RELEASING MICRODATA: DISCLOSURE RISK ESTIMATION, DATA MASKING AND ASSESSING UTILITY

Natalie Shlomo[1]
[1]Southampton Statistical Sciences Research Institute, University of Southampton, SO17 1BJ, UK
N.Shlomo@soton.ac.uk

Abstract

Statistical Agencies need to make informed decisions when releasing sample microdata from social surveys with respect to the level of protection required in the data according to the mode of access. These decisions should be based on objective quantitative measures of disclosure risk and data utility. We assume microdata that contain individuals investigated in a survey and the population is unknown. Disclosure risk is a function of both the population and the sample counts in cells of a contingency table spanned by identifying discrete key variables, i.e. place of residence, sex, age, occupation, etc. Disclosure risk measures are estimated using probabilistic modeling. Based on the disclosure risk assessment, appropriate Statistical Disclosure Limitation (SDL) methods are chosen depending on the access mode, user requirements and the contents of the data. Information loss measures are defined to quantify the effects of SDL methods on statistical analysis. We demonstrate a Disclosure Risk-Data Utility assessment on a sample drawn from a Census where the population is known and can be used to validate procedures.

Key Words: Log-linear models; Goodness of fit; Measurement error; Additive noise; Micro-aggregation; Random rounding; PRAM; Information loss

1. Introduction

Statistical Agencies release sample microdata from social surveys under different modes of access. Access methods range from Public Use Files (PUF) in the form of tables or highly perturbed datasets to Microdata Under Contract (MUC) for researchers and licensed institutions where levels of protection are less severe. Statistical Agencies also often have on-site datalabs where registered researchers can access unperturbed statistical data. Microdata Review Panels (MRP) need to make informed decisions when releasing microdata based on objective disclosure risk measures, and set tolerable risk thresholds according to the access mode. They also provide quality guidelines and initial rules for data masking based on recoding variables.

We assume that the microdata contain individuals investigated in a survey and the population is unknown (or only partially known through some marginal distributions). The disclosure risk is a function of both the population and the sample, and in particular the cell counts of a contingency table defined by combinations of identifying discrete key variables, i.e. place of residence, sex, age, occupation, etc. Using probabilistic models, we estimate per-record disclosure risk measures which can be used to target high-risk records for Statistical Disclosure Limitation (SDL) techniques. Consistent global file-level disclosure risk measures are aggregated from per-record risk measures. Global risk measures are used by MRPs to inform decisions on the release of microdata. Section 2 provides an overview of disclosure risk assessment for sample microdata using probabilistic modeling.

Based on the disclosure risk assessment, Statistical Agencies must choose appropriate SDL methods either by perturbing, modifying, or summarizing the data. The choice of the SDL method depends on the access mode, requirements of the users and the impact on quality and information loss. Choosing an optimal SDL method is an iterative process where a balance must be found between managing disclosure risk and preserving the utility in the microdata. SDL methods for microdata include perturbative methods that alter the data and non-perturbative methods which limit the amount of information released. Each SDL method impacts differently on information loss and they should be combined and optimized to preserve the consistency and integrity of the perturbed microdata. In Section 3, we present improvements to some standard perturbative SDL methods for sample microdata. In Section 4, we define information loss measures to quantify the effects of SDL methods on bias and variance and other statistical analysis tools. In Section 5, we demonstrate the Disclosure

Risk-Data Utility assessment on sample data drawn from a Census (where the population is known). Section 6 concludes with a discussion.

2. Disclosure Risk Assessment

Identifying key variables for disclosure risk assessment are determined by a disclosure risk scenario, i.e. assumptions about available external files and IT tools that can be used by intruders to identify individuals in released microdata. For example, key variables may be chosen which allow linking the released microdata to a publicly available file containing names and addresses. Disclosure risk is assessed on the contingency table of counts spanned by these identifying key variables. The other variables in the file are sensitive variables.

