Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada's Symposium 2005
Methodological Challenges for Future Information Needs

DATA SWAPPING IS NOT THE PANACEA

Jean-René Boudreau[1]

ABSTRACT

Data swapping introduces noise in a dataset to improve the protection of statistical confidentiality. The noise is generated by permuting selected characteristics (variables) among a small number of units (records). We demonstrate in this article that this technique introduces a bias in the estimates. We provide explicit formulas for the mean square error. We will see that the error is proportional to the size of the cell in a tabulation. Then we compare that technique to the random rounding used in the Canadian Census of Population. We determine the permutation rate that produces the same amount of noise. Having obtained that rate, we discuss the effectiveness of data swapping.

KEYWORDS: Data swapping; confidentiality, disclosure risk; protection of statistical confidentiality.

1. INTRODUCTION

Statistics agencies typically have two conflicting mandates. On one hand, their mission is to publish statistical information about the activities and condition of the country's people. On the other, they set stringent rules for themselves to preserve the confidentiality of the responses provided by respondents. They have to find methods of ensuring that respondents' information will never be disclosed without their permission to anyone in any form that would make it possible to identify them. If respondents know that they will not be identified with the information they provide, they will be more likely to provide accurate information.

What are those methods? First, there are physical security methods: providing secure storage locations for paper questionnaires, having intrusion-proof communications systems, restricting access to the information to employees who have taken the oath of office, and so on. There are also measures that can be applied to information intended for release. Such measures ensure that the reporting unit cannot be identified from the published information. Agencies typically restrict access to data when statistical confidentiality is not guaranteed. They also protect the information by modifying it, though they must simultaneously preserve its statistical value. Two examples of restricting access to data are (1) requiring users to comply with certain conditions, such as taking an oath or signing an agreement on how the released data will used, and (2) suppressing a variable in a microdata file. Random rounding of tabulations to a multiple of a whole larger than the unit is an example of data modification.

Clearly, the trend in dissemination programs is to issue products containing more detailed data. It is also true that users are increasingly sophisticated, using more precise methods of analysis than in the past. Statistical software automatically compensates for cluster effects, thereby reducing certain survey design biases in statistical analyses and tests. The value added is actually in the accuracy of the published information. From the standpoint of statistical confidentiality protection, the methodological challenge for future information needs is to find ways of distorting the data and calibrating the distortion to preserve the accuracy and reliability of the information. First, respondents must be reassured that their information will always be treated as confidential so that they will provide accurate data, and second, ways must be found of protecting the data without affecting their use. Consequently, the protection techniques must be visible and known, and the distortion of the data must not generate additional phenomena; in other words, the noise must be as "white" as possible.

There is a wide variety of methods for introducing noise in data. For example, sampling can be used to introduce variation in the estimates. This method is also used to restrict access to the data: the designer of a microdata file can resample to reduce the chances of identification. Two other ways of introducing noise are rounding aggregate data in

[1] Jean-René Boudreau, Statistics Canada, R. H. Coats Building, 15th Floor, Tunney's Pasture, Ottawa, K1A 0T6, jboudre@statcan.ca.

tables (tabulations) and creating dummy units in the survey database. One of them involves swapping the values of selected variables between records; this is known as "data swapping."

In this article, we will study the noise generated by data swapping. Is it white? Does it reduce the chances of identification? The title of the article casts doubt on the method's effectiveness. The method was tested because it is easy to analyze. The value of this presentation lies in the fact that it provides an example of how a protection technique can be analyzed. Identifying a protection technique's strengths and weaknesses can only benefit everyone concerned.

2. DEFINITION OF DATA SWAPPING

This section introduces the basic concepts needed for an informed discussion of data swapping. It also introduces the various notations that will be used throughout the presentation.

