=c_i, Y_l(a)=c_i\}|= F_i\cdot|\{Y_l\in\Omega_G^* : Y_l(a)=c_i\}|=$$
$$F_i\cdot\frac{\left(N-\sum_{k=1}^K\min(F_k,G_k)\right)!}{\prod_{k=1}^K(G_k-\min(F_k,G_k))!}\cdot\binom{F_i-1}{\min(F_i,G_i)-1}\cdot\prod_{k\neq i}\binom{F_k}{\min(F_k,G_k)}=$$
$$\min(F_i,G_i)\cdot\frac{\left(N-\sum_{k=1}^K\min(F_k,G_k)\right)!}{\prod_{k=1}^K(G_k-\min(F_k,G_k))!}\cdot\prod_{k=1}^K\binom{F_k}{\min(F_k,G_k)} \tag{4.4}$$

Combine this result with (4.2) and (4.3). We get that

$$E(Z_{ii})=\frac{\min(F_i,G_i)}{F_i}$$

Assume now that $i\neq j$. Select an a_0 individuals with $X(a_0)=c_i$ and $Y_l(a_0)=c_j$. We will show that $F_i>G_i$ and $F_j<G_j$ must satisfy. Indirectly, if $F_i\leq G_i$, then $\min(F_i,G_i)=F_i$, and therefore $X(a_0)=c_i$ implies that $Y_l(a_0)=c_i$ since $Y_l\in\Omega_G^*$.

On the other hand, if $F_j \geq G_j$, then $Y_l(a_0) = c_j$ implies that $X(a_0) = c_j$. These are contradictions, therefore $F_i > G_i$ and $F_j < G_j$.

We need to determine the $|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_j\}|$ frequency.

$$|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_j\}| =$$
$$\frac{\left(N - 1 - \min(F_i - 1, G_i) - \min(F_j, G_j - 1) - \sum_{k\neq i, k\neq j} \min(F_k, G_k)\right)!}{(G_i - \min(F_i - 1, G_i))! \cdot (G_j - 1 - \min(F_j, G_j - 1))! \cdot \prod_{k\neq i, k\neq j} (G_k - \min(F_k, G_k))!} \cdot$$
$$\binom{F_i - 1}{\min(F_i - 1, G_i)} \cdot \binom{F_j}{\min(F_j, G_j - 1)} \cdot \prod_{k\neq i, k\neq j} \binom{F_k}{\min(F_k, G_k)} \tag{4.5}$$

$F_i > G_i$ and $G_j > F_j$ yield that $\min(F_i - 1, G_i) = \min(F_i, G_i) = G_i$ and $\min(F_j, G_j - 1) = \min(F_j, G_j) = F_j$. Therefore (4.5) can be written as

$$|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_j\}| =$$
$$\frac{\left(N - 1 - \sum_{k=1}^{K} \min(F_k, G_k)\right)!}{(G_j - 1 - \min(F_j, G_j))! \cdot \prod_{k\neq j} (G_k - \min(F_k, G_k))!} \cdot$$
$$\binom{F_i - 1}{\min(F_i, G_i)} \cdot \binom{F_j}{\min(F_j, G_j)} \cdot \prod_{k\neq i, k\neq j} \binom{F_k}{\min(F_k, G_k)} =$$
$$\frac{G_j - \min(F_j, G_j)}{N - \sum_{k=1}^{K} \min(F_k, G_k)} \cdot \frac{\left(N - \sum_{k=1}^{K} \min(F_k, G_k)\right)!}{\prod_{k=1}^{K} (G_k - \min(F_k, G_k))!} \cdot$$
$$\frac{F_i - \min(F_i, G_i)}{F_i} \cdot \prod_{k=1}^{K} \binom{F_k}{\min(F_k, G_k)}$$

We have F_i choices to select the a_0 individual, therefore

$$|\{(a, Y_l) \in I \times \Omega_G^* : X(a) = c_i, Y_l(a) = c_j\}| = F_i \cdot |\{Y_l \in \Omega_G^* : Y_l(a) = c_i\}| =$$
$$\frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{(N - \sum_{k=1}^{K} \min(F_k, G_k))} \cdot \frac{\left(N - \sum_{k=1}^{K} \min(F_k, G_k)\right)!}{\prod_{k=1}^{K} (G_k - \min(F_k, G_k))!} \cdot \prod_{k=1}^{K} \binom{F_k}{\min(F_k, G_k)} \tag{4.6}$$

