WP. 15
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (ii): Disclosure risk, information loss and usability of data

EXPERIENCE OF USING A POST RANDOMISATION METHOD AT THE OFFICE FOR NATIONAL STATISTICS

Supporting Paper

Submitted by the Office for National Statistics, United Kingdom[1]

[1] Prepared by Christine Bycroft (Christine.Bycroft@ons.gsi.gov.uk) and Katherine Merrett (Katherine.Merrett@ons.gsi.gov.uk).

Experience of using a Post Randomisation Method at the Office for National Statistics

Christine Bycroft* and Katherine Merrett*[1]

* Statistical Disclosure Control Centre, Methodology Directorate, Office for National Statistics, UK, email Christine.Bycroft@ons.gsi.gov.uk or Katherine.Merrett@ons.gsi.gov.uk

1 Introduction

The release of microdata files from sample surveys or extracts from Censuses are of great analytical value to researchers. When surveys and Censuses are undertaken confidentiality guarantees are given to respondents usually saying that information that could lead to their identification will not be released. This means that when a statistical office is considering whether to release and how to release these microdata files they need to consider the risks of possible disclosure of confidential information. The first step in this process is to make an assessment of the disclosure risk. There are various methods that have been developed and other methods which are being developed to conduct a disclosure risk assessment. Once a risk assessment has been conducted, some protection methods can then be applied to the data. Typically at the Office for National Statistics (ONS) the main disclosure control method used is recoding. However there comes a point where further recoding causes a large decrease in the information released for little decrease in disclosure risk. When this point is reached the only way to protect the remaining high risk records will be to remove them, or to alter one or more of their characteristics. Our preference was for a method that would perturb values of high risk records in a manner that has a small impact on analysis outcomes.

As nearly all of the variables on social surveys are categorical, we have developed a method based on the Post Randomisation Method (PRAM). The Invariant PRAM method especially seemed attractive to us as it conserves the expected values of the frequencies for each category after the perturbation. The PRAM method has been applied to the 2001 Individual Samples of Anonymised Records (SAR) drawn from the Census. In this context we have adapted the Invariant PRAM in the following three ways:

[1] The PRAM methodology developed by the ONS for the use on the 2001 Individual SAR was joint work between Bill Gross, Philippe Guiblin, Joshua Fox and Katherine Merrett. Thanks to Natalie Shlomo who provided helpful comments on an earlier draft.

- We were interested in preserving relationships between variables and developed an adaptation of the method which enables some control on the joint distributions between the variables perturbed by the PRAM and other variables in the microdata.
- Our implementation of the method conserves the exact frequencies of the categories after the perturbation and not just the expected values.
- In contrast to the typical PRAM described in Willenborg and De Waal (2001), we are perturbing only the high risk records within the sample, not the whole sample.

In section 2 of the paper we provide a description of the PRAM method as introduced by Kooiman et. al. (1997) and the way in which ONS has adapted the methodology to suit its needs. In section 3 we provide details of how the methodology was applied in practice to the 2001 individual SAR. Section 4 of the paper describes some of the methods that were used to examine the effects that the perturbation has had on data quality. Finally, in section 5, we summarise our experiences from applying the new PRAM methodology and draw some conclusions.

2 Method: PRAM

PRAM is a disclosure control technique for microdata. It was introduced in 1997 by Kooiman et. al. as a disclosure control method to be applied to categorical data in microdata files. The values of a categorical variable for certain records in the microdata file are changed according to a prescribed probability. Each new value may or may not be different from the original value. For example, a person who is classified as a widow may be re-classified as single under PRAM.

The probability mechanism is described by an invertible transition matrix P per variable. Let $P = (p_{ij})$ be an L x L matrix for a variable having L categories. The entries of the matrix are the conditional probabilities

p_{ij} = Pr(New_value = j | Old_value = i).

The resulting perturbed file is released along with information about the probability mechanism (transition matrix) used. The researcher can use this information to adjust his or her analysis regarding the perturbation caused by PRAM. The perturbation can be seen as a form of prior misclassification and methods to deal with misclassification can be found in the literature, see Kuha and Skinner (1997).

As explained in Willenborg and De Waal (2001, Section 5.5.1), PRAM offers protection by inflow and outflow: inflow from safe combinations of values to risky combinations, and outflow from risky combinations to safe combinations. The resulting perturbed file will retain some unusual or high risk combinations, but there will be uncertainty over whether these have been created through the perturbation process or are original values from a respondent.

