# Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure

Natalie Shlomo, Ton de Waal

## Abstract

Before releasing statistical outputs, data suppliers have to assess if the privacy of the statistical units is endangered and apply Statistical Disclosure Control (SDC) methods if necessary. SDC methods perturb, modify or summarize the data, depending on the format for releasing the data, whether as micro-data or tabular data. The goal is to choose an optimal method that manages disclosure risk below a tolerable risk threshold while ensuring high utility and high quality statistical data. In this article we first overview several SDC methods for continuous and categorical micro-data. All the methods perturb the data in some way. Changing values, however, will cause fully edited records in micro-data to fail edit constraints (i.e., logical rules or edits), resulting in low utility data. Moreover, an inconsistent record will signal it as having been perturbed for disclosure control and attempts can be made to unmask the data. In order to deal with these problems, we develop new implementation methods for the perturbation and minimize record level edit failures as well as overall measures which assess information loss and utility. This is done by perturbing within control strata and imputing for failed edits, ensuring additivity constraints, and preserving totals, means and covariance matrices.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M06/16

# Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure

Natalie Shlomo[a] and Ton de Waal[b]

[a]Southampton Statistical Sciences Research Institute, University of Southampton and
Department of Statistics, Hebrew University; natalieshlomo@cc.huji.ac.il

[b]Statistics Netherlands; twal@cbs.nl

**Summary**. Before releasing statistical outputs, data suppliers have to assess if the privacy of the statistical units is endangered and apply Statistical Disclosure Control (SDC) methods if necessary. SDC methods perturb, modify or summarize the data, depending on the format for releasing the data, whether as micro-data or tabular data. The goal is to choose an optimal method that manages disclosure risk below a tolerable risk threshold while ensuring high utility and high quality statistical data. In this article we first overview several SDC methods for continuous and categorical micro-data. All the methods perturb the data in some way. Changing values, however, will cause fully edited records in micro-data to fail edit constraints (i.e., logical rules or edits), resulting in low utility data. Moreover, an inconsistent record will signal it as having been perturbed for disclosure control and attempts can be made to unmask the data. In order to deal with these problems, we develop new implementation methods for the perturbation and minimize record level edit failures as well as overall measures which assess information loss and utility. This is done by perturbing within control strata and imputing for failed edits, ensuring additivity constraints, and preserving totals, means and covariance matrices.

**Keywords.** Additive noise, Information loss, Micro-aggregation, Post-randomization method, Rank swapping, Rounding, Statistical disclosure control.

## 1. Introduction

The aim of statistical disclosure control (SDC) is to prevent sensitive information about individual respondents from being disclosed. SDC is becoming increasingly important due to the growing demand for information provided by Statistical Agencies. The information released by Statistical Agencies can be divided into two major forms of statistical data: tabular data and micro-data. Whereas tables have been released traditionally by Statistical Agencies, micro-data sets released to researchers is a relatively new phenomenon. Nowadays, many Statistical Agencies have provisions for releasing micro-data from social surveys for research purposes usually under special license agreements and through secure data archives. Micro-data from business surveys are typically not released because of their disclosive nature due to high sampling fractions and skewed distributions. In order to preserve the privacy and confidentiality of individuals responding to social surveys, Statistical Agencies must assess the disclosure risk in micro-data and if required choose appropriate SDC methods to apply to the data (see also Willenborg and De Waal, 2001; Shlomo and De Waal, 2005; Shlomo and Young, 2006; and Willenborg and Van den Hout, 2006). Measuring disclosure risk for the SDC decision problem involves assessing and evaluating numerically the risk of re-identifying statistical units. SDC methods perturb, modify, or summarize the data in order to prevent re-identification by a potential attacker. Higher levels of protection through SDC methods however impact negatively on the utility and quality of the data. The SDC decision problem involves finding the optimum balance between managing disclosure risk to tolerable thresholds depending on the mode for accessing the data and ensuring high utility in the data.

In any released micro-data set direct identifying key variables, such as name, address or identification numbers, have obviously been removed for else identification of units would be quite trivial. Disclosure risk typically arises from attribute disclosure where small counts on cross-classified indirect identifying key variables (such as: age, sex, place of residence, marital status, occupation, etc.) can be used to identify an individual and confidential information may be learnt. Generally, identifying variables are categorical ones. Sensitive variables are often continuous ones, but can also be categorical.

We will illustrate the problem by means of an example. Suppose, for instance, that a micro-data set containing a sample of the participants of the UN/ECE Work Session on Statistical Data Editing held in September 2006 in Bonn (where a version of the present article was presented) were released. Suppose furthermore that the micro-data set contains information on the affiliation of authors and their co-authors, and sensitive information on, for instance, the income (a continuous variable) of the authors. Now consider the record: “Affiliation author = Statistics Netherlands”, “Affiliation co-author = University of Southampton”, and “Income = 95,000 euro”. At the UN/ECE Work Session on Statistical Data Editing in Bonn there was only one author from Statistics Netherlands with a co-author from the University of Southampton. If the record were released in this form, it would be quite easy to re-identify this person and disclose that his income is 95,000 euro.

The above example shows that micro-data may need to be protected against disclosure. However, absolute prevention of disclosure of sensitive information about individual respondents can only be guaranteed if no or hardly any information is released. This aim would therefore be far too restrictive for Statistical Agencies. A more realistic aim is to limit the probability that sensitive information about individual respondents can be disclosed. This aim can be achieved by applying SDC techniques.

SDC techniques for micro-data include perturbative methods which alter the data and non-perturbative methods which limit the amount of information released in the micro-data without actually altering the data. Examples of non-pertubative SDC techniques are global recoding, suppression and sub-sampling (see e.g. Willenborg and De Waal, 2001). Perturbative methods for continuous variables (see Section 2) include adding random noise, micro-aggregation (replacing values with their average within groups of records), rounding to a pre-selected rounding base, and rank swapping (swapping values between pairs of records within small groups). Perturbative methods for categorical variables (see Section 3) include record swapping (typically swapping geography variables) and a more general post-randomization probability mechanism (PRAM) where categories of variables are changed or not changed according to a prescribed probability matrix and a stochastic selection process.

With non-perturbative SDC methods, the logical consistency of the records remain unchanged and so-called edit rules, or edits for short, will not begin to fail as a result of these methods. Edits describe either logical relationships that have to hold true, such as “a two-year old person cannot be married” or “the profit and the costs of an enterprise should sum up to its turnover”, or relationships that have to hold true in most cases, such as “a 12-year old girl cannot be a mother”.

Perturbative methods, however, alter the data, and therefore we expect consistent records to start failing edits due to the perturbation

In this article we focus on perturbative SDC techniques to protect micro-data against disclosure. We provide an overview of the most common perturbative SDC techniques found in the literature and show how they can be extended and modified so as to take edits into account. We also demonstrate new implementation methods that preserve sufficient statistics (totals, means and covariance matrices). This ensures a high level of utility in the data. We generally propose several alternatives for a given SDC method, and provide some results obtained from evaluation studies to illustrate the effects of these alternatives on the information loss.

The innovative aspect of our work is the extension to edits so that the protected data are consistent and have high utility. For some academic statisticians the emphasis on consistent data, i.e. the wish of Statistical Agencies to let the data satisfy specified edits, may be difficult to understand. Statistically speaking there is indeed hardly a reason to let a data set satisfy edits, apart from hoping that enforcing internal consistency results in data of higher statistical quality. Statistical Agencies, however, have the responsibility to supply data for many different, both academic and non-academic, users in society. For the majority of these users, inconsistent data are incomprehensible. They may reject the data as being an invalid source or make adjustments themselves. This hampers the unifying role of the Statistical Agency in providing data that are undisputed by different parties such as policy makers in government, opposition, trade unions, employer organizations etc. As mentioned by Särndal and Lundström (2005, p. 176) in the context of imputation “Whatever the imputation method used, the completed data should be subjected to the usual checks for internal consistency. All imputed values should undergo the editing checks normally carried out for the survey”. This holds even more true in our context of protecting micro-data against disclosure as inconsistent perturbed records may pinpoint to potential intruders that these records have been protected.

In order to protect a data set by means of perturbative techniques one can either perturb the identifying variables or perturb the sensitive variables. In the first case identification of a unit is rendered more difficult, and the probability that a unit is identified is hence reduced. In the second case, even if an attacker succeeds in identifying a unit by using the values of the indirectly identifying key variables, the sensitive variables would hardly disclose any useful information on this particular unit as they have been perturbed. One can also perturb both the identifying variables and the sensitive ones simultaneously. This offers more protection, but also leads to more information loss.

