Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 463–478

Post Randomisation for Statistical Disclosure Control: Theory and Implementation

J.M. Gouweleeuw, P. Kooiman, L.C.R.J. Willenborg, and P.-P. de Wolf[1]

The Post RAndomisation Method (PRAM) is a perturbative method for disclosure protection of categorical variables. Applying PRAM means that for each record in a microdata file the score on a number of variables is changed according to a specified probability mechanism. This article considers the effect of PRAM on both the safety of the data and the statistical quality of the data. When applying PRAM in practice, a number of decisions have to be made, as for example to which variables and in what way to apply PRAM. These issues are briefly discussed in this article. As an example, the result of an investigation performed at Statistics Netherlands into the possibility of protecting the Dutch National Travel Survey using PRAM is presented.

Key words: Post RAndomisation Method (PRAM); disclosure; perturbed data; randomised response; Markov matrix; invariant matrix; noise; Dutch National Travel Survey.

1. Introduction

This article introduces the Post RAndomisation Method (PRAM) as a method for disclosure protection of the categorical variables in a microdata file. Applying PRAM means that for each record in a microdata file the score on one or more categorical variables is changed (independently of the other records) according to a predetermined probability mechanism. Since the original data file is perturbed, it will be difficult for an intruder to identify records as corresponding to certain individuals in the population. The records in the original file are thus protected, which is the main goal of applying PRAM. On the other hand, since the probability mechanism that is used when applying PRAM is known, characteristics of the (latent) true data can be estimated from the perturbed data file. Hence it is still possible to perform all kinds of statistical analyses after PRAM has been applied.

Originally we developed PRAM as the categorical variable analogon of noise addition to continuous variables; see e.g., Fuller (1993), Hwang (1986), and Kim and Winkler (1995). Only after we had developed most of the theory did we become aware of the obvious relationship of our method with the randomised response technique applied in survey sampling; see e.g., Warner (1965, 1971) and Chaudhuri and Mukerjee (1988). This method is employed in the case of highly sensitive questions to which the respondent is not likely to respond truthfully in a face-to-face setting. By embedding the question in

[1] Statistics Netherlands, Division Research and Development, Department of Statistical Methods, P.O. Box 4000, 2270 JM Voorburg, The Netherlands.
The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

a pure chance experiment, the true score of the respondent is never revealed to the interviewer. By knowing the probabilities involved in the chance experiment, the analyst can nevertheless uncover population frequencies of the characteristics involved, be it with some loss in precision. Retracing the literature on randomised response methods, we also found out that the idea of using Markov randomisation techniques to protect data ex post, i.e., after the data has been collected, had predecessors as well: it is mentioned in Särndal *et al*. (1992, pp. 572–573), and also in Warner (1971). Neither of these have elaborated the idea in any detail, though. An important difference between PRAM and randomised response methods is that the probability mechanism in randomised response models is necessarily independent of the (unknown) true score, whereas in the case of PRAM we condition on the true score, which is known at the moment that the method is applied.

PRAM offers an alternative to data swapping as a technique for disclosure protection. In data swapping, individual scores on certain variables are interchanged between records, thus preserving the first moments of the data. For an overview of data swapping, see e.g., Dalenius and Reiss (1982). Another article related to PRAM is that by Adam and Wortman (1989). This article provides a rough sketch of an alternative method for fixed-data perturbation.

When it comes to analysis of data protected by PRAM, there is a close relationship with the literature on the analysis of categorical variables affected by misclassifications. Indeed, applying PRAM amounts to deliberately introducing a certain amount of misclassifications in the data set. Methods to correct for the presence of misclassifications are proposed by e.g., Kuha and Skinner (1997). It should be mentioned that in most of the errors-in-variables literature, the probability mechanism giving rise to misclassifications is unknown to the analyst, so that it has to be inferred from the data. In contrast, the probability mechanism underlying misclassifications introduced by PRAM is known by the analyst, which simplifies the subsequent analysis considerably, see e.g., Chen (1979).

This article is organised as follows. In Section 2 PRAM is introduced formally. Section 3 is concerned with the effect of PRAM on disclosure limitation. It discusses when a data file to which PRAM is applied can be called safe. Section 4 describes how results of certain analyses performed on the perturbed file can be translated back to the results that would have been obtained if these analyses were performed on the original file. This translation will in general imply performing a matrix multiplication as an extra step. It is also possible to apply PRAM in such a way that the perturbed file can be used as if it were the original file. This special case of PRAM is called invariant PRAM, and is discussed in Section 5. Section 6 is concerned with questions that arise when one wants to apply PRAM in practice, as for example how to construct the probability mechanism that is used in PRAM. As an example, Section 7 applies PRAM to the Dutch National Travel Survey. Section 8, finally, contains some concluding remarks and suggests some topics for further research.

2. How to Apply PRAM

Let ξ denote a categorical variable in the original data file to which PRAM is applied, and let X denote the same categorical variable in the perturbed file. Suppose that ξ has K

categories, numbered 1, ..., K. Let $p_{kl} = P(X = l|\xi = k)$ denote the probability that an original score $\xi = k$ is transformed into a score $X = l$ for all $k, l = 1, ..., K$, and let $P = \{p_{kl}\}$ denote the $K \times K$ matrix that has p_{kl} as its (k, l)-th entry. Note that P is a Markov matrix, i.e., $P\iota = \iota$ where ι is a K-vector of 1's. It is desirable for P not to have two equal rows, since in that case the categories corresponding to these rows cannot be distinguished from one another in the perturbed file. In this article the more general assumption that P is invertible is made. Strictly speaking, this is not necessary for applying PRAM, but it turns out that P^{-1} can be used to estimate the frequency distribution of ξ in the original file, as well as the additional variance introduced by PRAM, as will be seen in Section 4.