A problem with PRAM as described above is the possibility of creating invalid or highly unusual combinations, e.g. a 14 year old doctor or 17 year old widow. This is partly a result of allowing inflow as part of the confidentiality protection. Also using the transition matrix in the analysis may be a burden to some researchers as standard statistical applications may be more difficult to implement.

PRAM has not yet been applied extensively by statistical agencies. This may be because there is little practical knowledge available on the effect that it has on disclosure control and on the information loss that it introduces. Some results are available in de Wolf and van Gelder (2004) who carried out an empirical evaluation of PRAM.

2.1 Invariant PRAM

A specific form of PRAM introduced by Kooiman et. al. (1997) is the invariant PRAM method. In this form, applying PRAM is invariant with respect to the frequencies of the variables. Let $P = (p_{ij})$ be the transition matrix for a variable ? having L categories, and F be the vector of frequencies containing the sample counts of each category. The matrix P is chosen such that:

$$P^t F = F \qquad (1)$$

As a result, frequencies after the perturbation are in expectation equal to the original frequencies of ?. This relieves the user of the perturbed file from the extra effort of obtaining unbiased estimates of the original data. It is still important to release the transition matrices so that the user can compute the extra variance introduced by using invariant PRAM.

2.2 Adapting Invariant PRAM

The Individual SAR is a 3% sample of some 1.8 million individuals drawn from the 2001 Census. The geography area identification on the file is Government Office Regions for England and country for Wales, Scotland and Northern Ireland. In addition to applying the PRAM to the individual SAR it has also been applied to the Small Area Microdata file also drawn from the Census. This file is a 5% sample of individuals but the geographic information is more detailed Local Authority for England, Wales and Scotland and Parliamentary Constituencies for Northern Ireland.

The adaptations to PRAM were motivated by the need to protect the 2001 Individual SAR. In this situation it was possible for us to identify high risk records because we had the full population data from which the sample was derived. Our approach to reducing disclosure risk was to use recoding of variable categories to a point where further recoding would seriously impact on data use without much increase in protection. The remaining subset of high risk records were protected using perturbation through use of our adapted PRAM. PRAM as implemented here has the advantage of being able to target modification of the file directly to high risk records and to the particular variables within each high risk record that contribute most to disclosure risk.

In this situation where we are only perturbing high risk records, protection against disclosure is achieved by largely removing any inflow and relying only on outflow. In contrast to invariant PRAM, the transition matrix P will need to ensure that the probability of changing values is maximised. Thus we need to minimize the probabilities that are on the diagonal of the transition matrix (i.e., the probability that no change occurs). Other constraints on the transition matrix are that it is invariant and that statistical properties of the dataset stay similar after the perturbation. This means that we want to perturb categories to other categories that are both feasible and will not result in highly unusual combinations.

In summary, the method developed for obtaining the transition matrix P ensures three goals:

1. the probabilities of no change are minimised
2. in expectation, the output distributions are the same as the input distributions
3. transition to “similar values” are maximised

To obtain the transition matrix P we used the linear programming feature of SAS. The routine minimises an objective function, subject to constraints. The objective function is defined as follows:

$$\sum_{i} w_{ii} p_{ii} + \sum_{i \neq j} w_{ij} p_{ij} \qquad (2)$$

where $W = (w_{ij})$ is a Weight Matrix: a low weight for a preferred transition and a high weight for a non-preferred transition. The Weight Matrix is set up to avoid extreme transitions. Rather than having extreme changes that might create highly unusual individuals or invalid combinations, we prefer to keep the values as they are. Minimising the objective function will lead to minimum probabilities on the diagonals, subject to also avoiding extreme transitions.

Example: If we have an age variable with 9 categories (0-9; 10-19; 20-29;…; 70-79; 80+), and

- transition to the adjacent is preferred to any other transition;
- no transition is preferred as a second choice;
- transition to extreme values is not desired

A possible Weight Matrix is the following:

$$\begin{pmatrix} 2 & 1 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 2 & 1 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 1 & 2 & 1 & 3 & 4 & 5 & 6 & 7 \\ 4 & 3 & 1 & 2 & 1 & 3 & 4 & 5 & 6 \\ 5 & 4 & 3 & 1 & 2 & 1 & 3 & 4 & 5 \\ 6 & 5 & 4 & 3 & 1 & 2 & 1 & 3 & 4 \\ 7 & 6 & 5 & 4 & 3 & 1 & 2 & 1 & 3 \\ 8 & 7 & 6 & 5 & 4 & 3 & 1 & 2 & 1 \\ 9 & 8 & 7 & 6 & 5 & 4 & 3 & 1 & 2 \end{pmatrix}$$