We illustrate the above by returning to our simple example. As the value of income is considered relatively high (at least for statisticians working at Statistics Netherlands!) this record is considered to require protection. Say we decide to protect the record by adding noise to the variable “income”. Without the presence of edits, one would draw a value from an appropriate probability distribution. Say, we draw a value of 20,000 euro and obtain a perturbed income of 115,000 euro. Suppose, however, that it is well-known that statisticians working at Statistics Netherlands never earn more 100,000, which we use as an edit rule. As this edit is violated an attacker would be able to conclude that this records has been protected; probably because the true value of income was relatively high. In this very simple case we can take this edit into account by drawing again. Say, we now draw a value of -35,000 euro. We then obtain an income of 60,000 euro. The record seems sufficiently protected now.

We can also decide to perturb the categorical identifying variables by means of PRAM (see Section 3). The record “Affiliation author = Statistics Netherlands”, “Affiliation co-author = University of Southampton”, “Income = 95,000 euro” might then be modified into a record “Affiliation author = Statistics Canada”, “Affiliation co-author = University of Southampton”, “Income = 95,000 euro”. However, at the UN/ECE Work Session on Statistical Data Editing in Bonn there was no author from Statistics Canada with a co-author from the University of Southampton. This (edit) rule is violated by our “protected” record. This inconsistency might trigger a potential attacker to further examine and unmask this record. The record we have obtained after application of PRAM, “Affiliation author = Statistics Canada”, “Affiliation co-author = University of Southampton”, “Income = 95,000 euro”, can be processed further by imputing values for the non-perturbed data in such a way that a feasible record results. Suppose that we impute “Affiliation co-author” and obtain a record “Affiliation author = Statistics Canada”, “Affiliation co-author = Statistics Canada”, “Income = 95,000 euro”. This is a feasible record as there were couples of authors and co-authors from Statistics Canada at the UN/ECE Work Session on Statistical Data Editing. In fact, there were more than one couple, implying that the final record cannot be misused to falsely deduce that a specific author from Statistics Canada has an income of 95,000 euro. Note that sensitive categorical data in a micro-data set can also be protected by means of PRAM.

SDC techniques have received ample attention in the literature. However, SDC techniques for micro-data that take edits into accounts is a new topic that only recently has received attention by researchers from academia and official statistics. Willenborg and Van den Hout (2006) have examined an SDC technique that takes edits into account. A difference between their article and ours is that Willenborg and Van den Hout focus on one particular SDC technique, which they refer to as Peruco. This is a deterministic delete/impute method for only those records having unsafe combinations with respect to a frequency threshold. This method introduces bias into the micro-data and there is no guarantee that marginal distributions, means and variance estimates are preserved.

The application of SDC measures to prevent the disclosure of sensitive data leads to a loss of information. It is therefore important to develop quantitative information loss measures in order to assess whether the resulting disclosure controlled micro-data set is fit for purpose. Obviously, information loss measures should be minimized in order to ensure high utility. Information loss measures assess the impact on statistical inference: the effects on bias and variance of point estimates, distortions to distributions, effects on goodness of fit criteria for statistical modeling, etc. Assessing the information loss of the various SDC methods that we consider in this article is an important aspect of our article.

The article is split into two parts: Section 2 describes the perturbation of sensitive continuous variables and Section 3 describes the perturbation of identifying categorical key variables. In Section 2, Sub-Sections 2.1 through 2.4 describe the SDC methods under analysis in this article: additive noise, micro-aggregation, rounding, and rank swapping. This analysis is carried out on survey micro-data from the 2000 Israel Income Survey where the variables that are perturbed are all continuous income variables: gross income, net income and taxes. In Section 3, Sub-Section 3.1 describes the Post-randomization (PRAM) methodology which generalizes other SDC methods such as record swapping and impute/delete techniques. Sub-Section 3.2 describes the evaluation dataset based on the 1995 Israel Census sample, including the edit constraints. Sub-Section 3.3 presents the algorithm for implementing PRAM under various methods of controlling variables in order to minimize edit failures and maximize data utility. Sub-Section 3.4 presents results of the algorithm and the impact on the edits and information loss. Finally, we conclude in Section 4 with a discussion on the entire analysis and future work.

## 2. Perturbation of sensitive continuous variables

### 2.1 Protecting continuous variables by means of additive noise

Additive noise is an SDC method that is carried out on continuous variables. In its basic form random noise is generated identically independently distributed with a mean of zero in order to ensure that no bias is introduced into the original variable and a positive variance. The random noise is then added to the original variable. It has been shown that re-identification can occur using this SDC method based on probabilistic record linkage techniques (Yancey, Winkler and Creecy, 2002). This has led towards some development of mixture models for generating random noise which achieve higher protection levels. Adding random noise will not change the mean of the variable but may introduce more variance for the estimate of the mean of the variable. This will impact on the ability to make statistical inference, particularly for estimating parameters in a regression analysis. In this section we examine several methods for adding random noise which focus on preserving edits and minimizing information loss measures.

As mentioned in the Introduction, adding noise to a variable such as income may cause edit failures at the record level. For example, consider the edits:

E1a: gross income (*gross*) $\geq 0$,

E1b: net income (*net*) $\geq 0$,

E1c: taxes (*tax*) $\geq 0$

and

E2: IF age $\leq 17$ THEN gross income $\leq$ mean income.

Adding noise across the whole file may cause these edits to start failing. For example, in the 2000 Israel Income Survey, out of 32,896 individuals aged 15 and over surveyed, 16,232 individuals earned an income from salaries. The mean of their gross income from salaries was 6,910 IS (Israeli shekel) with a standard deviation of 7,180 IS. Random noise is generated using a normal distribution with a mean of 0 and a variance that is 20% of the variance of the income variable ($0.2 \times 7{,}180^2$). After adding the random noise to the income variable *gross*, 1,685 individuals failed the non-negativity edit E1a and out of 119 individuals under the age of 17, 6 individuals failed edit E2. It is clear that more control should be placed into the perturbation scheme in order to minimize the number of failed edits.

More control can be achieved by generating, for example, random noise for each strata defined by percentiles (for example, quintiles) of gross income as follows: sort the file according to the variable *gross*; define quintiles; generate random noise separately in each quintile using 20% of the variance of the variable *gross* in the quintile as described above. Based on this method, we obtain that now only 66 individuals fail the non-negativity edit E1a and no individuals under the age of 17 fail edit E2. Moreover, based on the first method using the overall variance of the variable *gross*, the resulting perturbed variable had a standard deviation of 7,849 compared to 7,180. However, when perturbing the variable *gross* within quintiles, this led to an increase in the standard deviation to only 7,487.

In order to reduce information loss, we can also carry out a method for generating additive random noise that is correlated with the variable to be perturbed, thereby ensuring that not only are means preserved but also the variance. Some methods for generating correlated random noise have been discussed in the literature based on transformations and fixed parameters (Kim, 1986, Fuller 1993, Brand, 2002, Yancey, Winkler and Creecy, 2002). We propose, however, an alternative method for generating correlated random noise that preserves means and variances that is very easy to implement. We demonstrate our method first on the univariate income variable *gross*. Define $\delta$ which controls the amount of random noise added and calculate: $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$ .

Now, generate random noise $\varepsilon$ with a mean of $\mu' = \frac{1-d_1}{d_2}\mu$ , where $\mu = 6{,}910$ is the mean of the original income variable, and a variance $\sigma^2 = 7{,}180^2$ of the original income variable. Calculate the perturbed variable as a linear combination: $gross' = d_1 \times gross + d_2 \times \varepsilon$ .

Note that $E(gross') = d_1 E(gross) + d_2[\frac{1-d_1}{d_2} E(gross)] = E(gross)$ and

$Var(gross') = (1-\delta^2)Var(gross) + \delta^2 Var(gross) = Var(gross)$ .

As defined earlier, the above method can also be carried out within quintiles in order to minimize the number of edit failures. Indeed, based on this method within quintiles and using $\delta` = 0.3$ (which is similar to the amount of noise generated earlier), we obtain that now only 9 individuals fail the non-negativity edit E1a and no individuals under the age of 17 fail edit E2. Moreover, the overall standard deviation of the perturbed variable has remained unchanged with a value of 7,198.

An additional problem when adding random noise is that there may be several variables to perturb at once, and these variables may be connected through an edit constraint. For example, again consider in the 2000 Israel Income Survey the three variables: *gross*, *tax* and *net*. The original micro-data set that has undergone edit and imputation processing will have ensured that no records fail the following edit:

E3: *net* + *tax* = *gross*.