Some methods for assessing disclosure risk rely on heuristics to identify special uniques on a set of cross-classified key variables, i.e. sample uniques that are likely to be population uniques (see Elliot, et al., 2005, Skinner and Elliot, 2002 and references therein) and probabilistic record linkage (see Yancey, Winkler, and Creecy, 2002, Domingo-Ferrer and Torra, 2003 and references therein). A drawback of these methods is that they do not take into account the protection afforded by the sampling and inconsistent record level and global level disclosure risk measures. We assess disclosure risk using probabilistic modeling.

We consider individual per-record risk measures in the form of a probability of re-identification. These per-record risk measures are aggregated to obtain global risk measures for the entire file. Let F_k be the population size in cell k of a table spanned by key variables having *K* cells and f_k the sample size. Also, $\sum_{k=1}^{K} F_k = N$ and $\sum_{k=1}^{K} f_k = n$. We focus our attention on the set of sample uniques, $SU = \{k : f_k = 1\}$ since these are potential high-risk records, i.e. population uniques. Two global disclosure risk measures (where *I* is the indicator function) are the following:

1. Number of sample uniques that are population uniques: $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$
2. Expected number of correct matches for sample uniques (i.e., a matching probability) $\tau_2 = \sum_k I(f_k = 1) 1 / F_k$.

The individual risk measure for τ_2 is $1/F_k$. This is the probability that a match between a record in the microdata and a record in the population having the same values of key variables will be correct. If for example, there are two records in the population with the same values of key variables, the probability is 0.5 that the match will be correct. Adding up these probabilities over the sample uniques gives the expected number (on average) of correctly matching a record in the microdata to the population when we allow guessing. We assume that population frequencies F_k are unknown and estimate from a probabilistic model the risk measures by:

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(F_k = 1 \mid f_k = 1) \quad \text{and} \quad \hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(1/F_k \mid f_k = 1) \qquad (1)$$

Skinner and Holmes (1998) and Elamir and Skinner (2006) propose a Poisson Model to estimate disclosure risk measures. In this model, they assume the natural assumption in contingency table literature: $F_k \sim Poisson(\lambda_k)$ for each cell *k*. A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell *k*: $f_k \mid F_k \sim Bin(F_k, \pi_k)$. It follows that:

$$f_k \sim Pois(\pi_k \lambda_k) \text{ and } F_k \mid f_k \sim Poisson(\lambda_k(1-\pi_k)) \qquad (2)$$

where $F_k \mid f_k$ are conditionally independent.

The parameters $\{\lambda_k\}$ are estimated using log-linear modeling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k \lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = \mathbf{x}'_k \beta$ where $\mathbf{x}_k$ is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood (MLE) estimator $\hat{\beta}$ may be obtained by solving the score equations:

$$\sum_k [f_k - \pi_k \exp(\mathbf{x}'_k \beta)]\mathbf{x}_k = 0 \qquad (3)$$

The fitted values are calculated by: $\hat{u}_k = \exp(\mathbf{x}'_k \hat{\beta})$ and $\hat{\lambda}_k = \hat{u}_k / \pi_k$.
Individual disclosure risk measures for cell k are:

$$P(F_k = 1 | f_k = 1) = \exp(\lambda_k (1 - \pi_k)),$$

$$E(1/F_k \mid f_k = 1) = [1 - \exp(\lambda_k (1 - \pi_k))]/[\lambda_k (1 - \pi_k)] \qquad (4)$$

Plugging $\hat{\lambda}_k$ for λ_k in (4) leads to the estimates $\hat{P}(F_k = 1 \mid f_k = 1)$ and $\hat{E}[1/F_k \mid f_k = 1]$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1). Rinott and Shlomo (2007b) consider confidence intervals for these global risk measures.