Let $\xi^{(r)}$ (respectively $X^{(r)}$) denote the score on ξ (respectively X) for the rth record in the microdata file. Applying PRAM then means that, given that $\xi^{(r)} = k$, the score on $X^{(r)}$ is drawn from the probability distribution $p_{k1}, ..., p_{kK}$. This procedure is performed for each record in the data file, independently of the other records.

We now illustrate these ideas with an example. Suppose that the variable ξ is gender, with scores 1 = male and 2 = female. PRAM is applied to the gender variable in such a way that $p_{kk} = 0.9$ for $k = 1, 2$. Suppose that the data file originally contained 100 males and 100 females. Then the perturbed data will (in expectation) also contain 100 males and 100 females. However, 10 of these males were originally female, and similarly 10 of the females were originally male (in expectation)!

Obviously PRAM can be applied independently to different variables, by applying the method sequentially. However it is also possible to apply PRAM to more than one variable simultaneously. Consider the case where we want to apply PRAM to two categorical variables ξ_1 and ξ_2, with K_1 and K_2 categories, respectively. Let X_s denote the value of ξ_s in the perturbed file, $s = 1, 2$. Furthermore, let

$$p_{(k_1,k_2),(l_1,l_2)} = P(X_1 = l_1; X_2 = l_2|\xi_1 = k_1; \xi_2 = k_2)$$

for $k_1, l_1 = 1, ..., K_1$ and $k_2, l_2 = 1, ..., K_2$. Applying PRAM now means that $\{\xi_1^{(r)} = k_1; \xi_2^{(r)} = k_2\}$ the scores on $X_1^{(r)}$ and $X_2^{(r)}$ are (simultaneously) drawn from a probability distribution $\{p_{(k_1,k_2),(l_1,l_2)}, l_1 = 1, ..., K_1, l_2 = 1, ..., K_2\}$. Again this is performed for each record independently of the other records. There are no essential differences between applying PRAM to one variable and applying PRAM to two (or more) variables. Indeed ξ_1 and ξ_2 can be considered as one compounded variable ξ with K_1K_2 categories, numbered $1, ..., K_1K_2$ (by letting $\xi_1 = k_1$ and $\xi_2 = k_2$ correspond with $\xi = k_1 + (k_2 - 1)K_1$). The corresponding Markov matrix $P = \{p_{(k_1,k_2),(l_1,l_2)}\}$ now is the $K_1K_2 \times K_1K_2$ matrix with $p_{(k_1,k_2),(l_1,l_2)}$ as its $(k_1 + (k_2 - 1)K_1, l_1 + (l_2 - 1)K_1)$th entry. If PRAM is applied to ξ_1 and ξ_2 independently, the transition probabilities can be rewritten as

$$p_{(k_1,k_2),(l_1,l_2)} = p^{(1)}_{k_1 l_1} p^{(2)}_{k_2 l_2}, \text{where} \, p^{(s)}_{k_s l_s} = P(X_s = l_s|\xi_s = k_s), \text{for} \, s = 1, 2$$

If we let $P^{(s)} = \{p^{(s)}_{k_s l_s}\}$ denote the matrix with transition probabilities for the sth variable, then we have

$$P = \left\{p_{(k_1,k_2),(l_1,l_2)}\right\} = P^{(2)} \otimes P^{(1)}$$

where $\otimes$ denotes the Kronecker product.

Even though it is computationally convenient to apply PRAM to different variables independently, it may give rise to some unpleasant side-effects, as will be illustrated in

the next example. Suppose that the microdata file contains the variables ξ_1 = gender (with categories 1 = male and 2 = female) and ξ_2 = number of pregnancies (for illustrational purposes with just two categories: 1 = 1 or more and 2 = none). In this case, it may be convenient to apply PRAM in such a way that the perturbed file does not contain any males with one or more pregnancies, since if a record of a male with a positive number of pregnancies appears in the file, then it is obvious to any intruder that this record has been affected by PRAM. If we want to apply PRAM to ξ_1 and ξ_2 independently then the only way to exclude males with a positive number of pregnancies from the perturbed data file would be to impose $p_{21}^{(1)} = p_{21}^{(2)} = 0$. But this does not give enough freedom to protect the data, since a male in the perturbed file was originally male, and a person with a positive number of pregnancies in the perturbed file originally had a positive number of pregnancies. However, if we do not restrict ourselves to applying PRAM to each variable independently, then any set of transition probabilities $\{p_{(k_1,k_2),(l_1,l_2)}\}$ can be used, as long as $p_{(k_1,k_2),(l_1,l_2)} = 0$ if $l_1 = l_2 = 1$. In this case the probabilities can be chosen in such a way that a male in the perturbed file can have been female (either with or without any pregnancies) in the original file, so the original scores are protected.

The example shows that structural zeroes in the cross-tabulation of two variables ξ_1 and ξ_2 can be maintained by applying PRAM simultaneously to these two variables. However, preservation of structural zeroes in the cross-tabulation of ξ_1 and some other variable in the data file (to which PRAM is not applied) cannot be guaranteed because of the stochastic character of PRAM. Thus in considering a primary set of variables to which PRAM will be applied, one might wish to include extra variables in the set only because they give rise to structural zeroes, or other analytical restrictions, in cross-tabulations with the primary variables to which PRAM will be applied.

The extension to the case where we want to apply PRAM to m variables $\xi_1, ..., \xi_m$ is straightforward, only notation will become more cumbersome. In this case, it is also possible to apply PRAM independently to some variables and not independently to others. This may be accomplished by partitioning the variables $\xi_1, ..., \xi_m$ into groups in such a way that PRAM is applied to each group independently of the other groups, and within groups, dependencies may occur.