However, after perturbing each variable separately, this edit constraint will not be guaranteed. Therefore, we considered two possibilities for adding noise, preserving means and the co-variance matrix and preserving the edit constraint of additivity:

- Split the procedure into two separate processes: (1) first carry out the perturbation method of adding random noise on each of the variables as described above; (2) implement an additional stage of post-editing for correcting the additivity of the variables based on linear programming under the minimum change paradigm. This linear programming can be carried out as follows.

  Let the number of continuous variables be given by *n*. Denote the perturbed continuous variables after the first step by $x_i$ (*i*=1,…,*n*) and the adjusted perturbed continuous variables by $\hat{x}_i$ (*i*=1,…,*n*). The linear programming problem for the second step can then be formulated as

  $$\text{minimize } \sum_i w_i \,|\, x_i - \hat{x}_i \,|,$$

  subject to the constraint that the $\hat{x}_i$ (*i*=1,…,*n*) satisfy all edits. Here the $w_i$ (*i*=1,…,*n*) are non-negative weights expressing how serious a change in the *i*-th perturbed value is considered to be. In our case, we perturb variables *tax* and *net*, so *n* = 2. The constraints are based on: non-negativity (edits E1a, E1b and E1c), and additivity to the fixed total (*gross*). We also aim to preserve the ratio $x_1/x_2$ before and after the adjustments, where $x_1$ denotes the value of *tax* and $x_2$ the value of *net*. Aiming to preserve the ratio $x_1/x_2$ before and after perturbation gives us another linear constraint, namely $x_1/x_2 = \alpha$ leads to $x_1 = \alpha x_2$. The resulting linear programming problem can easily be solved, e.g. by means of the EXCEL solver.

- Implement the procedure in only one step by a priori generating additive random noise variables that preserve the edit constraint of additivity. Therefore when combining the constrained additive random noise to the original values of the variables, the additivity of the final perturbed variables is preserved. Note also that we want to maintain same means and same co-variance structure before and after the perturbation. This technique is described as follows.

  Generate multivariate random noise in each quintile (note that we drop the index for quintile: $(\varepsilon_{(GROSS)}, \varepsilon_{(NET)}, \varepsilon_{(TAX)})^T \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma})$, where the superscript T denotes the transpose. The vector $\boldsymbol{\mu}'$ contains the corrected means of each of the three variables: gross income, net income and taxes: $\boldsymbol{\mu}'^{\mathrm{T}} = (\boldsymbol{\mu}'_{(GROSS)}, \boldsymbol{\mu}'_{(NET)}, \boldsymbol{\mu}'_{(TAX)}) = (\frac{1-d_1}{d_2}\boldsymbol{\mu}_{(GROSS)}, \frac{1-d_1}{d_2}\boldsymbol{\mu}_{(NET)}, \frac{1-d_1}{d_2}\boldsymbol{\mu}_{(TAX)})$. The matrix $\boldsymbol{\Sigma}$ is the original covariance matrix. For each separate variable, calculate the linear combination of the original variable and the random noise as described earlier, for example: $gross' = d_1 \times gross + d_2 \times \varepsilon_{(GROSS)}$ using a parameter $\delta$. The mean vector and the covariance matrix remain the same before and after the perturbation, and the additivity is exactly preserved.

We used the parameter $\delta = 0.3$ for the linear combination between the original variables and their generated noise. For our data set, there were only 3 individuals that failed the non-negativity edit E1c based on the variable *tax* and no individuals failed the non-negativity edits for the other income variables *net* and *gross* (edits E1a and E1b). No individuals failed edit E2 for any of the income variables. To correct for the negativity of the variable *tax*, the value was set to zero and the other variables *gross* and *net* adjusted accordingly to ensure the preservation of the additivity edit E3.

## 2.2 Protecting continuous variables by means of micro-aggregation

Micro-aggregation is another disclosure control technique for continuous variables. Records are grouped together in small groupings of size *k*. For each individual in the group, the value of the variable is replaced with the average of the values of the group to which the individual belongs. This method can be carried out both on a univariate or multivariate setting where the latter is implemented through sophisticated computer algorithms. In this article, we focus on the simple univariate case.

Replacing values of variables with their average in a small group will not initiate edit failures of the types described in E1 and E2, although there may be problems at the boundaries and the edits may have to be adjusted slightly. Micro-aggregation preserves the mean (and the overall total) of the income variable but will lead to a decrease in the variance of the mean because of the following reason.

Let *n* be the sample size, *m* the number of groups of size *p*. The variance components are:

$$SST: \quad \sum_{i=1}^{m}\sum_{j=1}^{p}(X_{ij}-\overline{X})^2 \qquad n\text{-1 degrees of freedom} \qquad (1)$$

$$SSB: \quad \sum_{i=1}^{m}p(\overline{X}_i-\overline{X})^2 \qquad m\text{-1 degrees of freedom} \qquad (2)$$

$$SSW: \quad \sum_{i=1}^{m}\sum_{j=1}^{p}(X_{ij}-\overline{X}_i)^2 \qquad n\text{-}m\text{ degrees of freedom} \qquad (3)$$

The total sum of squares *SST* of the income variable $X_i$ (for $i$=1,…,$n$) can be broken down into the “within” sum of squares *SSW* which measures the variance of the mean income variable within the groups and the “between” sum of squares *SSB* which measures the variance of the mean income variable between the groups. When implementing micro-aggregation and replacing values by the average of their group, the variance that is calculated is based on the *SSB* only and not *SST*. In general, there may not be that much difference between *SST* and *SSB* since the size of the groups *p* is small and this results in a very small *SSW*. In order to minimize this information loss measure of a decrease in the variance, we can generate random noise according to the magnitude of the difference between the two variances and add it to the micro-aggregated variable. Besides raising the variance back to its expected level, this method will also result in extra protection against the risk of disclosure since it was shown in Winkler (2002) that micro-aggregation (and in particular univariate micro-aggregation) can be “unpicked” by intruders using elementary software.

We demonstrate our algorithm of adding random noise to a micro-aggregated variable for the 15,708 individuals that paid tax from among the 16,232 individuals that earned an income in the 2000 Israel Income Survey. We define small groupings of size 5 where the last grouping may contain less than 5 units. We define the groupings within the quintiles as defined in Section 2 in order to ensure that edits of types E1 and E2 will not begin to fail as a result of adding random noise. In each small group, the value of the variable *tax* is replaced by the average of the group. To generate random noise for each quintile, we calculate the difference between the two variances *SST* and *SSB* and generate the random normal distributed noise with a mean of zero and a variance equal to the difference. Table 1 presents the standard deviations for the mean of the variable *tax* at the different stages of the micro-aggregation/additive random noise process. Note that 8 individuals failed edit E2 with a negative value for the perturbed variable *tax.* These individuals had their perturbed value changed to zero.

[PLACE TABLE 1 AROUND HERE]

To ensure the edit constraint E3 based on the additivity of the three income variables, note that carrying out the micro-aggregation on each of the three variables within group *i* will preserve the additivity since the sum of the means of the two variables *net* and *tax* will equal the mean of the total variable *gross*. In order to ensure the correct variance for the means of the variables, we can generate random noise separately for each variable as described earlier. However generating random noise separately will not result in preserving the additivity and therefore the linear programming technique will have to be applied.

Another method which will preserve the additivity edit E3 is to generate multivariate normal noise which a priori preserves the edit constraint as defined in Section 2.1: $(\varepsilon_{(GROSS)} \quad \varepsilon_{(NET)} \quad \varepsilon_{(TAX)})^T \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma})$. For each of the variables, we define the linear combination of the group mean $\mu_i$ where *i* is the small group. Let *r*(*i*) be the quintile of *i*. The random noise variable is generated within quintiles. For example, the perturbed variable *gross* in group *i* belonging to quintile *r*(*i*) is equal to: $gross_i' = d_1 \times \mu_i + d_2 \times \varepsilon_{r(i)}$ where $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$ as defined in Section 2.1. Since the random multivariate noise itself maintains the additivity property, the additivity will hold when combining the random noise with the group means for each of the three income variables. However, this algorithm will not completely return the original level of the true variance since:

$$Var(gross_i') = (1-\delta^2)Var(\mu_i) + \delta^2 Var(gross_i) = Var(\mu_i) + \delta^2[Var(gross_i) - Var(\mu_i)].$$

The last term is the "within" variance and therefore the only way to get back the full covariance structure is to define $\delta = 1$. This however is the definition of synthetic data which is out of scope of this article. By increasing $\delta$ slightly we can gain back most of the original variance, although if $\delta$ is too high then edits of types E1 and E2 will likely begin to fail.