Suppose we have a database consisting of a number of characteristics (variables) measured for respondents (units or records). Let m and n represent the number of variables and the number of records respectively, and let the database be represented by the matrix $Y = (y_{ij})$, where $i = 1, ..., n$ and $j = 1, ..., m$. y_{ij} represents the measured value of variable j of record i. The database has a special variable for the weight of each record (the weight is usually the inverse selection probability). That variable is denoted $w = (w_i)$.

Given database Y, a swap is the specification of two objects $E = \{(A \mid B), \Sigma\}$ where:
(a) $(A \mid B)$ is a partition of the set $\{1, ..., m\}$ of the variables of Y into two parts A and B;
(b) Σ is a probability distribution of the set of permutations of the set $\{1, \ldots, n\}$.

Performing the data swap $\{(A \mid B), \Sigma\}$ on Y will create a new database $Y' = (y'_{ij})$ where, in this case, $y'_{ij} = y_{ij}$ if $j \in B$, $y'_{ij} = y_{\sigma i\, j}$ if $j \in A$. Permutation σ is a realization of probability distribution Σ. All the estimates and analyses will apply to this database.

The choice of partition $(A \mid B)$ of E is crucial to its effectiveness in protecting statistical confidentiality. The most discriminating variables must be separated by the permutation. In addition, for the file Y' to remain useful, it is desirable for the cardinality of A to be much smaller than the cardinality of B. We note that the error caused by E comes into play only when the estimator formula includes at least one variable of A and one variable of B. Consequently, if the relation between two variables is important to the objectives of the survey that generated the database, the two variables should not be separated by the permutation of the variables. Then it must be determined whether w should be in A or B. If the database contains certain stratification variables, all of them will have to be in A or B along with w. Otherwise, an intruder might be able to determine that a particular record was swapped, which must be avoided. In this article, we assume that $w \in A$.

As noted earlier, it is not necessary to swap the values of the variables of A for all records. It is sufficient to swap the values among a smaller subset of records; this reduces the error caused by the swap. We use a distribution Σ that permutes only the values of $k > 1$ records. The permutation rate is defined as the proportion of records in which the values have been permuted. We denote it as $\tau = k / n$. The distribution that governs the choice of permutations may be a function of the values of selected variables. It is desirable *a priori* to preserve certain characteristics (such as the person's gender). These conditions characterize a particular allocation of the records into classes. Within each class, we permute the records at a preset rate. To illustrate a swap, let's start with the data in Figure 1. The selected permutation replaces the values of the fourth record's variables with the values of the sixth record's variables, the values of the fifth record's variables with the values of the fourth record's variables, the values of the sixth record's variables with the values of the seventh record's variables, and the values of the seventh record's variables with the values of the fifth record's variables.

Figure 1: Illustration of a Data Swap

Original database				Data swap	Modified database			
Obs	***w***	**1_P**	**1_F**		**Obs**	***w***	**1_P**	**1_F**

1	5.800281	0	1		1	5.800281	0	1
2	9.760256	1	1	$1_P, w \in A \quad 1_F \in B$	2	9.760256	1	1
3	6.531695	1	0		3	6.531695	1	0
4	8.829931	0	0	$\sigma = 1\ 2\ 3\ 6\ 4\ 7\ 5$	6	8.347917	1	0
5	9.805243	0	1		4	8.829931	0	1
6	8.347917	1	1	$\tau = 4 / 7$	7	5.952525	1	1
7	5.952525	1	1		5	9.805243	0	1

The designer has the choice of permutation types. Three types come to mind.

(a) For a number k ($1 < k \leq n$), we choose k whole numbers from $\{1, \ldots, n\}$ to form a vector with k components. We apply a permutation of the indices that cause them all to vary.
(b) For an even number k ($1 < k < n$), we choose k whole numbers from $\{1, \ldots, n\}$ to form a vector with k components. We substitute the vector's components two by two.
(c) For an even number k ($1 < k < n$), we divide the set $\{1, \ldots, n\}$ into two parts, G_1 and G_2, whose cardinality is greater than $k/2$. We choose $k/2$ whole numbers from each part to form two vectors with $k/2$ components. We apply a general permutation to the indices of the first part's components, and we permute the components of both parts with the same indices two by two.