The constraints of the optimization routine are the following:

- the rows of the transition matrix sum up to 1;
- all the probabilities are positive;
- expected output frequencies are equal to input frequencies;

Given the Weight Matrix, the optimisation routine finds p_{ij} values that minimise the objective function.
The constraints are mathematically expressed as:

$$\begin{cases} \sum_{j=1}^{L} p_{ij} = 1 \quad with \;\; i = 1,...,L \\ p_{ij} \geq 0 \quad \forall \quad (i,j) \\ \sum_{j=1}^{L} p_{ij} F_i = F_i \quad with \;\; i = 1,...,L \end{cases} \tag{3}$$

2.3 Preserving Univariate Distributions

In applications of the invariant PRAM, the movement of a record from category *i* to category *j* is applied in a way that ensures that the expected values of the frequencies of the categories will be preserved.

At the ONS a method was developed for obtaining the exact frequencies of the categories after the perturbation scheme and not just in the expected values. First, based on the transition matrix, we calculate how many records should be changed from category i to category j. Denote this number by r_{ij} . The records are then sorted randomly within category i and a sub-sample of size r_{ij} is drawn for each j. All records in the sub-sample have their category i changed to category j. The method is repeated for all categories, i = 1,..., L.

2.4 Stratification: A way to preserve multivariate distributions.

The modified PRAM method preserves the univariate distributions as shown above. Preservation of multivariate distributions is also important to users.

The method we have developed to preserve the relationship between the prammed variable (e.g. age) and another variable (e.g. marital status) is to PRAM the variable within strata defined by the values of the second variable (e.g. PRAM age within each stratum 'single', 'married', etc...). This provides some control on the transitions. To make this process more effective, the second variable, if prammed, should be prammed also within strata defined by the first variable. We call the variable used for the definition of the strata a 'control variable' Strata can also be defined as a combination of several 'control variables'. This ensures that the perturbations are constrained for the joint distributions of the prammed variables with the control variables.

2.5 Disclosure risks

As explained above in section 2 protection is provided by both inflow and outflow. In the ONS implementation of PRAM there is considerable outflow. However little protection is provided in the file by inflow as only those records which are high risk are perturbed and perturbed values are controlled to avoid creating unusual and therefore potentially risky combinations.

To counteract this reduced protection, we are not providing the transition matrices to the user. This in turn limits the information available to the user about the specific perturbation mechanism that has been applied, and means that the user will be unable to calculate the additional variance introduced by PRAM. However only a small proportion of records in the file have been perturbed and we are providing the analyst with some information by flagging records which have been perturbed with the same flag as used for marking imputed records. The implication from an intruder point of view is that an intruder does not know whether a flagged record is a true value, a perturbed value or an imputed value.

3 An example of the use of Stratified Invariant PRAM

Stratified Invariant PRAM as described above was implemented on the 2001 Individual Licensed SAR from the Census. A disclosure risk assessment was conducted on the data, see Gross et. al. (2004). The file is available for researchers to download under a License agreement, and the protection measures reflect ONS assessment of disclosure risk under these conditions. Our disclosure risk measure was the percentage of population uniques in the sample for the private database cross match scenario[2] as developed by Elliot and Dale (1998). We aimed to protect only against attempts at exact matching, so considered that perturbing the value of one variable in a high risk record provided sufficient protection.[3]

The risk of the Individual SAR after applying recodes was about 4% of records being population unique. Our aim was to bring the risk down to a tolerable threshold of approximately 1% of records being population unique. Further recoding would have rendered the file of little use to researchers, therefore it was decided that for the remaining high risk records some perturbation would be applied.

3.1 Record and variable selection for PRAM.

A special uniques analysis was run on the recoded file. For more information on the special uniques methodology see Elliot et. al. (2004). We used the results of the special uniques analysis to efficiently target the perturbation to the highest risk records and highest risk variables. The special uniques analysis ranks sample uniques in the file by what is called a DIS/SUDA score. PRAM was applied only to records that exceeded a threshold for the DIS/SUDA score and were population unique for the private database scenario.