The notation introduced in this section will be used throughout the remainder of this article, i.e., ξ is a variable in the original data file with K categories and X is the value of ξ in the perturbed file. $\xi^{(r)}$ denotes the score on ξ for the rth record and P denotes the matrix with transition probabilities used for PRAM. For notational convenience, most of the theory in this article is explained for the case where PRAM is applied to one single variable. Extensions of the theory to the general case where PRAM is applied to m variables $\xi_1, ..., \xi_m$ are always straightforward by considering $\xi_1, ..., \xi_m$ as one compounded variable.

3. The Effect of PRAM on Disclosure Limitation

In this section, the effect of PRAM on disclosure limitation is considered. We start with an example. Consider a microdata file containing n records. This file represents a simple random sample of a population of size N. The data set contains exactly one female surgeon. PRAM has been applied to the gender variable, though. Independently for each record, the gender score has remained unaltered with probability 0.9, and has been changed to the

opposite score with probability 0.1. Other variables in the file have not been perturbed. Suppose an intruder knows that the population contains one female surgeon and 99 male surgeons. He or she can derive that the probability that the female surgeon in the perturbed data file is indeed the female surgeon in the population equals 0.08. This probability is very small, hence the perturbed data can be considered safe. However, if the gender score had remained unaltered with probability 0.9999 and changed to the opposite score with probability 0.0001, then the probability that the female surgeon in the perturbed data file corresponds to the female surgeon in the population equals 0.99 and of course the perturbed data file can hardly be considered safe!

It can be concluded from the previous example that we need a methodology to establish whether the perturbed file obtained by applying PRAM is indeed safe. The outline for such a methodology is developed in this section. The ideas behind this methodology are based on the statistical disclosure control rules that Statistics Netherlands currently uses. These entail that certain rare combinations of scores on variables should not be released, since such a rare combination of scores could lead to spontaneous recognition by an intruder. (For example, if a record contains the scores {occupation = mayor} and {place of residence = Amsterdam}, then any intruder knows to whom this record corresponds. Obviously, the combination of scores {mayor} and {Amsterdam} is rare (there is only one mayor in Amsterdam) and hence not safe.) The rules prescribe which combinations of variables have to be checked. A combination of scores is considered rare if the number of times that this combination occurs in the populations is below a certain threshold (where the exact value of the threshold is specified in the rules). For further details, the reader is referred to Willenborg and de Waal (1996).

When a traditional disclosure control method, such as global recoding of variables or local suppression of certain scores, is applied to a file, the resulting file is considered safe if it does not contain any rare combination of scores. When a perturbative method such as PRAM is applied to the microdata file, it does not make sense to consider a perturbed file safe when it does not contain any rare combination of scores. Indeed, starting from a file that is perfectly safe (i.e., does not contain any rare combination of scores), applying PRAM may give rise to (artificial) rare combinations. Considering such a file unsafe is obviously not sensible, since by assumption the underlying data are safe and no one is vulnerable to disclosure. As a consequence, it cannot be judged by inspection of the contents of the perturbed file if a data file protected by PRAM is safe. More generally, it is desirable that the safety of the perturbed microdata file is determined by the way PRAM is applied (including the choice of the transition probabilities) and not by the coincidentally obtained realisation of the perturbed file.

Since the rare combinations of scores in the original file are vulnerable to disclosure, it seems natural to concentrate on these combinations when looking for a sensible quantity to determine the (un)safety of a perturbed data file. In particular it is critical that such a combination, when it appears in the perturbed file, has sufficiently small probability to be a true, i.e., unperturbed, rare combination. In other words, the application of PRAM should introduce enough confusion as to whether apparently rare combinations of scores represent truly rare combinations of scores.

An obvious quantity to represent this idea would be the so-called Posterior Odds ratio as defined in, for example, Zellner (1971). Here the posterior odds ratio of the score

k ($k = 1, ..., K$), $PO(k)$ is defined by

$$PO(k) = \frac{P(\xi = k|X = k)}{P(\xi \neq k|X = k)} = \frac{p_{kk}P(\xi = k)}{\sum_{l \neq k} p_{lk}P(\xi = l)}$$

Now $P(\xi = k)$ is the probability that $\xi = k$ for any member of the population, and this quantity is in general not known, which makes it difficult to base rules on the posterior odds. If the sampling design is known, then $P(\xi = k)$ can be estimated. In general this is a complicated formula, unless the sampling design is self-weighting, in which case $P(\xi = k)$ can be estimated by $T_\xi(k)/n$, where $T_\xi(k)$ denotes the number of records in the original file for which $\xi^{(r)} = k$, and n is the number of records in the microdata file. It is also possible that an intruder knows who participated in the survey. In that case $P(\xi = k)$ should refer to the probability that $\xi = k$ for any record in the microdata file, in which case $P(\xi = k) = T_\xi(k)/n$. This is a worst case scenario, since it is easier for an intruder to identify records if he or she knows who participated in the survey.

These ideas lead to the introduction of the so-called Expectation Ratio as a measure for the amount of confusion introduced by PRAM. The expectation ratio of the score k, $ER(k)$, is defined by

$$ER(k) = \frac{p_{kk}T_\xi(k)}{\sum_{l \neq k} p_{lk}T_\xi(l)}, \qquad \text{for } k = 1, ..., K \tag{3.1}$$

The numerator of $ER(k)$, $p_{kk}T_\xi(k)$, equals the expected number of records for which $X^{(r)} = k$ given that $\xi^{(r)} = k$ (conditional given the values of $\xi^{(r)}$ for all records r in the original file). Similarly, the denominator $\sum_{l \neq k} p_{lk}T_\xi(l)$ equals the expected inflow, i.e., the expected number of records for which $X^{(r)} = k$ given that $\xi^{(r)} = l$ with $l \neq k$ (given the values of $\xi^{(r)}$ for all records r in the original file). The sum of the nominator and the denominator is exactly the expected number of records in the perturbed data file for which $X^{(r)} = k$. If k was originally a rare score, then the expectation ratio $ER(k)$ shows the ratio between the average number of records that truly belong to the rare score k and the average number of records that obtained the score k as a result of applying PRAM. The smaller the value of $ER(k)$ is, the more likely it is that a record for which $X^{(r)} = k$ did not originally belong to this score, and thus the safer the perturbed file is.

