PRESERVING EDITS WHEN PERTURBING MICRODATA FOR STATISTICAL DISCLOSURE CONTROL

NTALIE SHLOMO, TON DE WAAL

ABSTRACT

To protect individuals in microdata from the risk of re-identification, a general perturbative method called PRAM (the Post-Randomization Method) is sometimes used for masking records. This method adds "noise" to categorical variables by changing values of categories for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. Changing values of categorical variables, however, will cause fully edited and clean records in microdata to start failing edit constraints resulting in data of low utility. In addition, an inconsistent record pinpoints to a potential attacker that the record was perturbed and attempts can be made to unmask the data. Therefore, the perturbation process must take into account micro edit constraints which will ensure that perturbed microdata satisfy all edits. Macro edit constraints which take the form of information loss measures also need to be defined in order to ensure that the overall utility of the data will not be badly compromised given an acceptable level of disclosure risk. This paper will discuss methods for perturbing microdata using PRAM while minimizing micro and macro edit failures.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/12

University of Southampton

Preserving Edits When Perturbing Microdata for Statistical Disclosure Control

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

[a]Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom

Tel: +44 23 8059 5732 Fax: +44 23 8059 3846 e-mail: n.shlomo@soton.ac.uk

[b]Statistics Netherlands, PO Box 4000, 2270 JM Voorburg, Netherlands

Tel: +31 70 337 4930 Fax: +31 70 387 7429 e-mail: twal@cbs.nl

Abstract. To protect individuals in microdata from the risk of re-identification, a general perturbative method called PRAM (the Post-Randomization Method) is sometimes used for masking records. This method adds "noise" to categorical variables by changing values of categories for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. Changing values of categorical variables, however, will cause fully edited and logical records in microdata to start failing edit constraints (*i.e.*, logical rules) resulting in data of low utility. Also, an inconsistent record will target the record as having been perturbed for disclosure control and attempts can be made to unmask the data. Therefore, the perturbation process must take into account per-record micro edit constraints through post-editing which will ensure that perturbed microdata satisfy all edits. In addition, file-level macro edit constraints, which take the form of information loss measures, are also defined in order to ensure that the overall utility of the data will not be badly compromised given an acceptable level of disclosure risk. This paper will discuss methods for perturbing microdata using PRAM while minimizing micro and macro edit failures.

Keywords. Post-randomization method, Statistical disclosure control, Disclosure risk, Information loss, Post-editing, Imputation, Microdata

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 microdata. Tabular data can be classified into tables containing frequency counts and tables containing aggregated data. Microdata can be seen as special cases of frequency tables where the cell count is one. Whereas tables have been released traditionally by Statistical Agencies, microdata sets released to researchers is a relatively new phenomenon. The dissemination of microdata creates non-trivial SDC-problems that remain to be solved, in particular how to assess disclosure risk based on realistic scenarios, how to measure the quality and utility of microdata that has undergone masking techniques and what is the optimum balance between minimizing the disclosure risk and maximizing the utility of the microdata.

As 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 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. Masking techniques on the microdata include perturbative methods which alter the data (*e.g.*, adding random noise to variables, data swapping, *etc*.) and non-perturbative methods which preserves the integrity of the data (*e.g.*, global recoding, suppression, sub-sampling, *etc*.). Sensitive information about an individual respondent might be disclosed if the respondent were re-identified by an attacker. Many SDC methods for protecting microdata therefore aim to prevent re-identification by a potential attacker.

A general perturbative method for masking records in microdata against the risk of re-identification is PRAM (Post-Randomization Method) for categorical variables (Gouweleeuw *et al*, 1998). PRAM 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 Anonymised Records (SARs) of the 2001 UK Census (Gross *et al*, 2004).