Combine this result with (4.2) and (4.3). We obtain that

$$E(Z_{ij}) = \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))}$$

This proves the theorem. □

Proof of (3.5).

$$H(X|Y) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \frac{\min(F_i, G_i)}{F_i} \cdot \log \frac{\frac{F_i}{N} \cdot \frac{\min(F_i,G_i)}{F_i}}{\frac{F_i}{N} \cdot \frac{\min(F_i,G_i)}{F_i} + \sum_{k\neq i} \frac{F_k}{N} \cdot \frac{(F_k - \min(F_k,G_k)) \cdot (G_i - \min(F_i,G_i))}{F_k \cdot \left(N - \sum_{m=1}^{K} \min(F_m,G_m)\right)}} -$$

$$\sum_{i=1}^{K} \sum_{j\neq i} \frac{F_i}{N} \cdot \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k))} \cdot$$

$$\log \frac{\frac{F_i}{N} \cdot \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k))}}{\frac{F_j}{N} \cdot \frac{\min(F_j,G_j)}{F_j} + \sum_{k\neq j} \frac{F_k}{N} \cdot \frac{(F_k - \min(F_k,G_k)) \cdot (G_j - \min(F_j,G_j))}{F_k \cdot (N - \sum_{m=1}^{K} \min(F_m,G_m))}} =$$

$$-\sum_{i=1}^{K} \frac{\min(F_i,G_i)}{N} \cdot \log \frac{\min(F_i,G_i)}{\min(F_i,G_i) + \frac{\sum_{k=1}^{K} (F_k - \min(F_k,G_k)) \cdot (G_i - \min(F_i,G_i))}{\left(N - \sum_{m=1}^{K} \min(F_m,G_m)\right)}} -$$

$$\sum_{i=1}^{K} \sum_{j\neq i} \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{N \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k))} \cdot$$

$$\log \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{\min(F_j,G_j) \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k)) + \sum_{k=1}^{K} (F_k - \min(F_k,G_k)) \cdot (G_j - \min(F_j,G_j))} =$$

$$-\sum_{i=1}^{K} \frac{\min(F_i,G_i)}{N} \cdot \log \frac{\min(F_i,G_i)}{G_i} -$$

$$\sum_{i=1}^{K} \sum_{j\neq i} \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{N \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k))} \cdot \log \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{G_j \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k))} =$$

$$-\sum_{i=1}^{K} \frac{\min(F_i,G_i)}{N} \cdot \log \frac{\min(F_i,G_i)}{G_i} -$$

$$\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{N \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k))} \cdot \log \frac{(F_i - \min(F_i,G_i)) \cdot (G_j - \min(F_j,G_j))}{G_j \cdot (N - \sum_{k=1}^{K} \min(F_k,G_k))} =$$

$$
\begin{aligned}
&-\sum_{i=1}^{K}\frac{\min(F_i,G_i)}{N}\cdot\log\frac{\min(F_i,G_i)}{G_i}-\\
&\sum_{j=1}^{K}\frac{G_j-\min(F_j,G_j)}{N}\cdot\sum_{i=1}^{K}\frac{F_i-\min(F_i,G_i)}{N-\sum_{k=1}^{K}\min(F_k,G_k)}\cdot\log\frac{F_i-\min(F_i,G_i)}{N-\sum_{k=1}^{K}\min(F_k,G_k)}-\\
&\sum_{i=1}^{K}\frac{F_i-\min(F_i,G_i)}{N-\sum_{k=1}^{K}\min(F_k,G_k)}\cdot\sum_{j=1}^{K}\frac{G_j-\min(F_j,G_j)}{N}\cdot\log\frac{G_j-\min(F_j,G_j)}{G_j}=\\
&-\sum_{i=1}^{K}\frac{\min(F_i,G_i)}{N}\cdot\log\frac{\min(F_i,G_i)}{G_i}-\\
&\sum_{i=1}^{K}\frac{F_i-\min(F_i,G_i)}{N}\cdot\log\frac{F_i-\min(F_i,G_i)}{N-\sum_{k=1}^{K}\min(F_k,G_k)}-\\
&\sum_{j=1}^{K}\frac{G_j-\min(F_j,G_j)}{N}\cdot\log\frac{G_j-\min(F_j,G_j)}{G_j}
\end{aligned}
\tag{4.7}
$$

□

Proof of (3.6).