For each sample unique record in the file, the special uniques method tells us the relative contribution of each variable to the disclosure risk. For each high risk record, we identified the variable that contributed the most to the risk and then prammed that variable on the record. The special uniques analysis also ranks variables in order of their overall contribution to disclosure risk. PRAM was applied to variables

[2] The Private Database Cross Match intruder scenario assumes an intruder is potentially able to match the following variables against available databases: Geography, age, sex, marital status, number of cars, number of dependent children, work place, distance of journey to work, number or residents, number of earners, tenure and primary economic status.

[3] Because Scottish data has a different risk profile, more than one variable in a record was sometimes prammed.

sequentially, beginning with the highest risk variable. Using 10 variables allowed us to perturb one variable on all the target high risk records.

Under the assumptions that an intruder accepts only exact matches, knows whether a combination is a population unique and only claims a match to a population unique, there are three possibilities for a prammed record:

1. the record has become a non-existent individual, so that a true match is impossible;
2. the record has turned into an existing population unique, in which case the match will be false;
3. the record has turned into a record which is not a population unique, for which no match will be claimed.

Imputed and non-applicable values originally present in the dataset were not considered for pramming. Imputed values were not prammed because we assume that there is sufficient protection from the uncertainty due to imputation. Non-applicable values were not prammed to avoid creating inconsistencies within a record.

To control the transitions of the PRAM method, Weight Matrices as defined in section 2.2 were used to avoid undesirable combinations. We preserved the univariate distributions exactly by systematically controlling the number of records moving from one category to another.

Multivariate distributions were controlled by pramming a variable within strata defined by a set of other (control) variables. A problem occurs when the subset of high risk records defined by a stratum becomes too small. PRAM only changes a category *i* to a category *j* value that appears within the stratum. Strata with too few values or records do not offer enough options for the transitions and may result in undesirable transitions. To avoid this we used broad categories to define some of the control variables and limited the number of control variables used.

PRAM was least successful for variables lowest in the sequence of application. Since only one variable was prammed for each record, relatively few records remained for the variables low down in the sequence. No control variables were used, and higher proportions remained unchanged.

3.2 After PRAM

After applying PRAM to the individual SAR we ran a series of edits. The edits were run to ensure that no invalid or extreme combinations had been created. Although the process of data perturbation had been controlled as much as possible some inconsistencies between variables could occur, particularly it they had not been used as control variables in the stratification. Edits were used, for example to check the consistency between age and work variables. Records which failed edits were adjusted. We found that only a very small number of records failed due to the perturbation applied.

As mentioned above, the individual SAR records containing imputed information were flagged. We used the same flag to indicate whether a record had been subject to PRAM. This informs the user that the value is not obtained directly from a true response, but does not allow them to distinguish between the two processes. Therefore if an intruder comes across a flagged record they do not know whether it is a true value, perturbed or imputed. No transition matrices were released.

4 Effect on Data Quality – Measuring the information loss.

It is important to examine the effect that the perturbation has on data quality so that users can take this effect into account when conducting their analyses. For the individual SAR we examined three aspects of data quality:

1. The invariance property – preservation of the univariate frequencies. This was checked by:
   - looking at univariate frequency tables pre- and post-PRAM
   - looking at the transitions (cross frequency between original variable and prammed variable)
2. Preserving the multivariate frequencies between prammed variables and variables used as controls
3. Preserving the relationship between prammed variables and non-prammed variables.
   (a) The number of records which failed the edit checks
   (b) The assessment of the damage, or information loss, on the distribution between variables involved and not involved in the PRAM process was measured by comparing tables before and after PRAM and by comparing the impact of pramming relative to the sampling error4.

[4] If we approximate the sampling process as simple random sampling with replacement, the relative sampling error for a cell is given by:
$RSE(\bar{N}_c) = \frac{\sqrt{Var(\bar{N}_c)}}{E(\bar{N}_c)} = \sqrt{\frac{(1-p_c)}{np_c}}$ where $\bar{N}_c$ is the estimated population total of individuals in category *c*, *n* is the sample size (of the SAR) and p_c the probability that a population member falls in this category. In practice this can be estimated by replacing p_c with the observed proportion $\bar{p}_c$ of cases falling in category *c*

The added error due to PRAM was measured as the relative absolute difference between perturbed and unperturbed cell estimates. The ratio between the PRAM error and the relative sampling error was calculated for each cell. This provided some assessment of the additional variance due to PRAM.