Changing the categories of variables will cause records in microdata, *i.e.* the data of individual respondents, to fail certain logical rules, called edit constraints or edits. Data collected by Statistical Agencies generally contain errors. In order to be able to publish reliable statistical information these errors have to be corrected. This correction process is referred to as statistical data editing. At Statistical Agencies, edit constraints are often used to determine whether a record is consistent or not. An example of such an edit is that the age of a married person must be over a minimum age as required by law. Inconsistent records, *i.e.* records that fail at least one edit, are considered to contain errors, while consistent records, *i.e.* records that satisfy all edits, are generally considered error-free.

For academic statisticians the emphasis on consistent data, *i.e.* the wish of Statistical Agencies to let the data satisfy specified edit constraints, 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.*

In general, microdata that have been through all phases of data processing, including editing and imputation, will be error-free. However, the perturbation process for SDC on logical and fully edited records will result in records again failing edit constraints since inconsistencies will reoccur between the perturbed and original variables. In particular, PRAM purposely introduces misclassification into the microdata which will cause perturbed records to fail edit constraints. Therefore, there is a need to develop the PRAM procedure which will simultaneously take into account edit constraints and ensure that the resulting perturbed microdata satisfy all edits. Although users of perturbed microdata are aware that certain variables in records have been misclassified, it is not advisable to release microdata with records that fail edits since this damages the utility of the data. In addition, an inconsistent and illogical record will immediately target the perturbed record and attempts can be made to unmask it. This is particularly problematic when microdata contain hierarchical data (*i.e.*, households and persons) and unperturbed variables can be used to identify perturbed variables and their original content. For example, the size of the household may be perturbed, yet the size of the household can be determined by the number of records in each household.

Statistical data typically undergo extensive edit and imputation at the data collection and data processing stages. Original edit constraints that are used for checking the data include edits for missing data, out-of-scope responses and faulty skip patterns in the questionnaire. These original edit constraints do not have to be re-checked after perturbing the microdata since the perturbation scheme will not create these types of edit failures. However, original edit constraints that check for illogical records involving interactions of variables with perturbed variables need to be re-examined at the post-editing stage. In addition, in order to correct inconsistent records, imputation procedures are implemented and other variables may be changed in the record. Therefore, original edit constraints involving imputed variables also need to be re-examined at the post-editing stage. Finally, new edit constraints need to be added which check the logical

consistency for all derived variables or a decision can be made to automatically recalculate all derived variables after the perturbation process.

As mentioned, records that fail edit constraints as a result of the perturbation need to undergo imputation procedures to correct inconsistencies. In this paper we will implement a hot-deck imputation method for correcting inconsistent records. Potential donors are found that have passed all edit constraints and also match on perturbed variables and other control variables. The record nearest to the failed record is chosen as a donor, and variables are transferred onto the recipient record until the failed record passes all edit constraints. The edit and imputation for perturbed microdata is more complex than edit and imputation carried out at the data processing stage since in addition to the control variables, perturbed variables must also remain fixed, while other variables need to be changed in order to obtain a logical and consistent record.

The goal when developing optimal SDC strategies is to assess and minimize disclosure risk while maintaining high utility data. In Gomatam *et al*, 2003, SDC methods are presented in a decision problem framework based on a disclosure risk – data utility assessment of the microdata. The decision problem finds the balance between the need for protection against the risk of re-identification and the amount of information loss incurred by the data masking techniques. The optimum trade-off is determined through the use of quantitative measures of disclosure risk and data utility. Given the same levels of disclosure risk, we need to find ways of obtaining higher utility data. One way of doing this when perturbing microdata with PRAM is to put more controls into the perturbation process. This causes less micro edit failures and therefore less imputation is needed to correct inconsistencies in the data. In addition, macro editing constraints which alert the data protector to loss of information beyond acceptable thresholds need to be taken into account. Macro editing constraints typically contain measures for data distortion in marginal and joint distributions (*e.g.*, Hellinger Distance, entropy, *etc*.) and the impact on various statistics that are used for statistical inferences (*e.g.*, R^2, χ^2, *etc*.).