A decision whether a perturbed file created by applying PRAM is indeed safe can now be based on the expectation ratios, as follows. First it has to be decided which combinations of variables have to be inspected. For all these combinations, the expectation ratios of the rare scores are inspected. If these ratios are ''small enough,'' the perturbed file is considered safe. It is hard to define what exactly is ''small enough.'' Just like the decision which (combinations of) variables have to be inspected, this is essentially a matter of policy and should be determined by the statistical office.

4. The Effect of PRAM on Statistical Analyses

When applying PRAM to a microdata file, an obvious question is what the effect of PRAM is on all kinds of statistical analyses. Of course, analyses that are based on the perturbed file will usually give different results from analyses performed on the original file. However, results of certain analyses performed on the perturbed file can be translated to the results that would have been obtained if these analyses were performed on the original

file. This subject has been treated in Kooiman *et al.* (1997). For the sake of completeness and for later reference, some of the results are presented in this section. First we consider cross-tabulations of categorical variables, and next we discuss regression analysis on the perturbed data file. All results are discussed briefly.

Let $T_\xi = (T_\xi(1), ..., t_\xi(K))^t$ be the K-vector of frequencies in the original file of the K categories of a categorical variable ξ and similarly let T_X be the vector of frequencies in the perturbed file. Here the superscript 't' indicates transposition. Let n denote the number of records in the microdata file. Then it is easy to verify that

$$E\left[T_X \middle| \xi^{(1)}, ..., \xi^{(n)}\right] = P^t T_\xi \tag{4.1}$$

Thus, T_ξ can unbiasedly be estimated by

$$\hat{T}_\xi = \left(P^{-1}\right)^t T_X \tag{4.2}$$

Note that the matrix P has to be non-singular in order for $\hat{T}_\xi$ to be well-defined. The conditional variance of $\hat{T}_\xi$ is given by

$$V\left(\hat{T}_\xi \middle| \xi^{(1)}, ..., \xi^{(n)}\right) = \left(P^{-1}\right)^t V\left(T_X \middle| \xi^{(1)}, ..., \xi^{(n)}\right)\left(P^{-1}\right)$$

Since PRAM is applied to each record in the microdata file independently of the other records,

$$V\left(T_X \middle| \xi^{(1)}, ..., \xi^{(n)}\right) = \sum_{k=1}^{K} T_\xi(k) V_k$$

where, for $k = 1, ..., K$, V_k is the $K \times K$ covariance matrix of the outcomes $l = 1, ..., K$ of the multinominal transition process of an element with true score k:

$$V_k(l,j) = \begin{cases} p_{kl}(1 - p_{kl}) & \text{if } l = j \\ -p_{kl}\, p_{kj} & \text{if } l \neq j \end{cases} \quad \text{for } l,j = 1, ..., K$$

Substituting the estimator $\hat{T}_\xi$ for the unknown true frequencies T_ξ, we obtain an estimator for the uncertainty introduced by the noise process:

$$\hat{V}\left(T_X \middle| \xi^{(1)}, ..., \xi^{(n)}\right) = \sum_{k=1}^{K} \hat{T}_\xi(k) V_k$$

The derivations show that univariate frequencies can straightforwardly be corrected for the perturbation applied to the file. It just requires pre-multiplication with the transpose of the inverted transition probability matrix. This matrix can be supplied along with the (perturbed) data file, so that analysis only requires a matrix multiplication as an extra step in the tabulation. Using the same information, it is also possible to estimate covariances of the estimated true frequencies.

As a consequence, tabular analysis of perturbed microdata sets consisting of categorical variables poses no fundamental problems. The frequency tables summarise all available information in the data set. The presence of a small amount of extra variance in these estimates will generally pose no problem to the analyst, as the extra variance just adds to the measurement errors that are present in the data anyhow. Moreover, there will also be sampling variance present in the data. Multivariate analyses for categorical data, like loglinear

modelling or correspondence analysis, can proceed from the estimates of the true tables. Chen (1979) shows that loglinear modelling of contingency tables can also directly proceed from the perturbed table by incorporating the perturbation design in the model.

The moment estimator (4.2) satisfies $\hat{T}_\xi \iota = n$, i.e., the estimated frequencies add to the number of records n in the data file. However, since P^{-1} is not a Markov matrix, the estimated frequencies themselves may fall outside the feasible range $[0, n]$. In particular, negative frequencies may occur when the corresponding frequencies in the original data file are sufficiently small. Of course this is undesirable. Moreover, (4.2) is hence not the maximum likelihood estimator (MLE) of T_ξ when maximising over all distribution functions that could have generated the original table T_ξ. This problem has been noticed in the literature on randomised response methods; see Chaudhuri and Mukerjee (1988) and the references cited there. For the randomised response model, it has been derived that in the case where ξ has two categories (say 1 and 2), the MLE is obtained as

$$\hat{T}_\xi^{\text{MLE}} = \begin{cases} \hat{T} & \text{if} \quad 0 < \hat{T}_\xi(1) < n \ \text{ and } \ 0 < \hat{T}_\xi(2) < n \\ (0, n)^t & \text{if} \quad \hat{T}_\xi(1) < 0 \\ (n, 0)^t & \text{if} \quad \hat{T}_\xi(2) < 0 \end{cases}$$

The MLE is biased, but it has a smaller mean squared error than $\hat{T}_\xi$.