$$
\begin{aligned}
&H(X|Y)=-\sum_{i=1}^{K}\frac{\min(M\cdot F_i,N\cdot G_i)}{N\cdot M}\cdot\log\frac{\min(M\cdot F_i,N\cdot G_i)}{N\cdot G_i}-\\
&\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(M\cdot F_i-\min(M\cdot F_i,N\cdot G_i))\cdot(N\cdot G_j-\min(M\cdot F_j,N\cdot G_j))}{N\cdot M\cdot(N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k,N\cdot G_k))}\cdot\\
&\log\frac{(M\cdot F_i-\min(M\cdot F_i,N\cdot G_i))\cdot(N\cdot G_j-\min(M\cdot F_j,N\cdot G_j))}{N\cdot G_j\cdot(N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k,N\cdot G_k))}=
\end{aligned}
$$

$$
\begin{aligned}
&-\sum_{i=1}^{K}\frac{\min(M\cdot F_i, N\cdot G_i)}{N\cdot M}\cdot\log\frac{\min(M\cdot F_i, N\cdot G_i)}{N\cdot G_i}- \\
&\sum_{j=1}^{K}\frac{N\cdot G_j-\min(M\cdot F_j, N\cdot G_j)}{N\cdot M}\cdot \\
&\sum_{i=1}^{K}\frac{M\cdot F_i-\min(M\cdot F_i, N\cdot G_i)}{N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k, N\cdot G_k)}\cdot\log\frac{M\cdot F_i-\min(M\cdot F_i, N\cdot G_i)}{N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k, N\cdot G_k)}- \\
&\sum_{i=1}^{K}\frac{M\cdot F_i-\min(M\cdot F_i, N\cdot G_i)}{N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k, N\cdot G_k)}\cdot \\
&\sum_{j=1}^{K}\frac{N\cdot G_j-\min(M\cdot F_j, N\cdot G_j)}{N\cdot M}\cdot\log\frac{N\cdot G_j-\min(M\cdot F_j, N\cdot G_j)}{N\cdot G_j}
\end{aligned}
$$

□

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	$R_1(F, \mathbf{w})$	$R_2(F, G, \mathbf{w})$	$HD(F, G)$
1	230	5	7	3	0	0	0	1	19	0	1	13	19	4	0	0	0	0.5847	0.1383	1.0169
2	154	1	2	8	0	5	0	1	5	0	3	1	11	1	4	3	0	0.5584	0.1720	1.5053
3	188	9	6	2	0	5	0	2	1	0	2	10	18	6	0	3	6	0.5125	0.0940	1.2258
4	278	3	1	1	0	1	0	1	8	0	0	3	8	8	0	0	0	0.7085	0.2142	1.3384
5	173	4	8	0	0	7	0	2	4	1	1	6	14	5	0	4	0	0.5472	0.1570	1.2188
6	161	10	1	2	0	1	1	1	9	1	1	7	13	0	7	3	1	0.5092	0.1497	1.8428
7	151	8	1	4	0	1	0	0	7	1	0	8	27	5	0	0	1	0.5422	0.1495	1.3765
8	208	7	1	3	1	9	1	0	2	6	1	5	29	3	1	2	1	0.5178	0.1249	1.8226
9	174	12	0	4	0	3	0	1	5	1	0	3	3	3	0	2	2	0.6095	0.1478	1.2813
10	171	5	0	4	0	4	0	1	8	0	0	6	10	8	2	3	1	0.5558	0.1538	1.1956
$R_1(F, \mathbf{w})$	0.0175	0.0874	0.2714	0.1435	0.9900	0.1744	0.7344	0.1844	0.1158	0.4915	0.3273	0.0930	0.0756	0.1309	0.5151	0.2185	0.3928	0.3242	-	-
$R_2(F, G, \mathbf{w})$	0.0110	0.0463	0.1295	0.0742	0.1893	0.0791	0.4139	0.1347	0.0538	0.3911	0.2192	0.0463	0.0322	0.0619	0.2747	0.1065	0.2733	-	0.0459	-
$HD(F, G)$	0.1031	0.8224	1.4450	1.1257	0.6416	1.2123	0.9152	1.7978	1.0099	1.2946	1.4278	0.7659	0.3677	0.8118	0.9048	0.7980	1.4238	-	-	4.4800

Table 1: Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ country of birth. The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise.