We will illustrate the problem and method sketched so far by means of an example. Suppose, for instance, that a microdata set containing a sample of the participants of the UN/ECE Work Session on Statistical Data Editing held in May 2005 in Ottawa (where the present article was presented for the first time) were released. Suppose furthermore that the microdata set contains information on the affiliation of authors and their co-authors, and sensitive information on, for instance, the criminal past of the authors. Now consider the record: "Affiliation author = Statistics Netherlands", "Affiliation co-author = University of Southampton", "Criminal past = has stolen candy". At the UN/ECE Work Session on Statistical Data Editing in Ottawa 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 he has stolen candy. We therefore apply PRAM to protect our example microdata set. The record "Affiliation author = Statistics Netherlands", "Affiliation co-author = University of Southampton", "Criminal past = has stolen candy" might then be modified into a record "Affiliation author = Statistics Netherlands", "Affiliation co-author = *Statistics Canada*", "Criminal past = has stolen candy". However, at the UN/ECE Work Session on Statistical Data Editing in Ottawa there was no author from Statistics Netherlands with a co-author from Statistics Canada. 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 Netherlands", "Affiliation co-author = *Statistics Canada*", "Criminal past = has stolen candy" is, now further processed by imputing values for the non-perturbed data in such a way that a feasible record results. Suppose that we impute "Affiliation author" and obtain a record "Affiliation author = *Statistics Canada*", "Affiliation co-author = *Statistics Canada*", "Criminal past = has stolen candy". 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 mis-used to falsely deduce that a specific author from Statistics Canada has ever stolen candy.

The paper will be developed as follows. Section 2 describes the PRAM methodology for perturbing categorical variables in microdata (note that this method can also be used to perturb frequency tables and maintain marginal totals). Section 3 describes the evaluation dataset, including the corresponding micro edits constraints, that will be used to demonstrate the perturbation method and the analysis. Section 4 presents the algorithm for implementing PRAM under various methods of controlling variables in order to minimize edit failures and maximize data utility. Section 5 defines the macro edit constraints which will be used on the evaluation dataset. Section 6 presents results of the algorithm and the impact on the edit constraints. Section 7 discusses the trade-off between data utility and the disclosure risk of re-identification. Finally, Section 8 contains a short discussion.

2. PRAM (Post-Randomization Method)

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. 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: $p_{ij} = p(\textit{perturbed category is } j \,|\, \textit{original category is } i)$. 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 microdata 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_{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 $E(\mathbf{t}^* \,|\, \mathbf{t}) = \mathbf{tP}$. 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 (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 $r_{ij} = \sum_{k} \dfrac{v_j p_{ik} p_{jk}}{\sum_{l} p_{lk} v_l}$ 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 α.

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

- 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.
- Divide $1 - p_d$ evenly among the other columns of the row in the $L \times L$ transition matrix **P**.
- 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.
- Choose α for $\mathbf{R}^*$ that will bring the diagonals back to their approximate desired level. For instance, one can choose α 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:

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

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

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

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 placing controls in the perturbation process, we can also 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 paper we will not explore the possibilities of applying PRAM as an SDC masking technique for frequency tables any further.

PRAM is a perturbative method of disclosure control and therefore will distort important joint distributions between perturbed and unperturbed variables, 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 microdata, and these are generally 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 microdata.

3. Evaluation Dataset and Micro Edit Constraints

The method that is described in this paper for preserving micro and macro edit constraints when perturbing microdata for SDC and the resulting analysis of the algorithm are demonstrated on a file drawn from the 1995 Israel Census sample data which comprised 20% of all households in Israel. The evaluation 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 will perturb the variable age. Age has 86 categories since the evaluation dataset includes only individuals aged 15 and over.