If ξ has more than two categories, the conditional distribution function of T_X given T_ξ and P is a convolution of multinomial distributions, which is nontrivial. It is therefore difficult to verify whether the truncation at the boundary of the feasible region generalises to the case where ξ has more than two categories. Nevertheless, a practical solution to get rid of negative entries in $\hat{T}_\xi$ suggests itself: truncate these entries at zero and subsequently renormalise the remaining entries to add to n again.

A quite different approach is to apply Bayesian methods, using uniform or Dirichlet priors restricted to the feasible region. Given the complicated likelihood function, the posterior distribution of T_ξ is bound to be quite intractable in the case where ξ has more than two categories, so that analytical results will not easily be obtained. Still another solution is discussed below in Section 5, where we consider invariant matrices P. Using such a matrix to protect the data file entails that T_X itself is an unbiased estimator of T_ξ. Since T_X is admissible by construction, we will never end up with negative frequencies.

A natural further question is what the effect of PRAM is on different types of multivariate analysis, for example regression analysis. Suppose we want to perform a regression of some numerical variable y on a categorical variable ξ, and suppose that PRAM has been applied to ξ. We introduce the dummy variables $\delta_1, ..., \delta_K$, where for each record in the data file $\delta_k = 1$ if $\xi^{(r)} = k$ and $\delta_k = 0$ otherwise. Furthermore, let

$$T_\xi^y = \left(T_\xi^y(1), ..., T_\xi^y(K)\right), \text{ with } T_\xi^y(k) = \sum_{r=1}^{n} y^{(r)} I_{\{\xi^{(r)}=k\}}$$

where $y^{(r)}$ denotes the value of y for the rth record in the microdata file and I denotes the indicator function. Let T_X^y be defined similarly. It was shown in Kooiman *et al.* (1997) that T_ξ^y can unbiasedly be estimated by

$$\hat{T}_\xi^y = (P^{-1})^t T_X^y$$

Note that the elements of T_ξ^y divided by the number of records in the data file n are in fact

the (empirical) second moments (at zero) of the joint distribution of the dummy variables $\delta_1, ..., \delta_K$ and the numerical variable y. Now a regression of y on ξ amounts to a regression of y on $\delta_1, ..., \delta_K$. The regression coefficient is given by $(D^tD)^{-1}D^ty$, where D is the $n \times K$ matrix which has its (r, j)th entry equal to the value of δ_j for the rth record in the original data file. (D^tD) can unbiasedly be estimated by using $\hat{T}_\xi$ (note that (D^tD) is the diagonal matrix which has $T_\xi(k)$ as its (k, k)th entry) and D^ty can unbiasedly be estimated by $\hat{T}^y_\xi$. Since these estimates are unbiased the resulting regression estimator is consistent. This conclusion holds true for all statistical analysis techniques based on second moments of the data, such as discriminant analysis and analysis of variance.

In summary, we have demonstrated in this section that both tabulation and standard multivariate analysis techniques for mixed categorical/numerical variable models can easily be adapted to data sets with randomised categorical variables. It only requires pre-multiplication by the inverse of the transposed Markov transition matrix involved in the randomisation process.

5. Invariant PRAM

So far, the only restriction that has been imposed on the Markov matrix P used for PRAM is non-singularity. In this section, it will be shown that the analyses of the perturbed file can be simplified if a special choice is made for P. The general idea is that the perturbed file should be close to the original file. This can be achieved by choosing P in such a way that $\|P^tT_\xi - T_\xi\| < \epsilon$, for some pre-specified $\epsilon > 0$, where $\|.\|$ denotes a norm. In this section we consider the simpler case where the matrix P is chosen in such a way that the distribution of ξ over the different categories is invariant with respect to P, i.e., P should in fact satisfy the stronger condition that

$$P^tT_\xi = T_\xi \tag{5.1}$$

The identity matrix always satisfies this equation, but this is not very interesting, since the perturbed data file will be the same as the unperturbed data file. A non-trivial solution P of (5.1) can be constructed as follows. Assume without loss of generality that $T_\xi(k) \geq T_\xi(K) > 0$, for $k = 1, ..., K$, and let, for some $0 < \theta < 1$:

$$p_{kl} = \begin{cases} 1 - (\theta T_\xi(K)/T_\xi(k)) & \text{if } l = k \\ \theta T_\xi(K)/((K-1)T_\xi(k)) & \text{if } l \neq k \end{cases} \tag{5.2}$$

It is easy to verify that $P = \{p_{kl}\}$ is indeed a Markov matrix satisfying (5.1).
Now suppose that P is chosen in such a way that (5.1) is satisfied. In that case

$$E(T_X|\xi^{(1)}, ..., \xi^{(n)}) = P^tT_\xi = T_\xi$$

where the first equality follows from (4.1) and the second equality follows from (5.1). This means that T_ξ can unbiasedly be estimated by

$$\hat{T}_\xi = T_X \tag{5.3}$$

hence the estimator for T_ξ can directly be obtained from the perturbed file. Of course this simplifies the analysis, since no premultiplication by a matrix is needed. When P satisfies (5.1), we may also consider another estimator for T_ξ that may perform better. This estimator is discussed at the end of the present section.