We compare these two methods of preserving additivity and ensuring correct variance estimation. Adding random noise separately to each variable, *gross*, *net* and *tax* resulted in correcting the variance but large discrepancies occurred between the sum of variables *net* and *tax* and the total variable *gross*. In this process, 9 records failed edit E1b with a negative perturbed value for *tax*. These values were changed to zero. Table 2 presents the absolute difference between the perturbed variable *gross* and the sum of the perturbed variables *net* and *tax*.

[PLACE TABLE 2 AROUND HERE]

In order to correct these differences, the linear program technique as described in Section 2.1 (first bullet point) was applied. This resulted in the preservation of the additivity constraint, no additional edit failures and also preserved the original ratio between the adjusted perturbed variable *tax* and the adjusted perturbed variable *net*.

Adding correlated noise with a slightly higher $\delta = 0.5$ preserved the additivity constraint E3. Some edit failures occurred using this high value for $\delta$: 47 out of the 16,232 records had negative values in one of the variables. These were corrected automatically by setting them to zero and adjusting the additivity of the other variables.

Table 3 summarizes the standard deviations of the means of the variables *gross*, *tax* and *net* at the different stages of micro-aggregation for both procedures: additive noise and the linear programming to preserve the edits and adding correlated random noise.

[PLACE TABLE 3 AROUND HERE]

Comparing these two methods in Table 3 for improving the micro-aggregation with respect to maintaining edits and the preservation of the variance, it appears that the first procedure based on adding random noise and the linear programming to preserve additivity achieves these aims with the final variance structure closer to the original variance structure. Both of the methods are similar with respect to the resulting correlation structure between the three income variables.

### 2.3 Protecting continuous variables by means of rounding

Rounding to a predefined base is a form of adding noise, although in this case the exact width of the perturbation is known a priori and can be controlled. Therefore, it is likely that edits of types E1 and E2 will not fail due to the rounding. However, rounding continuous variables separately may cause edit failures of the type defined by E3 since the sum of rounded variables will not necessarily equal their rounded total. Indeed, there are some software applications (and in particular the Tau-Argus Statistical Disclosure Control Software Package developed within the framework of the European Initiative CASC. see Salazar-González, Bycroft and Staggemeier, 2005) that have a controlled rounding option based on sophisticated linear programming which preserves the additivity of the rounded numbers. This method however is biased and in addition, the option is not always available to data suppliers.

In our case, where we are dealing with micro-data with rather simple edit restrictions, rounding procedures can be relatively easy to implement, similar to the problem of rounding one or two dimensional tables. In this example, we describe a one dimensional random rounding procedure which not only has the property that it is stochastic and unbiased, but it can be carried out in such a way as to preserve the exact overall total (and hence the mean) of the variable being rounded. The algorithm is as follows. Let *x* be the value to be rounded and let $Floor(x)$ be the largest multiple *k* of the base *b* such that $bk < x$. In addition, define the residual of *x* according to the rounding base *b* by $res(x) = x - Floor(x)$. For an unbiased random rounding procedure, *x* is rounded up to $(Floor(x) + b)$ with probability $res(x)/b$ and rounded down to $Floor(x)$ with probability $(1 - res(x)/b)$. If *x* is already a multiple of *b*, it remains unchanged. The expected value of the rounded entry is the original entry. The rounding is usually implemented with replacement in the sense that each entry is rounded independently, i.e. a random uniform number *u* between 0 and 1 is generated for each entry. If $u < res(x)/b$ then the entry is rounded up, otherwise it is rounded down. The expectation of the rounding is zero and no bias should remain in the table. However, the realization of this stochastic process on a finite number of values in micro-data may lead to overall bias since the sum of the perturbations (i.e., the difference between the original and rounded value) going down may not necessarily equal the sum of the perturbations going up. In order to preserve the exact total of the variable being rounded, we define a simple algorithm for selecting (without replacement) which entries are rounded up and which entries are rounded down: for those entries having $res(x)$, randomly select a fraction of $res(x)/b$ of the entries and round upwards, the rest of the entries round downwards. Repeat this process for all $res(x)$.

Rounding as described above should be carried out within sub-groups in order to benchmark important totals. For example, rounding income in each group defined by age and sex will ensure that the total income in that group will remain unchanged. This may, however, distort the overall total across the whole file, although users are generally more interested in smaller sub-groups for analysis and therefore preserving totals for sub-groups is more important than the overall total. Reshuffling algorithms can be applied for changing the direction of the rounding for some of the records in order to correct the totals. This algorithm will be described in the paragraph below.

For our data set of 16,232 individuals that earned an income in the 2000 Israel Income Survey, we randomly round each of the variables *net* and *tax* to base 10. The method is carried out separately for each of the variables using the algorithm that controls and preserves the overall total. In order to ensure the edit of additivity E3, we calculate the rounded variable *gross* by summing the rounded variables *net* and *tax*. The rounded variable *gross* now has its overall total preserved (since the individual variables *net* and *tax* had their totals preserved), however since it is derived by adding the two rounded variables, this has caused the resulting sum to jump a base on some of the records. We carry out a reshuffling algorithm to correct this as follows:

1. Select the records with more than a difference of 10 (in absolute value) between the original variable *gross* and the rounded variable *gross* that was obtained by summing the rounded variables, *net* and *tax*;
2. Determine and select which of the variables *net* or *tax* had the most difference from its original value;

3. If the summed rounded variable *gross* was jumped to a higher base, drop the selected variable down a base and if the summed rounded variable *gross* was jumped to a lower base, raise the selected variable up a base.

The results of this procedure are presented in Table 4 and include the impact on the overall totals of each of the variables. Note that ensuring that the summed rounded variable gross is within the base has distorted slightly the controlled total. However the distortion is not large, especially when compared to the alternative of no controls in the totals.

[PLACE TABLE 4 AROUND HERE]

### 2.4 Protecting continuous variables by means of rank swapping

In its simplest version, rank swapping is carried out by sorting the continuous variable and defining groupings of size *k*. In each group, random pairs are selected and their values swapped. If the groupings are small, this method will not likely initiate edits to fail. In particular, the concern is for edits that are based on the logical consistency between highly correlated variables, such as edit E2 relating the level of income to age. This is because the method introduces bias on joint distributions that involve the swapped variable. Information loss measures that need to be minimized are based on minimizing the distortions to distributions and the effects on statistical inference. The larger the size of the groupings *k* the more possibilities of edit failures and loss of information, however the size of the groupings also impacts inversely on the disclosure risk, i.e. the larger the groupings the less disclosure risk. Therefore, a balance must be struck based on the parameter *k* which minimizes edit failures and information loss and also manages the disclosure risk to a tolerable risk threshold. Note that in order to preserve the edit of additivity as defined in edit E3, all variables involved in the edit would need to be swapped using the same paired record. Otherwise, adjustments could be carried out as defined by the linear programming approach described in Section 2.1 for preserving the additivity.

We demonstrate this method on the 16,232 individuals that earned an income in the 2000 Israel Income Survey based on the income variable gross. After sorting the variable, we define groupings of size 10 and of size 20, select random pairs in each group and swap the values of gross between each pair. No edits failed for either size grouping, and the original means and variances for the univariate variable gross are preserved. Next we examine some information loss measures based on the distortion to a particular joint distribution defined by cross classifying age groups (14), sex (2) and income groups (22).

The following information loss measures are used for our evaluation study.

Hellinger Distance: Let $x_i$ be the original cell count for a joint distribution and $\hat{x}_i$ the perturbed cell count. Also, let *n* again be the sample size. The Hellinger Distance metric is defined as:

$HD = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{\frac{x_i}{n}} - \sqrt{\frac{\hat{x}_i}{n}}\right)^2}$ . This is a symmetrical distance metric and measures how different two probability distributions are. Note that this measure takes into account the relative sizes of the original cell counts, i.e. the smaller the original cell count, the more impact on the Hellinger Distance. We use the Hellinger Distance to measure the distortion to the distribution defined by age groups × sex × income groups (616 cells) before and after rank swapping of gross income. The smaller the Hellinger Distance, the less information loss.