These permutations swap exactly k out of n records. The distribution of the permutations is specified by the experience of a random choice of one permutation from among those which preserve the classes. For a given class, they are uniform distributions whose probability mass is the inverse of the number of possible permutations. The last two types of swaps simply reduce the number of possibilities. In general, parts G_1 and G_2 of the third type represent contiguous geographies.

Let $E = \{(A \mid B), \Sigma\}$ be a data swap and let X be any estimate. We are interested in finding a measure of the error in using X' (the value found for X after the data swap) instead of X. To measure the error, we will compute the mean square error,

$$MSE_{\Sigma}\ X' = \sqrt{V_{\Sigma} X' + B_{\Sigma}^2 X'}\ ,$$

where $V_{\Sigma} X'$ and $B_{\Sigma} X'$ are the variance and the bias of distribution Σ applied to X. Returning to the example shown in Figure 1, if X is the estimate of the total number of units that have attributes P and F ($\mathbf{1}_P = 1$, $\mathbf{1}_F = 1$), we have $X = 24.06$. Table 1 shows the analysis of all X' estimates that we can obtain by letting σ run through all 315 possibilities.

Table 1: Analysis of the Sample Data Swap

X′	Frequency	(X′ - mean X′)^2	Frequency
12.48422	4	0.118036	30
14.30044	22	2.168725	99
14.87961	4	3.083192	30
15.71278	22	4.210004	30
16.29195	4	20.06919	22
18.10817	22	39.64072	4
20.83214	30	47.26917	22
22.24448	30	59.41984	4
24.0607	99	64.06968	48
24.63987	30	68.68426	22
30.59239	48	102.08716	4
Mean = 22,58804	315	Mean = 23.20468	315

By performing this swap, we generate a bias of 1.47 (24.06 – 22.59), with a variance of 23.2047. That yields a mean square error of 5.03. In what follows, we will consider only swaps of the first type. We can show that reducing the number of possibilities does not change the conclusions of this article.

3. ERROR INTRODUCED BY DATA SWAPPING

Let X be an estimate from a database in which a data swap has been performed. We assume that X is a weighted sum over domain D defined by discrete variables. In other words, X is a tabulation. We are examining only domains defined as an intersection of sets of records that are defined on the basis of a single variable (e.g., married men). If the variables (including w) that define D are all in either A or B, there is no swap error. Otherwise, domain D is written as the intersection of two specific domains P and F ($D = P \cap F$). P and F are the domains specified by D, but only with variables of A and B respectively. P and F are the out-of-phase and fixed domains of D. Hence, if

$$X = X_D = \sum_i w_i 1_D(i) = \sum w_i 1_P(i)\, 1_F(i),$$

the expression for X' will be

$$X' = \sum_i w_{\sigma i} 1_P(\sigma i)\, 1_F(i) = \sum_{i \in P} w_i 1_F(\sigma^{-1} i),$$

where the last term is simply a rearrangement of the order of the sum. We want to show in this section that the square of the mean square error of a data swap is proportional to n. Let's start with the number of possibilities. The general binomial coefficient will be denoted C_k^n.

Result 1. *The number of permutations of k (k > 1) objects that leave no object fixed is given by*

$$k!\, e_k = k! \sum_{r=0}^{k} (-1)^r / r! = k! \left(1/2! - 1/3! + \cdots + (-1)^k / k! \right).$$

Hence, the number of permutations of n objects that leave n – k objects fixed is given by $(e_k n!)/(n-k)!$.