The micro edit constraints involve the original edits from the data processing phase of the microdata that check for inconsistencies based on the variable being perturbed, age. The micro edit constraints used for the evaluation dataset are:

E_1 = {Under 16 and ever married}=Failure

E_2 = {Age of marriage under 14}=Failure

E_3 = {Age difference between spouse over 25}=Failure

E_4 = {Age of mother under 14}=Failure

E_5 = {Year of immigration less than year of birth}=Failure

E_6 = {Age of father under 14}=Failure

E_7 = {Under 16 and relation is spouse or parent}=Failure

E_8 = {Under 30 and relation is grandparent}=Failure

E_9 = {Under 16 and academic}=Failure

E_{10} = {Under 16 and higher degree}=Failure

E_{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 micro edit constraints:

E_{12} = {Single and year of marriage not *null*}=Failure

E_{13} = {Single and has spouse in household}=Failure

E_{14} = {Relation is spouse and not married}=Failure

4. Methods of Perturbation and Preserving Edit Constraints

In our evaluation dataset, the variable age containing 86 categories is randomly perturbed using PRAM as explained in Section 2. If no controls are taken into account in the perturbation process, many 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 micro and macro edit failures 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 micro and macro edit constraints will be minimal. The methods for controlling the perturbation are the following:

- 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.
- 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.

To flag failed edits, appropriate editing software needs to be developed and implemented both before and after the perturbations as an integral part of the perturbation software. All programs used for our evaluation study were written in SAS.

The control variables in the perturbation process will minimize the amount of micro 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 micro 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 edit constraints and are not control variables are imputed. The need for further imputation to satisfy micro editing constraints means that more perturbation is introduced into the microdata 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 micro 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 6 presents results of the effectiveness of putting into place controls in the perturbation of the microdata, thereby minimizing failed micro editing constraints.

5. Macro Edit Constraints

In this section we define macro edit constraints which will serve as overall measures of data utility. We need to ensure that not only are all records consistent in the final perturbed microdata, but also that the usefulness of the data for statistical analysis is preserved by ensuring that the macro edit constraints do not fall below acceptable thresholds.

Macro edit constraints take several forms: distance metrics that measure the distortion to joint distributions with the perturbed variable age; weaknesses in measures of association between target variables and the perturbed variable age; and the amount of shrinkage towards a common mean as expressed by the "between" variance of target variables within perturbed age groups. Note that this latter measure reflects the reduction of the measure R^2 for regression analysis. Some of these measures are described in Gomatam *et al* (2003).

Let *D* represent the dataset of size *n* and let *D(c)* be the cell frequency associated with a cell *c* for a distribution having $c = 1,...,C$ cells. The macro edit constraints used for this evaluation dataset are based on the following information loss measures.

- Hellinger Distance: $HD(D_{orig}, D_{pert}) = \frac{1}{\sqrt{2n}} \sqrt{\sum_c \left(\sqrt{D_{orig}(c)} - \sqrt{D_{pert}(c)}\right)^2}$. The Hellinger Distance 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 district × sex × age before and after PRAM. The smaller the Hellinger Distance, the less information loss.

- Cramer's V: Let: $V_{1,2} = \sqrt{\frac{\chi^2}{N \times \min((C_1 - 1),(C_2 - 1))}}$ where χ^2 is the standard test statistic for independence on two variables having C_1 and C_2 number of cells. Cramer's V lies between 0 for no association and 1 for full association. The measure that defines the loss in the association between two variables is $CV_{i,j}(D_{orig}, D_{pert}) = V_{i,j}(D_{orig}) - V_{i,j}(D_{pert})$. We use the reduction in Cramer's V statistic on joint distributions between the perturbed variable age to other target variables not perturbed: labor force characteristics and years of education. The smaller the difference in Cramer's V, the less information loss.

- Impact on R^2: The categorical variables that are perturbed are generally used in the analysis of variance on target variables or as covariates in regression models where the goodness of fit is expressed in the measure R^2. For a univariate analysis of variance, the total sum of squares, *SST*, of a numerical target variable can be broken down into two components: the "between" sum of squares, *SSB*, which measures the variance of the target variable between groupings and the "within" sum of squares, *SSW*, which measures the variance of a target variable within the groupings . Let *m* define the number of groups based on a categorical variable, each group of size n_i (i=1,…,m). The variance components are:

SST	$\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\overline{X})^2$	*n*-1 degrees of freedom
SSB	$\sum_{i=1}^{m}(\overline{X}_i-\overline{X})^2$	*m*-1 degrees of freedom
SSW	$\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\overline{X}_i)^2$	*n*-*m* degrees of freedom

where X_{ij} is the value of the target variable for the *j*-th unit in the *i*-th group (i=1,…,m), $\overline{X}_i$ is the mean value of the target variable in the *i*-th group, and $\overline{X}$ is the overall mean of the target variable.