In the previous paragraph, the matrix P was invariant with respect to the distribution in the data file. Alternatively, we can choose P to be invariant with respect to the distribution in the population (that is, with respect to the weighted frequencies), i.e., P should be such that

$$T_\xi^w = P^t T_\xi^w$$

where T_ξ^w is the K-vector of weighted frequencies. In that case, T_ξ^w can be directly estimated from the perturbed file by T_X^w, where T_X^w is the K-vector of weighted frequencies in the perturbed file. (For details, the reader is referred to Kooiman *et al.* 1997.)

When invariant PRAM is applied, T_X equals T_ξ in expectation. The two quantities are not exactly the same, though. Since PRAM uses perturbations that are independent for the different records it is only possible to preserve entries of T_ξ by switching off PRAM for these entries altogether. Otherwise, identically preserving some entries requires dependency of the perturbations for the different records, as e.g., in data swapping. A method that protects tables by random perturbations while identically preserving its marginals is given in Duncan and Fienberg (1998). Their approach implies misclassifications that are (singularly) dependent between records.

When the Markov matrix P satisfies (5.1), in some cases there is an estimator superior to (5.3). For, if the invariant matrix P is shipped to the analyst along with the perturbed microdata file, then the analyst can unbiasedly estimate T_ξ as the eigenvector of P corresponding to the eigenvalue 1. In fact, if the eigenvalue 1 has multiplicity 1, then this estimator is uniquely determined and has variance equal to 0. This implies that if P is such that all possible cross-tabulations in the original microdata file are preserved, then the original microdata file can be retrieved from P. In practice, this situation will not occur. First of all, the preceding argument only holds if the eigenspace corresponding to the eigenvalue 1 is one-dimensional. It follows from the theorem of Perron-Frobenius (see e.g., Seneta 1981) that this only holds if P is irreducible. It is easy to make sure that P is not irreducible, by letting P be a block diagonal matrix. Secondly, it will not be tractable to preserve all possible cross-tabulations, as we will argue below.

The obvious advantage of using invariant PRAM is that we can just work with the perturbed data file, no pre-multiplication by a matrix being needed. As a consequence, there will never be negative estimated frequencies, for which a correlation has to be made. A drawback is that there is less freedom in the choice of the Markov matrix P. Moreover, the matrix P may become very large if we want to preserve the distribution of all possible cross-tabulations. Indeed, suppose that the data file contains m variables $\xi_1, ..., \xi_m$ with $K_1, ..., K_m$ categories, respectively. Then we can preserve all possible cross-tabulations by considering $\xi_1, ..., \xi_m$ as one compounded variable ξ with $K = K_1 \times ... \times K_m$ possible categories. Note that K may indeed be very large. Moreover, many of the K categories of this compounded variable ξ will contain no observations, and this makes a non-trivial choice for P that satisfies (5.1) almost impossible.

In practice, it is probably best to apply PRAM in such a way that some distributions are preserved while others are not. Some analyses can then be performed directly on the perturbed file, while others require an extra matrix multiplication. It is a topic for further research whether it is possible to choose P in such a way that all analyses can (unbiasedly) be performed on the perturbed file, even if not all possible cross-tabulations are preserved.

In choosing an invariant matrix P, we probably have to choose which distribution should be kept invariant (the distribution in the data file or the distribution in the population). It is another topic for further research whether it is possible to choose P in such a way that both distributions are preserved. Finally it should be noted that an analyst always needs the Markov transformation matrices, be they invariant or not, once he or she wants to compute the extra variance introduced by using PRAM.

6. Problems When Applying PRAM in Practice

Now that we have introduced PRAM, we turn to the question of how to apply this technique in practice. This involves several aspects. First of all, it has to be decided to which variables PRAM should be applied. Next, for each of the variables to which PRAM will be applied, it should be decided which category can be changed into which category and with what probability. We will discuss these problems one by one.

First of all, it has to be decided which variables will be perturbed. If PRAM is applied to (some of the) identifying variables in the microdata file, then as a result it becomes more difficult for an intruder to recognise a record as corresponding to some individual in the population. Alternatively, PRAM can also be applied to the sensitive variables in the microdata file. In that case, an intruder may recognise the record of an individual in the microdata file, but he or she cannot be sure as to whether the sensitive information obtained from the data file is correct. There is some discussion on whether a file can be called safe when PRAM would only be applied to sensitive variables. In that case, records can be identified as belonging to certain individuals, and thus an intruder has discovered that the individual has participated in the survey. This may itself be sensitive information. From the point of view of statistical analyses, it does not matter whether PRAM is applied to identifying or sensitive variables.

A microdata file will usually contain many variables that are candidates for applying PRAM. A choice has to be made whether it is preferable to perturb only a few variables (and thus in order to obtain a safe file, perturb each of them a lot) or to perturb many variables (and perturb each of them a little bit). Both strategies could lead to a safe file. A choice of either of the strategies could be motivated by the information loss due to the application of PRAM. It is a reasonable approach to produce a safe microdata file from an unsafe one by applying disclosure control techniques in such a way to the original microdata set that the resulting file is safe (according to the criteria applied) while the amount of information loss due the modification is minimised. This information loss would then have to be quantified. A straightforward measure is the increase in variance of the estimates due to the measurement error introduced by PRAM. Appropriate variance formulas have been derived in Section 4. Another approach that could be used for this purpose involves using the concept of entropy as introduced by Shannon in communication theory in the 1940s (see e.g., Shannon and Weaver 1949), or the more general measures of disclosure that are introduced in Duncan and Lambert (1986).