Proof. We will prove this by induction. It is true for $k = 2$. Now, the total number of permutations can be expressed as the sum of permutations that leave exactly zero objects fixed, plus the permutations that leave exactly one object fixed, plus those which leave exactly two objects fixed, and so on up to k objects. Since we have C_r^n ways of choosing $r > 0$ objects that remain fixed and $(k-r)!\, e_{k-r}$ possibilities of permuting the rest (induction hypothesis), the number of possibilities we are seeking is written as

$$k! - \sum_{r=1}^{k} C_r^n (k-r)!\, e_{k-r} = k! \left(1 - \sum_{r=1}^{k} \frac{e_{k-r}}{r!} \right) = k! \left(\sum_{r=0}^{k} \sum_{l=0}^{k-r} \frac{(-1)^l}{(r+l)!} C_l^{r+l} \right) = k! \left(\sum_{l=0}^{k} \frac{(-1)^l}{(0+l)!} C_l^{0+l} \right) = k!\, e_k .$$

Let n_D and δ_D be the number of elements in D and its proportion in the database. The next result is the equation for the bias.

Result 2. *Let $D = P \cap F$, where P and F are the out-of-phase and fixed domains of D. The bias generated by using X' instead of X is given by*

$$B_\Sigma(X'_D) = \frac{k}{n-1}(X_D - \delta_F X_P).$$

If $w_i \sim w$, *the equation becomes* $B_\Sigma(X'_D) = \frac{n}{n-1} wn\tau\,(\,\delta_D - \delta_P\delta_F\,)$.

Proof. The bias is given in general by

$$B_\Sigma(X'_D) = \sum_{i\in D} w_i\ P_\Sigma(\sigma : \sigma i \notin F \mid i \in F) - \sum_{i\in P-D} w_i P_\Sigma(\sigma : \sigma i \in F \mid i \notin F). \tag{1}$$

Through a symmetry argument, we can show that these probabilities do not depend on i. Take the first probability. Let $i \in F$. To determine the number of ways, we remove i from the database, swap the remaining elements and replace i with an element from outside F. If we want to make sure that exactly k records have been swapped by the end of the operation, we need to know whether the record being replaced with i has already been swapped or not. The number of possible ways of swapping i with a record from outside F while keeping the number of swapped records at k is given by

$$e_{k-1}(k-1)!\sum_{j=0}^{k-1}(k-1-j)\,C_j^{n_F-1}\,C_{k-1-j}^{n-n_F} + e_{k-2}(k-2)!\sum_{j=0}^{k-2}(n-n_F-(k-2-j))\,C_j^{n_F-1}\,C_{k-2-j}^{n-n_F}\ .$$

(We control the number of records swapped in F.) After dividing by $(e_k n!)/(n-k)!$ and simplifying with $ke_k = (k-1)e_{k-1} + e_{k-2}$, we obtain $k/(n-1)\times\delta_{\bar{F}}$. Taking the complement of F, we find the second probability, i.e., $k/(n-1)\times\delta_F$. Substituting these probability terms in (1), we obtain the result.

The probabilities in (1) are important. They are known as the first-order outgoing and incoming probabilities. Thus, the bias is the outgoing proportion of the domain's mass minus the incoming proportion of the out-of-phase domain's mass. Clearly, an unbiased data swap for an estimate is a rare possibility. The quotient of the two domains has to be exactly equal to the quotient of the incoming and outgoing probabilities. When $w_i \sim w$, the relative bias becomes approximately the product of the permutation rate and the covariance between P and F. Hence, as soon as there is a non-negligible linear relation between the two domains, a bias will be present.

In what follows, we will denote $k/(n-1)$ as f_1, the swap's first-order factor. We have shown that the incoming and outgoing probabilities are equal to $f_1\delta_F$ and $f_1\delta_{\bar{F}}$ respectively. To write the second-order incoming and outgoing probabilities, which will be defined later, we set $f_2 = \frac{k(k-2)-(k-1)\,e_{k-1}/e_k}{(n-2)(n-3)}$, the swap's second-order factor. With these new notations, we can find the variance.