Cramer's *V*: Let *T* define a 2-dimensional frequency table spanned by two variables each having $C_1$ and $C_2$ number of cells and *n* is again the sample size. Define Cramer's *V* by: $V_{1,2} = \sqrt{\frac{\chi^2}{n \times \min((C_1 - 1),(C_2 - 1))}}$ where $\chi^2$ is the standard test statistic for independence. Cramer's *V* lies between 0 for no association and 1 for full association. The measure that defines the loss in the association when comparing $T_{orig}$ and $T_{pert}$ is $CV_{1,2} = V_{1,2}(T_{pert}) - V_{1,2}(T_{orig})$. We use the difference in Cramer's *V* statistic on the frequency table defined by combined age groups × sex on the rows and the income groups on the columns. The smaller the difference in Cramer's *V*, the less information loss. Moreover, the sign of the difference is important since this tells us whether we are attenuating a target variable or adding more artificial association into the table.

Impact on $R^2$: For a univariate analysis of variance (ANOVA), we assess the impact on the "between" variance, i.e. the impact on the $R^2$ statistic. $R^2$ is the ratio of the between sum of squares *SSB* to the total sum of squares *SST* (see Section 2.2). In this ANOVA analysis, we define the dependent variable as income *gross* and the independent variables as the cross-classified age groups × sex. The information loss measure is the ratio of the "between" variance of the perturbed distribution and the "between" variance of the original distribution, where the "between" variance is defined by: $BV = \frac{1}{p-1}\sum_{i=1}^{p} n_i(\bar{x}_i - \bar{\bar{x}})^2$, and *p* is the number of cells in age groups × sex (28 cells), $n_i$ is the sample size in cell *i*, $\bar{x}_i$ is the mean of *gross* in cell *i* and $\bar{\bar{x}}$ is the overall sample mean of *gross*. Note that an information loss measure below one indicates attenuation, i.e., the means in cells *i* (*i*=1,…,*p*) are flattening towards the overall mean of the distribution whereas a value above one indicates more of a dispersion in the cell means. Table 5 presents the results of these information loss measures.

[PLACE TABLE 5 AROUND HERE]

In general, as the size of the groupings increases, we obtain slightly more distortion to the distribution examined. There is almost no impact on the measures of association for the frequency table examined nor on the ratio of the between variance for the ANOVA analysis. The negative sign for the Cramer's *V* and the ratio of *BV* smaller than one indicates that as the size of the groupings increases, we are indeed attenuating target variables across the distribution.

## 3. Perturbation of identifying categorical key variables

### 3.1 Protecting categorical variables by means of PRAM

PRAM is a method used for changing values of categorical variables for certain records in the original data to other categories according to a prescribed probability mechanism (Gouweleeuw et al, 1998). It is analogous to adding random noise to continuous variables. In this method, values of categories are changed or not changed according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. The prescribed probability matrix can be developed in such a way as to preserve the expected marginal frequencies of the original variable and thus minimize the information loss. Indeed, using a more deterministic approach in the actual perturbation process, the exact marginal distributions can also be maintained. This method was used to perturb the Sample of Anonymized Records (SARs) of the 2001 UK Census (Gross, Guiblin and Merrett, 2004).

The probability mechanism can be taken into account when making statistical inferences. We define a perturbation method in which a value in a record is moved from category $i$ to category

$j$ with probability: $p_{ij} = p(\text{perturbed category is } j \mid \text{original category is } i)$. Let $\mathbf{P}$ be a $L \times L$ transition matrix containing the conditional probabilities $p_{ij}$ for a categorical variable with $L$ categories. Let $\mathbf{t}$ be the vector of frequencies and $\mathbf{v}$ the vector of its relative frequencies: $\mathbf{v} = \mathbf{t}/n$ , where $n$ is the number of records in the micro-data set. On each record of the data set, the category of the variable is changed or not changed according to the prescribed transition probabilities in the matrix $\mathbf{P}$ and the result of a draw of a random multinomial variate u with parameters $p_{\mathrm{ij}}$ ($j$=1,…,$L$). If the $j$-th category is selected, category $i$ is moved to category $j$. When $i = j$, no change occurs**.**

Let $\mathbf{t}^*$ be the vector of the perturbed frequencies. We note that $\mathbf{t}^*$ is a random variable and $\mathrm{E}(\mathbf{t}^* \mid \mathbf{t}) = \mathbf{tP}$. Assuming that the transition probability matrix $\mathbf{P}$ has an inverse $\mathbf{P}^{-1}$, this can be used to obtain an unbiased moment estimator of the original data: $\hat{\mathbf{t}} = \mathbf{t}^*\mathbf{P}^{-1}$. Statistical analysis can be carried out on $\hat{\mathbf{t}}$. In order to ensure that the transition probability matrix has an inverse and to control the amount of perturbation, the matrix $\mathbf{P}$ is chosen to be dominant on the main diagonal, i.e. each entry on the main diagonal is over 0.5.

Another method of applying PRAM is described in Willenborg and De Waal (2001) and is called invariant PRAM since it places the condition of invariance on the transition matrix $\mathbf{P}$, i.e. $\mathbf{tP} = \mathbf{t}$. This releases the users of the perturbed file of the extra effort to obtain unbiased moment estimates of the original data, since $\mathbf{t}^*$ itself will be an unbiased estimate of $\mathbf{t}$. Note that the property of invariance means that the expected values of the marginal distribution of the variable being perturbed are maintained. The invariance applies to the variable being perturbed, so to do a full invariant PRAM on several variables at once means that all of the variables would have to be compounded into a single variable, i.e. the variables are cross-classified. An example is given by Van den Hout and Elamir (see Chapter 6 in Van den Hout, 2004).

To obtain an invariant transition matrix, the following two stage algorithm given in Willenborg and De Waal (2001) is described below. Let $\mathbf{P}$ be any transition probability matrix: $p_{ik} = p(c^* = k \mid c = i)$ where $c$ represents the original category and $c^*$ represents the perturbed category. Now calculate the matrix $\mathbf{Q}$ using Bayes formula by $Q_{kj} = p(c = j \mid c^* = k) = \dfrac{p_{jk}\, p(c = j)}{\sum_l p_{lk}\, p(c = l)}$. We estimate the entries $Q_{kj}$ of this matrix by $\dfrac{p_{jk} v_j}{\sum_l p_{lk} v_l}$, where $v_j$ is the relative frequency of category $j$. For $\mathbf{R} = \mathbf{PQ}$ we obtain an invariant matrix where

$\mathbf{vR} = \mathbf{vPQ} = \mathbf{v}$ since $r_{ij} = \sum_k \frac{v_j p_{ik} p_{jk}}{\sum_l p_{lk} v_l}$ and $\sum_i v_i r_{ij} = \sum_k v_j p_{ik} = v_j$. The vector of the original frequencies $\mathbf{v}$ is the eigenvector of $\mathbf{R}$. In practice, $\mathbf{Q}$ can be calculated by transposing matrix $\mathbf{P}$, multiplying each column $j$ by $v_j$ and then normalizing its rows so that the sum of each row equals one. We define $\mathbf{R}^* = \alpha\mathbf{R} + (1-\alpha)\mathbf{I}$ where $\mathbf{I}$ is the identity matrix of the appropriate size. $\mathbf{R}^*$ is also invariant and the amount of perturbation is controlled by the value of $\alpha$.

In this article, the general method for invariant PRAM on a categorical variable having $L$ categories is as follows:

1. Choose the minimum diagonal entry for the $L \times L$ transition probability matrix **P**, $p_d$, and generate $L$ random numbers between $p_d$ and 1 to be placed on the main diagonal of $P$. Note that the probability on the main diagonal determines the amount of perturbation that will be carried out on the variable and it typically is over 80% in order to minimize information loss to the variable.
2. Divide $1 - p_d$ evenly among the other columns of the row in the $L \times L$ transition matrix **P**.
3. Calculate the invariant matrix **R** as described above. This will distort the original probabilities in the transition matrix, and in particular the diagonals will not necessarily meet the requirement of having a value between $p_d$ and 1.
4. Choose $\alpha$ for $\mathbf{R}^*$ that will bring the diagonals back to their approximate desired level. For instance, one can choose $\alpha$ so that the average value of the entries on the main diagonal of $\mathbf{R}^*$ equals the desired level.