	1	2	3	4	5	6	7	8	9	10	11	$R_1(F, \mathbf{w})$	$R_2(F, G, \mathbf{w})$	$HD(F, G)$
1	8	11	7	1	44	5	3	3	47	0	126	0.3291	0.0533	0.7576
2	5	4	2	0	50	2	0	7	15	1	70	0.3670	0.0829	1.1139
3	3	1	8	0	18	5	0	8	26	0	122	0.4417	0.0916	0.7442
4	7	1	10	0	18	4	0	5	24	0	135	0.4536	0.0922	0.7665
5	3	2	4	0	17	6	0	2	26	0	107	0.4563	0.1000	0.8289
6	8	2	7	1	97	9	0	8	28	1	54	0.3157	0.0573	1.1356
7	5	0	8	0	29	2	1	3	14	0	88	0.4252	0.1019	0.9364
8	14	1	22	1	30	4	0	4	30	0	93	0.3214	0.0615	0.9884
9	10	0	8	1	23	2	0	0	17	2	78	0.3946	0.0996	1.0818
10	17	9	5	1	96	3	1	2	17	4	52	0.3003	0.0579	1.1450
$R_1(F, \mathbf{w})$	0.0850	0.2862	0.0944	0.3715	0.0927	0.0847	0.6206	0.1335	0.0474	0.5107	0.0309	0.2016	-	-
$R_2(F, G, \mathbf{w})$	0.0425	0.1363	0.0464	0.2868	0.0219	0.0536	0.4924	0.0655	0.0252	0.3283	0.0147	-	0.0295	-
$HD(F, G)$	0.5190	1.4289	0.7948	1.4492	0.2165	1.1111	0.9151	0.9035	0.2844	1.1103	0.1221	-	-	3.1133

Table 2: Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ mode of travel (age: 16-74).
The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise.

	1	2	$R_1(F, \mathbf{w})$	$R_2(F, G, \mathbf{w})$	$HD(F, G)$
1	161	141	0.0247	0.0222	0.0376
2	105	94	0.0276	0.0259	0.0486
3	142	116	0.0294	0.0237	0.0612
4	158	154	0.0220	0.0219	0.0539
5	139	90	0.0512	0.0252	0.0398
6	129	90	0.0434	0.0250	0.0000
7	107	107	0.0252	0.0252	0.0660
8	133	147	0.0243	0.0228	0.0402
9	98	115	0.0289	0.0253	0.0666
10	136	87	0.0529	0.0254	0.0409
$R_1(F, \mathbf{w})$	0.0170	0.0209	0.0150	-	-
$R_2(F, G, \mathbf{w})$	0.0127	0.0135	-	0.0100	-
$HD(F, G)$	0.1227	0.1032	-	-	0.1611

Table 3: Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ sex. The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise.

	1	2	3	4	5	6	7	8	9	Ran. Rou. $R_1(F, \mathbf{w})$	Ran. Rou. $R_2(F, G, \mathbf{w})$	Ran. Rou. $HD(F, G)$
1	181	0	0	1	17	1	1	83	18	0.4626	0.0634	1.1356
2	138	2	4	2	0	0	1	36	16	0.4973	0.1028	1.0752
3	130	0	0	0	22	4	1	61	40	0.3939	0.0742	0.7054
4	173	0	0	1	14	4	1	97	22	0.4403	0.0669	0.9668
5	142	2	5	0	15	6	1	37	21	0.3869	0.0568	0.8948
6	129	0	0	0	0	0	1	69	20	0.5460	0.1011	0.6535
7	118	2	0	2	24	9	1	38	20	0.3456	0.0562	1.0288
8	130	0	0	0	34	1	1	82	32	0.3974	0.0673	0.9218
9	148	3	0	0	0	2	1	38	21	0.5243	0.0894	0.8468
10	136	1	2	0	13	0	0	55	16	0.4692	0.0917	0.8783
$R_1(F, \mathbf{w})$	0.0152	0.3770	0.5763	0.4754	0.2029	0.2892	0.1166	0.0393	0.0404	0.2315	-	-
$R_2(F, G, \mathbf{w})$	0.0123	0.2235	0.2944	0.3282	0.0690	0.1304	0.0986	0.0178	0.0255	-	0.0327	-
$HD(F, G)$	0.1176	1.1877	0.6261	1.2280	0.2664	1.1223	1.9488	0.1920	0.2832	-	-	2.9751

Table 4: Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ religion. The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise.

References

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, 2nd edition.

Csiszár, I. (1967). Information-Type Measures of Difference of Probability Distributions and Indirect Observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Csiszár, I. and Shields, P. (2004). *Information Theory And Statistics: A Tutorial*, volume 1 of *Foundations and Trends in Communications and Information Theory*. Now Publishers Incorporated.

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review*, 75(2):199–217.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Lecture notes in statistics. Springer.