Result 3. *With the same assumptions as for result 2, the variance is given by*

$$\begin{aligned} V_\Sigma(X'_D) = {} & (f_1 - f_2)(1-2\delta_F)\sum_{i\in D} w_i^2 + \delta_F(f_1 - f_2\delta'_F)\sum_{i\in P} w_i^2 + \left(f_2 - f_1^2 + 2(f_1 - f_2)/n\right)X_D^2 \\ & - 2\left((f_1 - f_2)/n + (f_2 - f_1^2)\delta_F\right)X_P X_D + \delta_F\left(f_2\delta'_F - f_1^2\delta_F\right)X_P^2\,, \end{aligned}$$

where $\delta'_F = (n_F - 1)/(n-1)$. *If $w_i \sim w$, the formula can be written as*

$$\begin{aligned} V_\Sigma(X'_D) = {} & w^2 n\left(\delta_D(f_1 - f_2)(1 - 2(\delta_F + \delta_P - \delta_D)) + (\delta_F\delta_P(f_1 - \delta'_F f_2))\right) \\ & + w^2n^2\left(\delta_D(f_2 - f_1^2)(\delta_D - 2\delta_P\delta_F) + \delta_P^2\delta_F(f_2\delta'_F - f_1^2\delta_F)\right). \end{aligned} \tag{2}$$

Proof. As in the case of the bias, the expectation of the square of X' can be written as sums of weights over D or its complement in P, and those sums are multiplied by the joint probabilities of two records being either outgoing if

they are already in the domain or incoming if they are not. Since the permutation is chosen at random, the joint probabilities do not depend on the elements but only on whether they are in D or in $P - D$. They are referred to as second-order outgoing, mixed and incoming probabilities respectively. Denoting the probabilities as $\pi^{\overline{FF}}$, $\pi^{F\overline{F}}$ and π^{FF}, we can express the variance as

$$V_{\Sigma}(X') = \left(\pi^{\overline{F}} - \pi^{\overline{FF}}\right)\sum_{i\in D} w_i^2 + \left(\pi^{F} - \pi^{FF}\right)\sum_{i\in P-D} w_i^2 + \left(\pi^{\overline{FF}} - \pi^{\overline{F}}\pi^{\overline{F}}\right)X_D^2 - 2\left(\pi^{F\overline{F}} - \pi^{F}\pi^{\overline{F}}\right)X_D X_{P-D} + \left(\pi^{FF} - \pi^{F}\pi^{F}\right)X_{P-D}^2 .$$

Here, $\pi^{\overline{F}}$ and π^{F} are the first-order outgoing and incoming probabilities. The second-order probabilities are equal to $\pi^{FF} = f_2\delta_F\delta'_F$, $\pi^{\overline{FF}} = f_2\delta_{\overline{F}}\delta'_{\overline{F}}$ and $\pi^{\overline{F}F} = k/n(n-1) + f_2(\delta_F - 1/n)\delta'_{\overline{F}}$. Let's examine the formula for π^{FF}. For i, $j \notin F$, using the same technique as for result 1, if we have to choose two records in F to be replaced at the end of the process with i and j while controlling the total number of swaps, we have to control the number of elements in F that will be swapped. For example, the number of possibilities for the case in which i and j are replaced with two previously swapped elements is given by

$$(k-2)!\,e_{k-2}\sum_{i=0}^{k-2} i(i-1)\,C_i^{n_F}\,C_{k-2-i}^{n-n_F-2} = \frac{e_{k-2}(k-2)(k-3)\,n_F(n_F-1)}{(n-2)(n-3)}.$$

To conclude the proof, we have to find the sum of this number of possibilities plus all the possibilities of i and j being swapped with one fixed element and one swapped element of F plus all the possibilities of i and j being swapped with fixed elements of F. If we divide this number of possibilities by $(e_k n!)/(n-k)!$, we obtain the expression $f_2\delta_F\delta'_F$.