For instance, assume a variable having four categories: $\mathbf{X}' = (25, 30, 50, 10)$. A typical transition probability matrix would be generated as follows with a minimal diagonal of 0.80:

$$\mathbf{P} = \begin{pmatrix} 0.8264 & 0.0579 & 0.0579 & 0.0579 \\ 0.0427 & 0.8718 & 0.0427 & 0.0427 \\ 0.0479 & 0.0479 & 0.8563 & 0.0479 \\ 0.0598 & 0.0598 & 0.0598 & 0.8207 \end{pmatrix}$$

Following the above algorithm, the invariant matrix **R*** with $\alpha = 0.5$ is as follows:

$$\mathbf{R}^* = \begin{pmatrix} 0.8478 & 0.0496 & 0.0740 & 0.0287 \\ 0.0413 & 0.8764 & 0.0598 & 0.0225 \\ 0.0370 & 0.0359 & 0.9058 & 0.0213 \\ 0.0716 & 0.0674 & 0.1067 & 0.7543 \end{pmatrix}$$

Note that $\mathbf{X}'\mathbf{R}^* = \mathbf{X}'$.

As shown above, invariant PRAM can be carried out so that the expected marginal distribution of the variable being perturbed is preserved. By using a more deterministic approach and selecting (without replacement) those records which will have their value of the variable transformed from category *i* to category *j* based on the probability $p_{ij}$, we can obtain the exact marginal distribution of the variable. This method can also be implemented as an SDC data masking technique for frequency tables where high utility is gained by preserving the exact totals and sub-totals of the table and only the internal cells of the table are perturbed. In this article we will not explore the possibilities of applying PRAM as an SDC masking technique for frequency tables any further.

PRAM is a generalization of other perturbative methods of disclosure control such as record swapping and delete/impute techniques. As in all perturbative SDC methods, joint distributions between perturbed and unperturbed variables will be distorted, in particular for variables that are highly correlated with each other. An initial analysis of the dependencies between the categorical variables can provide insight into which variables should be perturbed for SDC. In particular those variables that are highly dependent should be compounded and treated as a single variable in the perturbation process. As more perturbation is introduced, the utility of the data will be compromised. Variables that are typically perturbed are the demographic and geographic identifiers in the micro-data, and as mentioned in Section 2.4 these are typically used for statistical analysis as explanatory independent variables (e.g., regression models, ANOVA). Therefore, the perturbation of these variables will have an impact on the ability to make statistical inferences based on the perturbed micro-data.

### 3.2 Evaluation dataset for PRAM

The dataset that has been used for evaluating the SDC techniques for continuous data is less suited for evaluation of PRAM. For the evaluation of PRAM we have therefore used a file drawn from the 1995 Israel Census sample data which comprised 20% of all households in Israel. The dataset for this analysis contains 35,773 individuals aged 15 and over in 15,468 households across all geographical areas and household characteristics. For this analysis, we perturb the variable age. Age has 86 categories since the evaluation dataset includes only individuals aged 15 and over.

The edits involve the original edits from the data processing phase that check for inconsistencies with the variable under perturbation, age. The edits used for the evaluation dataset are:

$E_{PRAM,1}$ : {Under 16 and ever married}=Failure;
$E_{PRAM,2}$ : {Age of marriage under 14}=Failure;
$E_{PRAM,3}$ : {Age difference between spouse over 25}=Failure;
$E_{PRAM,4}$ : {Age of mother under 14}=Failure;
$E_{PRAM,5}$ : {Year of immigration less than year of birth}=Failure;
$E_{PRAM,6}$ : {Age of father under 14}=Failure;
$E_{PRAM,7}$ : {Under 16 and relation is spouse or parent}=Failure;
$E_{PRAM,8}$ : {Under 30 and relation is grandparent}=Failure;
$E_{PRAM,9}$ : {Under 16 and academic}=Failure;
$E_{PRAM,10}$ : {Under 16 and higher degree}=Failure;
$E_{PRAM,11}$ : {Age inconsistent with year of birth}=Failure.

In addition, since other variables may be changed in the post-editing imputation stage for correcting inconsistent records resulting from the perturbation, we add the following edits:

$E_{PRAM,12}$ : {Single and year of marriage not *null*}=Failure;
$E_{PRAM,13}$ : {Single and has spouse in household}=Failure;
$E_{PRAM,14}$ : {Relation is spouse and not married}=Failure.

The subscript "PRAM" indicates that these edits refer to the data set that is used to evaluate PRAM.

We need to ensure that not only are all records consistent in the final perturbed micro-data, but also that the usefulness of the data for statistical analysis is preserved by ensuring that the information loss measures do not fall below acceptable thresholds. Information loss measures were described in Section 2.4. We use the following distributions to assess information loss.

Hellinger Distance: We use the Hellinger Distance to measure the distortion to the distribution defined by district (27) × sex (2) × age (86) before and after PRAM. The smaller the Hellinger Distance, the less information loss.

Cramer’s *V*: We use the difference in Cramer’s *V* statistic on two dimensional tables where the rows contain the variable age (86) and the columns contain the following target variables: labour force characteristics (4) and years of education (26). We compare the Cramer’s V before and after the perturbation. The smaller the difference in Cramer’s *V*, the less information loss. Moreover, the sign of the difference is important since this tells us whether we are attenuating the association between variables.

Impact on $R^2$: The information loss is expressed as the ratio of the “between” variance *BV* for a target variable in groupings defined by the perturbed variable age compared to the “between” variance *BV* for a target variable in groupings defined by the original variable age. For this analysis we banded age into 9 groupings: 15-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-69, 70-74, and 75+. The target variables selected for this analysis are: percentage of academics, percentage belonging to the labour force and percentage unemployed out of those belonging to the labour force. An information loss measure below one indicates attenuation, i.e., the percentages in cells *i* defined by the age groupings are flattening towards the overall percentage of the distribution.

### 3.3 PRAM and edit constraints

If no controls are taken into account in the perturbation process, edit failures will occur resulting in inconsistent and “silly” combinations, such as married children, children earning income, or an unfeasible age difference between a child and parents. Methods need to be developed for implementing PRAM that will place controls on the perturbation process and will avoid as much as possible edit failures, reduce information loss and raise the overall utility of the data. The controls in the perturbation are defined by control variables which define groupings within which perturbations will be allowed. These control variables are typically highly correlated with the variable being perturbed and ensure a priori that failed edits and information loss will be minimal. The methods for controlling the perturbation are the following:

1. Before applying PRAM, the variable to be perturbed is divided into subgroups, $g = 1,...,G$. The transition (and invariant) probability matrix is developed for each subgroup *g*, $R_g$. The transition matrices for each subgroup are placed on the main diagonal of the overall final transition matrix where the off diagonal probabilities are all zero, i.e. the variable is only perturbed within the subgroup and the difference in the variable between the original value

and the perturbed value will not exceed a specified level. An example of this is perturbing age within broad age bands.

2. The variable to be perturbed may be highly correlated with other variables. Those variables should be compounded into one single variable. PRAM should be carried out on the compounded variable. Alternatively, the variable to be perturbed is carried out within subgroups defined by the second highly correlated variable. An example of this is when age is perturbed within groupings defined by marital status.

The control variables in the perturbation process will minimize the amount of edit failures, but they will not eliminate all edit failures, especially edit failures that are out of scope of the variables that are being perturbed. Remaining edit failures need to be manually or automatically corrected through imputation procedures depending on the types of edit failures and the amount.

We have applied a hot-deck imputation method for correcting inconsistent records and edit failures. This hot-deck imputation method was implemented by choosing a neighboring donor matching on control variables: district, number of persons in the household, marital status, sex and perturbed age. All variables that are included in the edits and are not control variables are imputed. The need for further imputation to satisfy edits means that more perturbation is introduced into the micro-data for other variables in the file interacting with the perturbed variable age. For example, the ages of the spouse and/or parents may also need to be changed as well as marital status. Therefore, the lower the number of overall edit failures resulting from the perturbation process, the less need for imputation to correct inconsistencies and the higher the utility maintained in the data. Section 3.4 presents results of the effectiveness of putting into place controls in the perturbation of the micro-data, thereby minimizing failed edits.

### 3.4 Results of PRAM on evaluation data

The perturbation of age by PRAM was carried out using an invariant transition probability matrix as described in Section 3.1. As mentioned, there are 86 categories of age in the evaluation data for individuals aged 15 and over. To perturb age we use the following methods:

1. Random perturbation across all ages, i.e. the transition probability matrix is of size $86 \times 86$, the diagonal $p_d$ is generated randomly and all other columns are given equal entries: $(1-p_d)/85$. The matrix is then made to be invariant and the diagonals controlled through the use of $\alpha$ as explained in Section 3.1.

2. Perturbation carried out within categories of marital status (4 categories – married, divorced, widowed and single), i.e. four separate invariant transition probability matrices are developed for perturbing age in each of the categories of marital status and the perturbation is carried out separately within each category. In other words, the final probability transition matrix is block diagonal containing the four matrices on the diagonals and all other parts of the transition probability matrix are zero.