R^2 is the ratio of *SSB* to *SST*. By perturbing the categorical variable age, the groupings lose their homogeneity: *SSB* becomes smaller, and *SSW* becomes larger. In other words, the averages within each grouping are shrinking towards the overall mean. The information loss is expressed as the ratio of the "between" *SSB* for a target variable in groupings defined by the perturbed variable age compared to the "between" *SSB* for a target variable in groupings defined by the original variable age. The target variables selected for this analysis are: percent of academics, percent belonging to the labor force and percent unemployed out of those belonging to the labor

force within groupings defined by age. The larger the ratio of *SSB*'s between the perturbed and original age groupings, the less information loss.

6. Results on Evaluation Data

The perturbation of age by PRAM was carried out using an invariant transition probability matrix as described in Section 2. 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 86×86, 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 α as explained in Section 2.

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 1 presents the number of records that

failed the micro edit constraints as presented in Section 3 after perturbing age according to the above methods. Note the large reduction in the number of micro 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 1 AROUND HERE]

For each of the perturbation methods above, the edit failures were corrected using the hot-deck donor imputation method described in Section 4. 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 micro edit constraints. However, the macro editing constraints are also affected and we need to choose the method of perturbation that will minimize the macro edit constraints and obtain high utility data. Table 2 presents the results of the macro editing constraints as defined in Section 5.

[PLACE TABLE 2 AROUND HERE]

It is shown in Table 2 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 labor force characteristics (employed, unemployed and out of the labor 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. 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 labor force using age as an explanatory variable. Before perturbing age, the value of *SSB* between age groupings is 70.5. However, when age is perturbed across all possible ages, *SSB* decreases 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 *SSB* when age is perturbed within narrow age bands.

[PLACE FIGURE 1 AROUND HERE]

7. Disclosure Risk – Data Utility Trade-Off

As mentioned in Section 1, the decision problem framework for SDC methods is to minimize the disclosure risk while maximizing the utility of the data. Higher utility data however, as shown in Table 2, typically comes at the expense of increased disclosure risk in the microdata. Therefore, the disclosure risk of the perturbed microdata needs to be assessed in order to ensure that the risk of re-identification is not greatly increased as a result of placing controls in the perturbation process.

Assessing the risk of re-identification is problematic for microdata containing samples where the full characteristics of the population are unknown. In this case, we need to use probabilistic modeling to estimate the disclosure risk. However, in this paper we are perturbing census data and therefore the population is fully known and quantitative measures for the risk of re-identification can be calculated. Disclosure risk measures depend on disclosure risk scenarios which determine potential attacks on the microdata through the availability of public use files and software tools. The disclosure risk scenario determines a set of identifying key variables which when compounded together form a key that can be used to make a re-identification of an individual. Typical disclosure risk measures based on the key include the number of population uniques and the expected number of correct matches were the microdata matched back to the population. For example, suppose one record is chosen in the microdata having a value *k* of the key. The chance of a correct match to the population would be $1/F_k$ where F_k is the size of the population for the value *k* of the key. Disclosure risk measures, however, are not so straightforward when using perturbative methods of disclosure control since we need to take into account the level of

uncertainty introduced into the microdata which decreases the potential of an attacker to make a correct re-identification.