It is easy to see that when $\delta_F = 1$, $\delta_D = \delta_P$ and the variance vanishes. However, if $\delta_P = 1$, $\delta_D = \delta_F$ but there is still variation. This is due to the fact that w is a member of P. Similarly, if $w_i \sim w$, we can see, albeit after considerable algebraic manipulation, that $\delta_P = 1$ is sufficient to make the variance zero. If we set $\delta'_F \cong \delta_F$, an assumption shown to be true for non-negligible domains, the final term of (2) can be written as $w^2n^2\left(f_2 - f_1^2\right)\left(\delta_D - \delta_P\delta_F\right)^2$. Since $f_2 - f_1^2 \approx -\tau / n$ for sufficiently large n, it follows that the variance is proportional to n, which leads to the result we set out to prove.

Result 4. *With the same assumptions as for result 2, if $w_i \sim w$, the mean square error can be written as*

$$MSE_{\Sigma}(X'_D) \approx nw\tau\left|\,\delta_D - \delta_P\delta_F\,\right| + O\left(\sqrt{n}\right).$$

Proof. Under the assumptions mentioned, the variance is proportional to n. Hence, the dominant term of the error under the radical is the square of the bias. □

4. DATA SWAPPING AND THE EXTREMES

As we saw in the previous section, if w is nearly constant (which is the case in the Canadian Census), the mean square error is approximately the absolute value of the bias. Since the bias is proportional to n, if we apply this result to tabulations, data swapping will cause greater distortion in the data in tables for higher-level geographies (e.g., geographies with millions of units) than in the data in tables for small communities. However, an efficient statistical confidentiality protection technique should distort data only when the possibility of identification is non-negligible, which is the case only for tabulations for small communities. This leads to the following questions: are we generating enough noise when n is small, and are we generating too much noise when n is large? We will illustrate this point by comparing the noise produced by data swapping to the noise produced by random rounding in tabulations for two communities of very different sizes in Canada. We have chosen random rounding as the reference point be-

cause it is used by the Canadian Census. We will take age-marital status tabulations and subject them to data swapping of the second type with a permutation rate of 5% (AGE $\in A$ and STATUS$\in B$). Table 2 shows the tabulations before the protection measures were applied. Region II requires a substantial amount of protection (especially for widowed persons aged 65 and over). Region I needs essentially no protection.

Table 2: Age Groups by Marital Status for Two Regions

Region I ($w_i \approx w$)	Never married	Married	Separated	Divorced	Widowed	Total
0 - 15	165,268	–	–	–	–	**165,268**
16 - 35	148,313	21,316	1,524	973	120	**172,246**
36 - 65	73,508	294,438	21,130	44,795	10,051	**443,922**
65 +	8,830	65,701	2,170	6,248	43,614	**126,563**
Total	**395,919**	**381,455**	**24,824**	**52,016**	**53,785**	**907,999**
Region II ($w = 1$)	**Never married**	**Married**	**Separated**	**Divorced**	**Widowed**	**Total**
0 - 15	20	–	–	–	–	**20**
16 - 35	14	1	–	–	–	**15**
36 - 65	19	18	3	8	–	**48**
65 +	–	3	1	1	1	**6**
Total	**53**	**22**	**4**	**9**	**1**	**89**

Source: Canadian Census of Population. The geographies are not identified because confidentiality measures require rounding of all frequency counts.

After the data swap, we have Table 3. This process is somewhat unrealistic, of course, as we have just created widowed persons under the age of 15. It shows that we have to make sure that the newly created persons can pass all the survey controls. Data swapping is not as easy to implement as one might think. This way of implementing it is also unrealistic because it separates age from marital status. Ordinarily, the data swap designer will want to control this cross-tabulation of variables (e.g., swap selected variables within the unmarried 15-and-under group). However, we want to illustrate the damage caused by the proportionality of noise to cell size.