The conclusions were:

1. The univariate distributions for prammed variable were not damaged. This means that the optimisation process and the selection process worked well.
2. The multivariate distributions between variables involved in the PRAM process (prammed and control variables) worked well too. Very little difference was observed in cells of prammed variables by control variables.
3. Frequency table based information loss.

(a) Very few perturbed values failed subsequent edits.

(b) When the ratio between the relative error due to pram and relative sampling error is lower than 1 the additional error due to PRAM can be considered as acceptable.

The table below summarises the results of examining information loss due to the perturbation. We measure the ratio between the error due to PRAM and the sampling error for 2891 cells from 15 tables. Tables were created to reflect variable combinations that were important to data users. A ratio of greater than 1 indicates that the additional error due to PRAM is greater than the relative sampling error for that cell. Table 1 shows the percentage of cells with a ratio of greater than 1 and greater than 2. The effect of perturbation relative to sample error decreases as the cell size increases. Thus the damage done by PRAM is greater for cells with low frequencies.

	Cell Frequency Before PRAM								
	0-5	6-10	11-20	21-40	41-90	91-150	150-500	500+	**Total**
Percentage of cells with a ratio >1	35	25	24	13	15	10	17	10	16
Percentage of cells with a ratio >2	9	8	6	4	5	4	7	4	5

Table 1 Percentage of Cells across all tables with a ratio of the error due to PRAM and the sampling error of greater than 1 and 2

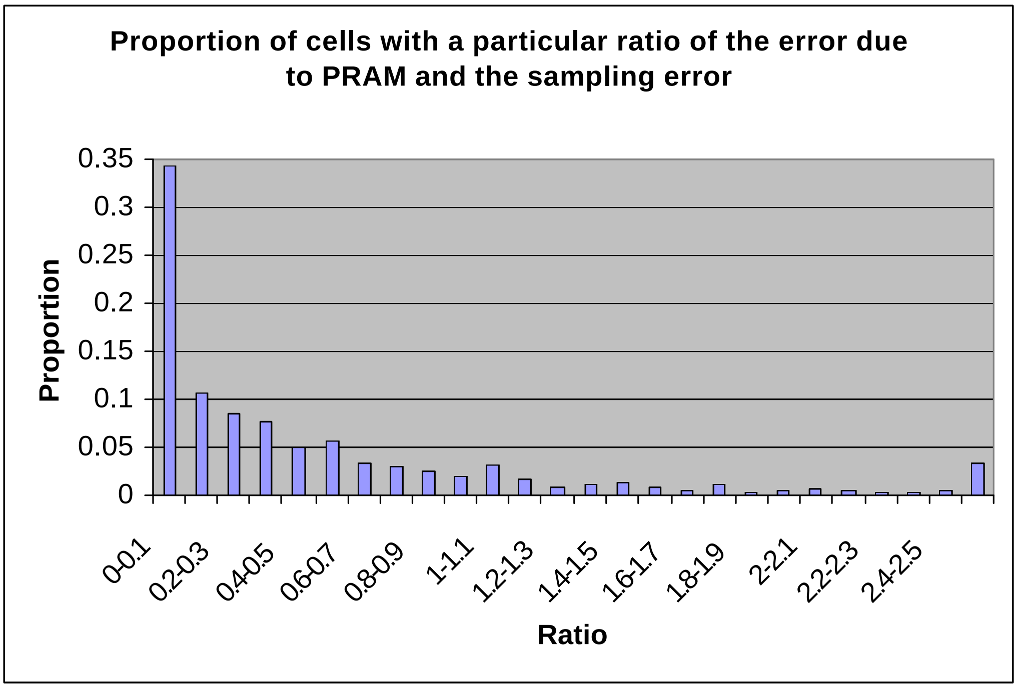

Figure 1 Bar chart showing the number of cells with a particular ratio of the error due to PRAM and the sampling error, across all tables.

Figure 1 shows the distribution of the ratio across all cells. The majority (84%) of the cells in our test have a ratio of less than 1. Just over one third have a ratio of less than 0.1, and for these cells the additional error due to PRAM is negligible. The proportion of cells with a ratio of 2 or more was small at just 5% of the total cells.

These are some basic tests that we conducted to look at the effect PRAM has on data quality. Further work should be carried out to look in more detail at the effect the PRAM has on typical analyses that users are likely to conduct.