We define the identifying key as: district (27) × sex (2) × original age (86) × marital status (4). There are 5,556 non-zero cells for this key in the evaluation dataset. We use the following quantitative measures for assessing the risk of re-identification in the microdata:

- Percent unperturbed records in small cells of the key: Out of the 5,556 non-zero cells in the key, 2,672 cells have a count of one or two. These cells contain 3,659 records. We calculate the percentage of those records that were not perturbed in any way. The higher this percentage, the higher the disclosure risk.

- Expected number of correct matches: Because of the perturbation that was introduced, the chance of choosing an unperturbed record in which to carry out a matching procedure is defined by the entries on the diagonal of the probability transition matrix used for PRAM, p_d. Therefore, the probability of a correct match to the population as explained above for a record having a probability p_d of not being perturbed is p_d / F_k . Summing up these probabilities for all records in the microdata gives the overall expected number of correct matches. The higher this expected number of correct matches, the higher the disclosure risk.

- Probabilistic record linkage: Setting up a probabilistic record linkage framework for assessing disclosure risk is out of the scope of this paper. However, since only the variable age was perturbed, it is likely that the closer the perturbed value is to the original value, the higher the probability of obtaining a correct link in a probabilistic record linkage procedure. This measure calculates the proportion of perturbed records that resulted in the variable age being perturbed within a 5 year age difference. The higher the percentage, the more likely to obtain a correct link and the higher the disclosure risk.

Table 3 contains the results of these disclosure risk measures. Since the first two measures depend on whether the value was perturbed or not, all methods of perturbation contain approximately the same levels of disclosure risk. The third disclosure risk measure is much higher for the method of perturbing age within narrow age bands since the variable age is a priori perturbed within 5 or 10 year age bands. The data protector must weigh this increased disclosure risk against the benefits of obtaining much higher utility in the data using the controls in the perturbation scheme as seen in Table 2.

[PLACE TABLE 3 AROUND HERE]

8. Discussion

In this paper, we explained how edit constraints can be taken into account when applying PRAM. In particular, we showed how placing controls in the perturbation procedure will raise the overall utility of the data by minimizing the number of micro and macro edit failures. Depending on the measure for assessing the disclosure risk, the risk of re-identification may be increased by limiting the perturbations within narrow ranges of possible values. In general, Statistical Agencies have to set thresholds to find the optimal balance between acceptable levels of disclosure risk and high utility data based on policies and protocols governing the release of microdata. It should be noted, however, that protecting microdata solely by PRAM leaves high disclosure risk in the microdata and therefore PRAM should be combined with other non-perturbative methods of disclosure control such as global recoding which would lower the disclosure risk and still preserve the integrity of the data.

Standard hot-deck imputation methods are typically used for correcting categorical data in social surveys and censuses. In this paper we have adopted this approach as well and have used a standard hot-deck imputation method for correcting perturbed data failing micro edit constraints at the post-editing stage. More sophisticated methods for imputing variables which, for instance, follow the Fellegi-Holt principle of minimum change (Fellegi and Holt, 1976) can be applied. The Fellegi-Holt 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 categorical 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 a standard hot-deck imputation method. Our intuition remains to be confirmed by future work.

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*, www.unece.org/stats/documents/1999/06/sde/24.e.pdf.

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*, pp. 383-402.

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

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

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

Gross, B., P. Guiblin and K. Merrett (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

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

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

Table 1: Number of Records Failing Micro Edit Constraints 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	1,303	800	19	5
3	59	3	0	0
4+ edit failures	84	0	0	0