Table 3: Age Groups by Marital Status for Two Regions ($\tau = 5\%$)

Region I (unequal w)	Never married	Married	Separated	Divorced	Widowed	Total
0 - 15	161,108	3,021	187	411	541	**165,268**
16 - 35	144,962	23,557	1,655	1,339	733	**172,246**
36 - 65	79,115	289,028	20,609	44,047	11,123	**443,922**
65 +	11,019	65,003	2,242	6,192	42,107	**126,563**
Total	396,204	380,609	24,693	51,989	54,504	**907,999**
Region II (equal w)	**Never married**	**Married**	**Separated**	**Divorced**	**Widowed**	**Total**
0 - 15	20	–	–	–	–	**20**
16 - 35	13	2	–	–	–	**15**
36 - 65	20	17	3	8	–	**48**
65 +		3	1	1	1	**6**
Total	**53**	**22**	**4**	**9**	**1**	**89**

Let's look at the count of married persons aged 36 to 65 for region I. There is a difference of 5,410 (the calculated error is 5,800). The bias accounts for 99% of the theoretical value for the error. In each age group, the modal status loses mass to the other statuses. This phenomenon can be identified statistically, revealing a fictitious macroscopic property. If we look at the table for region II, on the other hand, we have to ask ourselves whether data swapping provides adequate protection for respondents. The error for the number of married persons aged 36 to 65 is 0.66, 42% of which is explained by the bias. Clearly, in this case, it is not the possibility of identifying a new phenomenon that causes the problem but the inability to protect the data.

Table 4 shows the same tabulations computed with unbiased random rounding. The noise generated by this method is less than 2.5 for a given frequency count. Thus, when the count is large, the noise becomes marginal. In the case of married persons aged 36 to 65 in region I, for instance, noise accounts for less than 0.0008% of the count. By

comparison, noise accounts for 14% of the corresponding count for region II. Since we know the formulas for the mean square error of such swaps, we can work backwards and determine how many records have to be swapped to obtain the same amount of noise as random rounding generates for the same count. For regions I and II, the values are $k = 2$ and $k = 30$ respectively. The obvious conclusion is that data swapping has difficulties when the frequencies are either very small or very large.

Table 4: Age Groups by Marital Status for Two Regions (Rounding with $b = 5$)

Region I (unequal w)	Never married	Married	Separated	Divorced	Widowed	Total
0 – 15	165,265	–	–	–	–	**165,270**
16 - 35	148,310	21,315	1,525	975	120	**172,245**
36 - 65	73,510	294,440	21,130	44,795	10,055	**443,925**
65 +	8,830	65,700	2,170	6,250	43,615	**126,565**
Total	**395,920**	**381,455**	**24,825**	**52,015**	**53,785**	**908,000**
Region II (equal w)	**Never married**	**Married**	**Separated**	**Divorced**	**Widowed**	**Total**
0 - 15	20	–	–	–	–	**20**
16 - 35	15	–	–	–	–	**15**
36 - 65	15	20	–	5	–	**45**
65 +	–	–	–	–	–	**10**
Total	**55**	**25**	**5**	**10**		**85**

5. CONCLUSION

As defined in this article, data swapping introduces noise in the data; fictitious persons are created by swapping the values of selected variables for a small number of records. We have shown that, for tabulations, the noise introduced is proportional to cell size and is therefore highly coloured. This results in a substantial loss of efficiency at the upper and lower ends of the distribution of cell sizes: significant erosion of the modes for large frequency counts and inability to protect the data for small counts. Designers of confidentiality measures can always use this method to complement other measures. In the case of a microdata file, for example, after applying all of the standard techniques, such as collapsing categories, the designer has the option of swapping the geographies of certain records which, in his view, are still at risk.

It is important to obtain an accurate measurement of the distortion added to the data by any confidentiality protection measure. Because of its probabilistic properties, data swapping is a good example because its efficiency can be analyzed methodically.