5 Conclusions and Future work

The PRAM methodology we developed was adapted from the original PRAM in order to meet ONS objectives and constraints when releasing microdata. The purpose of PRAM is to perturb data without damaging the statistical properties. The difficulty resides in finding the right balance between safety and damage.

We feel that the PRAM methodology used on the 2001 Individual SAR worked wel1. The use of recoding allowed us to reduce the number of high risk records to a small percentage (4%). We were then able to target the perturbation only to those records which we considered to be of high risk. We were also able to preserve some of the most important multivariate distributions in the datasets through the use of stratification and preserve exactly univariate distributions of the variables being prammed. This method has enabled us to minimise the amount of damage done to the file but still achieve the required level of protection.

User groups were consulted about the perturbation process and their views incorporated in deciding the preferred moves between variable categories and the most important multivariate distributions to control for. We have provided as much information as possible about the perturbative method applied, although we have not released the values of the transition matrix. Users are therefore not able to adjust analyses or calculate the impact of added noise or extra variance.

Below are some conclusions from implementing the new PRAM methodology on the individual SAR:

1. As used in this context on targeted records, PRAM is an efficient method of data perturbation, which is well controllable.
2. The selection of the targeted record depends on the information provided by the risk model. The special uniques analysis provided rich information on the source of the risk since it gives for each sample unique record the variable/value combination responsible for the potential uniqueness of the record. This was used to decide which variable to PRAM for each high risk record.
3. The actual selection of records to be prammed was controlled in order to preserve the exact frequencies of the categories after the perturbation.
4. Stratification was used to control relationships between variables. Careful choice of control variables was needed to avoid creation of strata that were too small to carry out PRAM effectively.
5. Transition weights were used to control the changes affecting each variable considered for PRAM.
6. PRAM may still create inconsistent combinations. A very small number of records failed the edits due to the perturbation applied. These records were adjusted.
7. High risk records that are not changed through the perturbation stay risky in principle. However these records would still be flagged. If a user comes across a flagged record they cannot distinguish between a true value, a perturbed value or an imputed value.

Further work should be conducted to investigate the effect that PRAM has on data quality. Although we have conducted some investigation into this, it should be expanded to investigate to look at the type of typical analyses that users conduct on the microdata released.

Applying PRAM to a small proportion of the file has allowed us to strike a good balance between recoding and minimising the damage from perturbation.

References

De Wolf, P.-P., Gouweleeuw, J.M., Kooiman, P., and Willenborg, L.C.R.J (1997). Reflections on PRAM, Research paper no. 9742, Voorburg/ Heerlen: Statistics Netherlands.

De Wolf, P.P., Gouweleeuw, J.M., Kooiman, K. and Willenborg, L.C.R.J. (1998), Reflections on PRAM, proceedings of the conference "Statistical Data Protection", March 25-27 1998, Lisbon, Portugal.

De Wolf, P.-P. and Van Gelder, I (2004). An empirical evaluation of PRAM, Discussion paper 04012, Voorburg/ Heerlen, September 2004: Statistics Netherlands.

Elliot, M.J and Manning, A (2004) The methodology used for the 2001 SARs Special Uniques Analysis, University of Manchester.

Gouweleeuw, J.M., Kooiman, K. and Willenborg, L.C.R.J. and De Wolf, P.P. (1998a), Post Randomisation for Statistical Disclosure Control: Theory and Implementation, Journal of Official Statistics, Vol.14, 4, pp. 463-478.

Gross, B, Guiblin, P and Merrett, K (2004) Risk Assessment of the Individual Sample of Anonymised Records (SAR) from the 2001 Census**,** Office for National Statistics. http://www.ccsr.ac.uk/sars/events/2004-09-30/slides/index.html

Kooiman, P., Willenborg, L.C.R.J., and Gouweleeuw, J.M. (1997). PRAM: a method for disclosure limitation of microdata, Research paper 9705, Voorburg/Heerlen: Statistics Netherlands.

Kuha, J., and Skinner, C. (1997). Categorical data analysis and misclassification, in *Survey Measurement and Process Quality*, (L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C.Dippo, N. Schwarz, and D. Trewin, eds.), New York: Wiley.

Skinner, C.J and Elliot, M.J. (2002): A Measure of Disclosure Risk for Microdata, Journal of the Royal Statistical Society B, 64, 4, pp. 855-867

Willenborg, L.C.R.J., and De Waal, T. (2001). Elements of Statistical Disclosure Control, New York: Springer.