3. Perturbation carried out on marital status (4 categories – married, divorced, widowed and single) × age bands (5 bands – 15-17, 18-24, 25-44, 45-64, 65-74 , 75+) as explained above.

4. Perturbation only allowed within broad age bands (9 bands – 15-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-69, 70-74, 75+) as explained above.

Because of the stochastic nature of the process, each method above results in a different number of records being perturbed. The number of perturbations for method 1 was 7,316 records. For methods, 2, 3, and 4, 6,822, 7,535, and 8,068 records were perturbed, respectively. Table 6 presents the number of records that failed the edits as presented in Section 3.2 after perturbing age according to the above methods. Note the large reduction in the number of edit failures as a result of placing controls on the perturbation processes. In particular, perturbing within narrow age bands (which is highly correlated with marital status) produced the best results.

[PLACE TABLE 6 AROUND HERE]

For each of the perturbation methods above, the edit failures were corrected using the hot-deck donor imputation method described in Section 3.3. In method 1, 37 records could not be imputed since no suitable donor was found so these records were unperturbed. In some cases, the control variables for the hot deck imputation had to be collapsed in order to be able to find a suitable donor for the failed record. After the imputation process, all records satisfy the edits. However, the information loss measures are also affected and we need to choose the method of perturbation that will minimize the information loss measures and obtain high utility data. Table 7 presents the results of the information loss measures as defined in Sections 2.4 and 3.2.

[PLACE TABLE 7 AROUND HERE]

It is shown in Table 7 that putting more controls in the perturbation process raises the level of the utility of the data. For example, the original value for Cramer's *V* which measures the association between labour force characteristics (employed, unemployed and out of the labour force) and age is 0.306. By perturbing the variable age, the measure of association decreases by 0.082 when age is perturbed across all possible ages, but only decreases by 0.008 when age is perturbed within narrow age bands. Note that all the information loss measures are negative based on the Cramer's *V* analysis. This indicates the attenuation of the target variables. In another example, we assume that the user is interested in carrying out an ANOVA analysis on the percentage of unemployed out of those belonging to the labour force using age groups as an explanatory variable. Before perturbing age, the value of the "between" variance *BV* was 8.8. However, when age is perturbed across all possible ages, the *BV* decreased by almost a half. This implies that the percentage of unemployed in each perturbed age grouping is tending towards the overall mean and we would obtain a lower $R^2$ as a result of the analysis. Figure 1 shows the shrinkage of the unemployment percentages within randomly perturbed age groups compared to the percentages within original age groups. Note that the unemployment percentages are much flatter across the randomly perturbed age groups. By contrast, there is only a minute change in the *BV* when age is perturbed within narrow age bands.

[PLACE FIGURE 1 AROUND HERE]

## 4. Discussion

In this article we have demonstrated how placing controls in the perturbation processes preserves the logical consistency of the records by minimizing micro edit failures. In addition, we focus also on minimizing information loss measures which are based on preserving the quality and utility of the data for statistical analysis and inference. While this article mainly discusses aspects of utility, quality and consistency, data suppliers and Statistical Agencies must also focus on minimizing disclosure risk. The trade off between managing the disclosure risk and ensuring high data utility must be carefully assessed before developing optimal SDC strategies. Future work will examine this trade off by measuring disclosure risk in micro-data before and after applying SDC methods (see: Elamir and Skinner, forthcoming; Skinner and Shlomo, 2005; Rinott and Shlomo, 2006 and references therein), and comparing the methods with respect to information loss and the preservation of edit constraints. By combining SDC methods and developing innovative methodologies for implementation, we can obtain consistent data, preserve totals, means and variance estimates,  and release statistical outputs with higher degrees of utility at little cost to the risk of disclosure.

We have applied relatively simple approaches to ensure that perturbed data satisfy the specified edits. More sophisticated methods for ensuring that variables satisfy edits are available from the area of statistical data editing and the area of imputation. For instance, the Fellegi-Holt principle of minimum change (Fellegi and Holt, 1976) can be applied. This principle determines that the data of an inconsistent record should be made to satisfy all edits by changing the fewest possible number of values. When applying the Fellegi-Holt principle, one first identifies the erroneous fields. These erroneous fields can subsequently be imputed by an imputation method. In a last step, the imputed values can be adjusted so all edits become satisfied. An algorithm for implementing the Fellegi-Holt principle for both categorical and continuous data is based on a branch-and-bound search (De Waal and Quere, 2003). Several alternative approaches and a method to adjust imputed fields so all edits become satisfied are described by De Waal (2003). Another approach, called NIM (Nearest-Neighbor Imputation Method) which is implemented in Statistic's Canada CANCEIS, has been successfully carried out for Canadian Censuses (Bankier, 1999). This approach implements a minimum change principle similar to Fellegi-Holt principle. Namely, the data in a record are made to satisfy all edits by changing the fewest possible number of values given the available potential donor records. Intuitively, using the Fellegi-Holt principle or the NIM approach leads to results that are closer to optimality than using the relatively simple method for ensuring consistencies that we have used. Our intuition remains to be confirmed by future work.

With respect to the evaluation of information loss owing to the application of SDC methods much more research remains to be done. An aspect of information loss that requires more attention is, for instance, the effect of SDC methods on regression parameters. Some work on this aspect has been carried out by Van den Hout and Kooiman (2006).

Based on a given threshold for disclosure risk, the "best" method to protect a micro-data set is hard to determine in general. For a particular micro-data set the "best" SDC method depends on the intended uses of the data by the users, the willingness of the statistical agency to disseminate this data set, the legal aspects of releasing these data, and on the structure of the data. For instance, homogeneous data require different SDC techniques than heterogeneous data. To some extent the "best" SDC method for a micro-data set will always be a subjective choice. Levels of protection and tolerable disclosure risk thresholds vary from country to country and depend on the different modes for accessing the micro-data. A prerequisite however for making a well-founded choice of SDC method is a solid understanding of a wide range of SDC methods. We hope that this article helps to improve our understanding of several of such SDC methods.

## References

Bankier, M. (1999) Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses. *U.N. Economic Commission for Europe Work Session on Statistical Data Editing*, Rome.

Brand, R. (2002) Micro-data Protection Through Noise Addition. In *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), 97-116. New York: Springer.

De Waal, T. (2003) *Processing of Erroneous and Unsafe Data*. Ph.D. Thesis, Erasmus University, Rotterdam.

De Waal, T. and R. Quere (2003) A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics, 19*, 383-402.

Elamir, E. and Skinner, C.J. (forthcoming) Record-Level Measures of Disclosure Risk for Survey Micro-data. *Journal of Official Statistics* (see: http://eprints.soton.ac.uk/8175/01/s3ri-workingpaper-m04-02.pdf )

Fuller, W. A. (1993) Masking Procedures for Micro-data Disclosure Limitation. *Journal of Official Statistics, 9*, 383-406.

Fellegi, I.P. and Holt, D. (1976) A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association, 71*, 17-35.

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

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998) Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, *14*, 463-478.

Gross, B., Guiblin, P. and Merrett, K. (2004) Implementing the Post-Randomisation Method to the Individual Sample of Anonymised Records (SAR) from the 2001 Census. http://www.ccsr.ac.uk/sars/events/2004-09-30/gross.pdf.

Kim, J.J. (1986) A Method for Limiting Disclosure in Micro-data Based on Random Noise and Transformation. American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 370-374.

Rinott, Y. and Shlomo, N (2006) A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation. In *PSD'2006 Privacy in Statistical Databases*, Springer LNCS proceedings, to appear.

Salazar-González, J. J., Bycroft, C. and Staggemeier, A. T. (2005) Controlled Rounding Implementation. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Geneva.

Särndal, C. E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse.* Chichester: John Wiley & Sons.

Shlomo, N. and De Waal, T. (2005) Preserving Edits When Perturbing Micro-data for Statistical Disclosure Control. *Statistical Journal of the United Nations ECE*, *22*, 173-185.

Shlomo, N and Young, C. (2006) Statistical Disclosure Control Methods Through a Risk-Utility Framework. PSD'2006 Privacy in Statistical Databases, Springer LNCS proceedings, to appear.

Skinner C. J. and Shlomo, N. (2005) Assessing Disclosure Risk in Micro-data Using Record Level Measures. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Geneva.

Van den Hout, A. (2004) *Analyzing Misclassified Data: Randomized Response and Post Randomization*. Ph.D. Thesis, University of Utrecht, Utrecht.