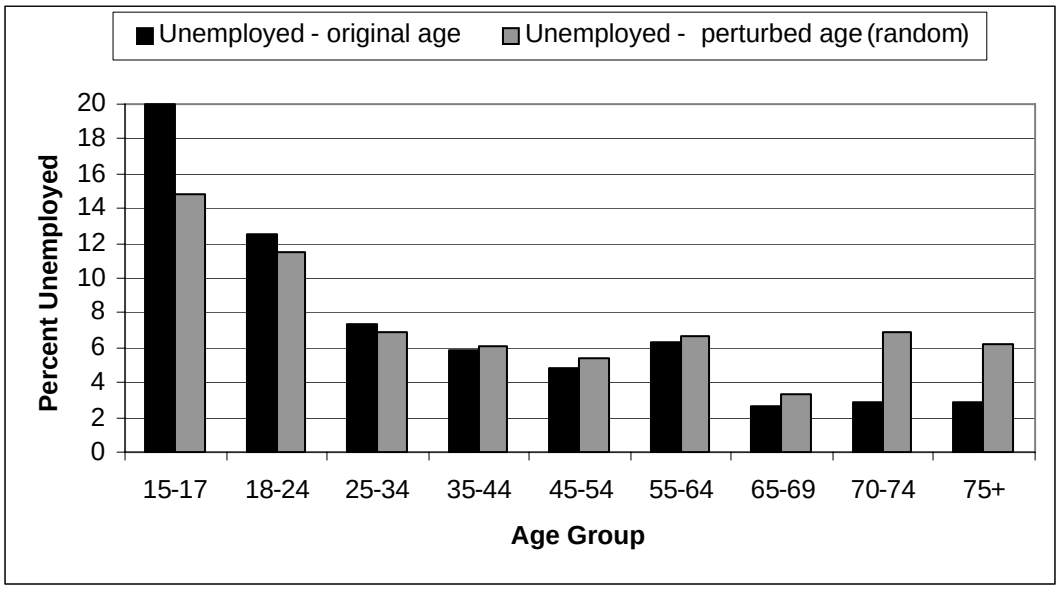
Table 2: Results of Macro Edit Constraints According to Perturbation Method

Macro-Edit Constraints		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 $V(D_{orig}) = 0.146$	0.0091	0.0099	0.0046	0.0037
	Labor Force Characteristics and Perturbed Age $V(D_{orig}) = 0.306$	0.0816	0.0686	0.0106	0.0076
Ratio of Between Variance	Percent Academics Within Perturbed Age Groupings $SSB(D_{orig}) = 159.5$	0.838	0.815	0.969	1.001
	Percent in Labor Force Within Perturbed Age Groupings $SSB(D_{orig}) = 2{,}164.3$	0.513	0.580	0.967	0.996
	Percent Unemployed Within Perturbed Age Groupings $SSB(D_{orig}) = 70.5$	0.486	0.557	0.982	0.998

Table 3: Results of Disclosure Risk Measures According to Perturbation Method

Disclosure Risk Measures	Original Dataset	Method of Perturbation			
		Random	Within Marital Status	Within Marital Status and Broad Age Groups	Within Narrow Age Groups
Percent Records Unperturbed in Small Cells	0	71.83	70.71	76.6	77.78
Expected Number of Correct Matches	2,702	2,042	1,924	2,073	2,090
Percent Perturbed Records With Age Changed Within 5 Years	0	11.76	16.27	58.00	79.80

Figure 1: Percent Unemployed according to Original Age Groups and Perturbed Age Groups



Natalie Shlomo is currently on leave of absence from the Israel Central Bureau of Statistics where she was director of the Census Methodology Sector. She is a visiting academic at the University of Southampton, United Kingdom and working as a consultant for the Statistical Disclosure Control Branch of the Methodology Directorate at the Office of National Statistics. Her main areas of research are data editing and imputation when incorporating administrative data and statistical disclosure control including disclosure risk assessment for microdata, data masking techniques and information loss measures. Since 2003 she is on the Organizing Committee for the UN/ECE Work Session on Statistical Data Editing.

Ton de Waal started working at Statistics Netherlands in 1993. His first area of research was statistical disclosure control on which he co-authored two books. Since 1996 his main areas of research are data editing and imputation, where his focus has been on algorithms for automatic editing. He has published several papers on statistical disclosure control, data editing and imputation in statistical journals. Currently, he works as Manager of Statistics Development at the Development and Support Department of the Division of Business Statistics at Statistics Netherlands. Like Natalie Shlomo, he is on the Organizing Committee for the UN/ECE Work Session on Statistical Data Editing since 2003.