When it has been decided to which variables PRAM should be applied, the next step is to decide which category can be replaced by which category and what probability mechanism should be used to do so. An algorithm for this will be described. The idea behind this algorithm is that for computational convenience we do not want it to be possible for each

category to be replaced by any other category, especially when the number of categories is large. Consider a variable ξ with categories $1, ..., K$ to which PRAM will be applied. First of all, a partition of the categories into groups $C_1, ..., C_G$ is made in such a way that each category can only be replaced by a category within the same group. More formally $C_1, ..., C_G$ are such that

$$C_g \cap C_h = \varnothing \text{ if } h \neq g, \qquad \cup_{g=1}^{G} C_g = \{1, ..., K\} \tag{6.1}$$
$$p_{kl} = 0 \text{ if } k \in C_g, l \in C_h \text{ and } h \neq g$$

It is up to the data protector to decide how these groups $C_1, ..., C_G$ should be constructed. This can be done in such a way that a group always contains categories that are (in some sense) similar, but this is by no means necessary. It can also be done in such a way that the expectation ratios become small.

The group structure immediately determines the structure of the Markov matrix. For, the categories $1, ..., K$ can now be reordered in such a way that P can be written as

$$P = \begin{pmatrix} P_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & P_G \end{pmatrix} \tag{6.2}$$

This follows immediately from (6.1). So the only thing left to choose are the values p_{kl} for k and l in the same group C_g. If general PRAM is applied, then any set of values for these p_{kl} will suffice, as long as the resulting matrix is an invertible Markov matrix. When invariant PRAM is applied, the matrix P should also satisfy (5.1). This still leaves a lot of possible choices.

7. A Case Study: The Dutch National Travel Survey

In this section we illustrate the application of PRAM to a real-life problem. This section is intended to illustrate the theoretical remarks made in the previous sections.

At Statistics Netherlands, it has been investigated how to apply PRAM to the Dutch National Travel Survey. In this survey, a 1% sample is drawn from the Dutch population and each person in the sample has to record all the travelling he or she does on one specific day. The microdata file contains information about these trips (such as location of departure, location of arrival, travelled time, travelled distance, means of transportation used), as well as background information on the persons in the file (such as gender, marital status, highest level of education, place of residence), and background information on the household to which the person belongs (such as size and composition of the household). The file contains a total of 167,923 records.

The microdata of the Dutch National Travel Survey is used by clients outside Statistics Netherlands. These clients have expressed the wish that not only the place of residence of each person is included in the microdata file, but also the postal code (a four-digit number) of place of residence. If we want to include the postal code in the microdata file, then it is no longer possible to arrive at a satisfactory level of disclosure protection using traditional methods (like recoding and data suppression). Therefore, it was considered to what extent PRAM would provide useful services.

According to the existing rules of Statistics Netherlands, eight combinations of two variables have to be inspected to see whether rare combinations of scores occur. In order to obtain a safe file, PRAM had to be applied to six variables: postal code (with 3,555 categories), mari-tal status, composition of the household, age (in classes), main activity status (indicating whether someone is working, unemployed, retired, and so on) and any household vehicle availability. It was decided to apply PRAM to these variables independently. In fact invariant PRAM was applied to each variable separately, so that all the analyses concerning one single variable can be performed directly on the perturbed data file. Invariant PRAM on all the six variables simultaneously is intractable, since this would lead to one compounded variable with 400 million categories. Since the data file contains 167,923 records, most of the categories of the compounded variable will contain no observations, which makes it difficult to determine an invariant matrix P. Moreover, from a computational point of view, it is impractical to work with a matrix of dimension 400 million.

Analyses concerning more than one variable would theoretically involve multiplication by $(P^{-1})^t$, as was shown in Section 4. For the moment, this may not be acceptable to the clients. Therefore, PRAM is applied in such a way that the original file is not too seriously perturbed, and can be used directly for all analyses, even when the analysis concerns multiple variables. In that case, estimates will not be unbiased, but it is likely that the bias is small. This assumption was tested by comparing several cross-tabulations in the original publication of the Dutch National Travel Survey with the same cross-tabulations based on the perturbed file. It turned out that the differences were all minor, and much smaller than the sampling margin.

Note that, since the place of residence can in fact be deduced from the postal code, applying PRAM to postal code can lead to inconsistencies in the perturbed file. This could give a researcher a clue as to which records postal codes have been affected by PRAM. This is not desirable. Therefore, the following approach was taken. The place of residence is first deleted from the original file. Next PRAM is applied to the postal code, and finally the place of residence is recomputed from the perturbed value of the postal code. This way, place of residence is always consistent with postal code. This idea is similar to the idea of matching additional variables to the perturbed file, as described in Kooiman, Willenborg, and Gouweleeuw (1997).

For each variable, the Markov matrix P that was used has the block diagonal form as in Formula (6.2). Furthermore within block P_g the probabilities p_{kl} were defined by Formula (5.2). For the postal codes, the value of θ in (5.2) is 0.9 for all the groups C_g. This implies that the value of p_{kk} (the probability that postal code k is not perturbed) varies between 0.1 and 1. For the other variables, the value of θ is chosen to be is 0.1, which implies that the value of p_{kk} varies between 0.9 and 1. When the value of θ is known, it can be computed what the expected number of records is in which the value of some variable is changed. These figures are given in Table 7.1.

Finally it had to be decided whether the perturbed file that was created by applying PRAM was indeed safe. A total of 38,466 expectation ratios were studied, to check whether they were small enough. It turned out that 15.8% of these ratios were larger than 20 and only 2.4% larger than 100. Furthermore, approximately 37% of the expectation ratios were smaller than 3. It still has to be decided whether this represents

Table 7.1. Expected number of changes per variable

Variable	Expected number of changes	Expected percentage of changes
Postal code	22,551	13.4
Marital status	1,041	0.6
Age (in classes)	2,803	1.7
Main activity status	914	0.5
Composition of the household	879	0.5
Household vehicle availability	889	0.5

a sufficient level of disclosure protection according to the current standards of Statistics Netherlands.