Van den Hout, A. and Kooiman, P. (2006) Estimating the Linear Regression Model with Categorical Covariates subject to Randomized Response. *Computational Statistic & Data Analysis*, *50*, 3311-3323.

Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

Willenborg, L. and Van den Hout, H. (2006) Peruco: A Method for Producing Safe and Consistent Micro-data. *International Statistical Review*, *74*, 271-284.

Winkler, W. E. (2002) Single Ranking Micro-aggregation and Re-identification. Statistical Research Division report RR 2002/08, at http://www.census.gov/srd/www/byyear.html.

Yancey, W.E., Winkler, W.E., and Creecy, R. H. (2002) Disclosure Risk Assessment in Perturbative Micro-data Protection. In: *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 135-151. New York: Springer.

*Table 1. Standard deviation (STD)  at different stages of micro-aggregation and additive random noise for variable* tax

| | Quintile 1 | Quintile 2 | Quintile 3 | Quintile 4 | Quintile 5 | Total |
|---|---|---|---|---|---|---|
| STD of *tax* | 79 | 149 | 253 | 555 | 2,998 | 2,119 |
| STD of micro-aggregated *tax* | 61 | 122 | 220 | 502 | 2,864 | 2,082 |
| STD for generating random noise[(*)] | 50 | 86 | 125 | 236 | 835 | 394 |
| STD of micro-aggregated *tax* with random noise | 78 | 149 | 252 | 552 | 2,981 | 2,126 |

[(*)] the value 50 in the cell defined by "STD for generating random noise" and Quintile 1 is obtained by taking the variance of *tax* ($79 \times 79$) minus the variance of the micro-aggregated *tax* ($61 \times 61$).

*Table 2. Number of individuals with an absolute difference (Diff) between the perturbed variable* gross *and the sum of perturbed variables* net *and* tax *based on micro-aggregation and additive noise*

| Diff | Number of Individuals |
|---|---|
| Total | 16,232 |
| No Difference | 641 |
| $1 < \text{Diff} \leq 10$ | 677 |
| $10 < \text{Diff} \leq 50$ | 2,859 |
| $50 < \text{Diff} \leq 100$ | 2,966 |
| $100 < \text{Diff} \leq 500$ | 6,239 |
| $\text{Diff} > 500$ | 2,850 |

*Table 3. Standard deviation (STD) at different stages of micro-aggregation for two procedures: adding random noise with linear programming and adding correlated random noise*

| Variable | STD Original Variable | STD Micro-aggregated Variable | Procedure 1 | | Procedure 2 |
|---|---|---|---|---|---|
| | | | STD Micro-aggregated Variable with Random Noise | STD Micro-aggregated Variable with Random Noise and Linear Programming | STD Micro-aggregated Variable with Correlated Random Noise |
| *tax* | 2,119 | 2,082 | 2,115 | 2,103 | 2,091 |
| *net* | 5,137 | 5,114 | 5,134 | 5,129 | 5,119 |
| *gross* (=*net*+*tax*) | 7,181 | 7,174 | 7,174 | 7,174 | 7,171 |

*Table 4. Results of the random rounding (RR) with and without controls and the re-shuffling algorithm on the totals of rounded variables* net, tax *and* gross

| Variable | True Total | RR - no controls on totals and no additivity | Difference | RR - controls on totals and additivity but not all within the base | Difference | RR - controls on totals and additivity and all within the base | Difference |
|---|---|---|---|---|---|---|---|
| *tax* | 25,443,623 | 25,444,410 | -787 | 25,443,630 | -7 | 25,443,710 | -87 |
| *net* | 86,724,755 | 86,725,330 | -575 | 86,724,770 | -15 | 86,724,860 | -105 |
| *gross* (=*net*+*tax*) | 112,168,378 | 112,169,740 | -1,362 | 112,168,400 | -22 | 112,168,570 | -192 |

*Table 5. Information loss measures for the joint distribution of age group, sex and gross income*

| | Groupings of 10 | Groupings of 20 |
|---|---|---|
| Number and Percent of Cells with Differences (out of 616 possible combinations) | 106 (22%) | 166 (34%) |
| Hellinger's Distance:<br>age groups × sex × income groups | 0.011 | 0.013 |
| Difference in Cramer's *V*:<br>income groups and age groups × sex<br>$V(T_{orig}) = 0.1300$ | 0 | -0.0001 |
| Ratio of *BV*:<br>mean of gross within<br>age groups × sex<br>$BV_{orig} = 3.83\times 10^9$ | 1.004 | 0.998 |

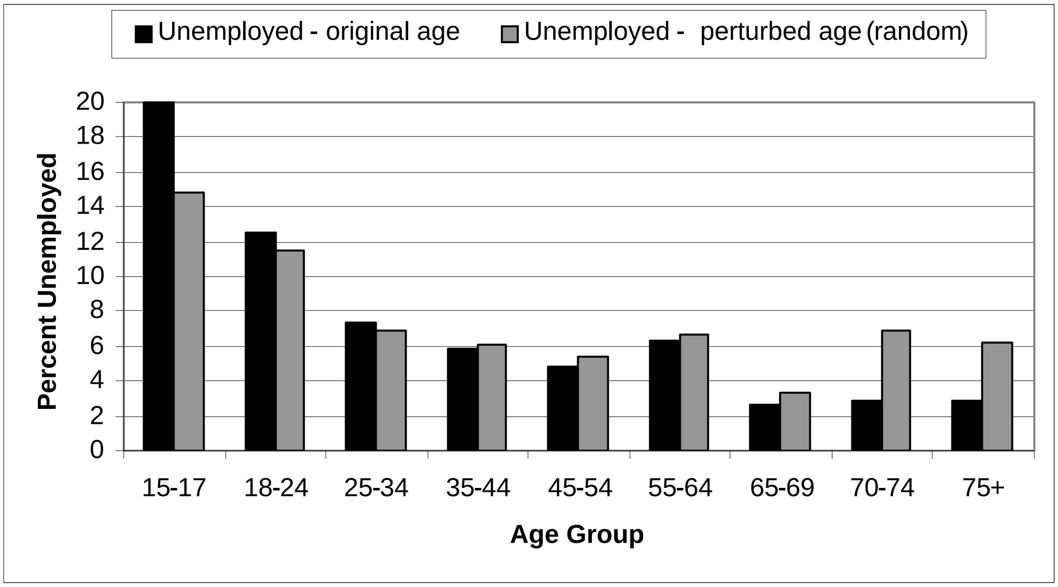
*Table 6. Number of records failing edits according to the method of perturbation*

| | Method of Perturbation | | | |
|---|---|---|---|---|
| | Random | Within Marital Status | Within Marital Status and Broad Age Groups | Within Narrow Age Groups |
| No edit failures | 31,983 | 33,143 | 35,023 | 35,440 |
| 1 edit failure | 2,344 | 1,827 | 731 | 328 |
| 2 edit failures | 1,303 | 800 | 19 | 5 |
| 3 edit failures | 59 | 3 | 0 | 0 |
| 4+ edit failures | 84 | 0 | 0 | 0 |

*Table 7. Results of information loss measures according to perturbation method*

| information loss measures | | Method of Perturbation | | | |
|---|---|---|---|---|---|
| | | Random | Within Marital Status | Within Marital Status and Broad Age Groups | Within Narrow Age Groups |
| Hellinger Distance | District*sex*age | 0.0995 | 0.0913 | 0.0844 | 0.0895 |
| Difference in Cramer's *V* | Years of Education and Perturbed Age<br>$V(T_{orig}) = 0.146$ | -0.0091 | -0.0099 | -0.0046 | -0.0037 |
| | Labour Force Characteristics and Perturbed Age<br>$V(T_{orig}) = 0.306$ | -0.0816 | -0.0686 | -0.0106 | -0.0076 |
| Ratio of Between Variance | Percent Academics Within Perturbed Age Groupings<br>$BV_{orig} = 19.9$ | 0.838 | 0.815 | 0.969 | 1.001 |
| | Percent in Labour Force Within Perturbed Age Groupings<br>$BV_{orig} = 270.5$ | 0.513 | 0.580 | 0.967 | 0.996 |
| | Percent Unemployed Within Perturbed Age Groupings<br>$BV_{orig} = 8.8$ | 0.486 | 0.557 | 0.982 | 0.998 |

*Figure 1: Percent unemployed according to original age groups and perturbed age groups*


■ Unemployed - original age
▨ Unemployed - perturbed age (random)
Percent Unemployed
Age Group