The conclusion which can be drawn from this application is that PRAM is a useful addition to the existing tools for disclosure protection. However, it is not possible to obtain a perturbed file with a high level of detail on a regional variable (for instance postal code) which is only slightly perturbed and can be analysed as if it were the original file. In the case of the Dutch National Travel Survey, some postal codes had to be perturbed significantly (i.e., with probability 0.9), and it still has to be decided whether the perturbed file is considered safe by the standards of Statistics Netherlands. At the time of writing this article, it was still unknown how the clients of Statistics Netherlands experienced working with a perturbed data file.

8. Concluding Remarks

In this article, PRAM has been described as a method for disclosure protection of categorical variables in a microdata file. There are a number of issues concerning PRAM that need further research.

In Section 3, the expectation ratio was introduced to determine whether the perturbed file is safe. Another possible way to do this is by using the theory of exact matching, when errors occur in the variables that have to be matched (see e.g., Fellegi and Sunter 1969). It could be checked to what extent the records in the original file could be matched successfully to those in the perturbed file, given some matching algorithm. For a given record in the original file, the number of records in the perturbed file that match this record can be determined. The larger this number is, the harder it is to match the original file to the perturbed file, and thus the safer the perturbed file is. An alternative measure of re-identification risk is described in Skinner (1997), where for each record the probability of re-identification is calculated.

Another important problem is the existence of dependencies between variables in the original microdata file. The variables in the original data file satisfy all kinds of edit rules. If PRAM is applied to different variables independently, then inconsistencies in the perturbed file may occur, i.e., the edit rules may not be satisfied anymore. This gives an intruder a clue as to which values are perturbed, and he or she can partially undo the perturbation process. For example, in the Dutch National Travel Survey, the data file contains postal code as well as place of residence. The place of residence can be deduced from the postal code. If PRAM should be applied to postal code and place of residence separately (or only to postal code without taking the place of residence into account), these two could be inconsistent in the perturbed file. This then gives an intruder a clue that in

such a record a score has been affected by PRAM. The question is how these dependencies should be dealt with routinely.

It was shown that it is possible to make cross-tabulations and perform regression analysis on the perturbed file. An open question is which other standard statistical techniques withstand PRAM, and how these techniques should be modified to account for the randomisation process.

Some microdata files have a hierarchical structure, containing information on for example households as well as on persons. If PRAM is applied to the records of the different persons, then it may occur that two persons belong to the same household but score differently on the same household variable. The question is how to take this structure into account when applying PRAM. This may be accomplished by abandoning the requirement that PRAM is applied to each record independently of the other records.

A more practical problem is how PRAM should be incorporated in a software package for disclosure protection of a microdata file. It seems wise to restrict the choices of the Markov matrix to a few special classes, in order to keep the number of possible options tractable. If a general Markov matrix is allowed, every user of the package has to have a thorough knowledge of PRAM before being able to use the package in a sensible way. It is a topic for further research how the Markov matrix should be chosen, and what the effects of any restrictions will be.

In this article, we have only considered PRAM as a method for disclosure protection. In general we want to apply a mixture of several disclosure protection methods: PRAM, global recoding, local suppression, etc. Generally speaking, it is still unclear what the implications for disclosure protection rules are and what the consequences for statistical analyses will be.

9. References

Adam, N.R. and Wortman, J.C. (1989). Security-control Methods for Statistical Databases: A Comparative Study. ACM Computing Surveys, 21, 4, 515–556.

Chaudhuri, A. and Mukerjee, R. (1988). Randomized Response, Theory and Techniques. Marcel Dekker Inc, New York.

Chen, T.T. (1979). Analysis of Randomized Response as Purposively Misclassified Data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 158–163.

Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference, 6, 73–85.

Duncan, G.T. and Fienberg, S.E. (1998). Obtaining Information While Preserving Privacy: A Markov Perturbation Method for Tabular Data. Proceedings of the Statistical Data Protection '98 Conference, Lisbon. Forthcoming.

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited Data Dissemination. Journal of the American Statistical Association, 81, 10–18.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 40, 1183–1210.

Fuller, W.A. (1993). Masking Procedures for Microdata Disclosure Limitations. Journal of Official Statistics, 9, 383–406.

Hwang, J.T. (1986). Multiplicative Errors-in-variable Models with Applications to Recent Data Released by the U.S. Department of Energy. Journal of the American Statistical Association, 81, 680–688.

Kim, J.J. and Winkler, W.E. (1995). Masking Microdata Files. Proceedings of the Section on Survey Research Methods, American Statistical Association, 114–119.

Kooiman, P., Willenborg, L.C.R.J. and, Gouweleeuw, J.M. (1997). PRAM: A Method For Disclosure Limitation of Microdata. Research paper no. 9705, Statistics Netherlands, The Netherlands.

Kuha, J. and Skinner, C. (1997). Categorical Data Analysis and Misclassification. In Lyberg, L., Biemer, P., Collins, M., DeLeeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.): Survey Measurement and Process Quality, Chapter 28, John Wiley and Sons, New York.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag, New York.

Seneta, E. (1981). Non-negative Matrices and Markov Chains. Springer-Verlag, New York.

Shannon, C.E. and Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press, Urbana.

Skinner, C. (1997). Estimating the Re-identification Risk per Record in Microdata. Proceedings of the Third International Seminar on Statistical Confidentiality, Bled, 1996.

Warner, S.L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 57, 622–627.

Warner, S.L. (1971). The Linear Randomized Response Model. Journal of the American Statistical Association, 66, 884–888.

Willenborg, L. and de Waal, T. (1996). Statistical Disclosure Control in Practice. Lecture Notes in Statistics 111, Springer-Verlag, New York.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons, New York.

Received August 1997
Revised April 1998