Skinner and Shlomo (2008) develop a method for selecting the log-linear model based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. Defining $h(\lambda_k) = P(F_k = 1 \mid f_k = 1)$ for τ_1 and $h(\lambda_k) = E(1/F_k \mid f_k = 1)$ for τ_2, they consider the expression: $B = \sum_k E[I(f_k = 1)][h(\hat{\lambda}_k) - h(\lambda_k)]$

A Taylor expansion of h leads to the approximation

$$B \approx \sum{}_k \pi_k \lambda_k \exp(-\lambda_k)[h'(\lambda_k)(\hat{\lambda}_k - \lambda_k) + h''(\lambda_k)(\hat{\lambda}_k - \lambda_k)^2 / 2]$$

and the relations $Ef_k = \pi_k \lambda_k$ and $E[(f_k - \pi_k \hat{\lambda}_k)^2 - f_k] = \pi_k{}^2 E(\lambda_k - \hat{\lambda}_k)^2$ under the hypothesis of a Poisson fit lead to a further approximation of B of the form

$$\hat{B} \approx \sum{}_k \hat{\lambda}_k \exp(-\pi_k \hat{\lambda}_k)[-h'(\hat{\lambda}_k)(f_k - \pi_k \hat{\lambda}_k) + h''(\hat{\lambda}_k)[(f_k - \pi_k \hat{\lambda}_k)^2 - f_k]/(2\pi_k)] \quad (5)$$

The method selects the model using a forward search algorithm which minimizes the standardized bias estimate $\hat{B}_i / \sqrt{\hat{v}_i}$ for $\hat{\tau}_i$, $i = 1,2$ where $\hat{v}_i$ are variance estimates of $\hat{B}_i$.

In the simple case where the sample is drawn under simple random sampling, $\pi_k = \pi = n/N$. Skinner and Shlomo (2008) address the estimation of disclosure risk measures under complex survey designs with stratification, clustering and survey weights. In this case, while the method assumes that all individuals within cell k are selected independently using Bernoulli sampling, i.e. $P(f_k = 1 \mid F_k) = F_k \pi_k (1 - \pi_k)^{F_k - 1}$, this may not be the case when sampling clusters (households). In practice, key variables typically include variables such as age, sex and occupation, and tend to cut across clusters. Therefore the above assumption holds in practice in most household surveys and does not cause bias in the estimation of the risk measures. Inclusion probabilities may vary across strata, the most common stratification is on geographies. Strata indicators should always be included in the key variables to take into account differential inclusion probabilities. Under complex sampling, the $\{\lambda_k\}$ can be estimated consistently using pseudo-maximum likelihood estimation (Rao and Thomas, 2003), where the estimating equation in (3) is modified as:

$$\sum_k [\hat{F}_k - \exp(x'_k \beta)]x_k = 0 \qquad (6)$$

and $\hat{F}_k$ is obtained by summing the survey weights in cell k: $\hat{F}_k = \sum_{i \in k} w_i$.

The resulting estimates $\{\hat{\lambda}_k\}$ are plugged into expressions in (4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k / \hat{F}_k$. Note that the risk measures in (4) only depend on sample uniques and the value of $\hat{\pi}_k$ in this case is simply the reciprocal of the survey weight. The test criteria $\hat{B}$ is also adapted to the pseudo-maximum likelihood method.

The probabilistic model presented as well as other probabilistic methods (see Bethlehem, Keller, and Pannekoek, 1990, Benedetti, Capobianchi, and Franconi, 1998, Rinott and Shlomo 2006, 2007a) assume that there is no measurement error in the way the data is recorded. Besides typical errors in data capture, key variables can also purposely be misclassified as a means of masking the data, for example through record swapping or PRAM. Skinner and Shlomo (2007) adapt the estimation of risk measures to take into account measurement errors. Denote the cross-classified key variables by X and assume that X in the microdata has undergone some misclassification or perturbation error denoted by the value $\tilde{X}$. Assume that the values of X in the population are fixed and suppose the values of $\tilde{X}$ for the records in the microdata are determined independently by a misclassification matrix *M*,

$$M_{kj} = P(\tilde{X} = k \mid X = j) \tag{7}$$

The per-record disclosure risk measure of a match with a sample unique under measurement error is:

$$\frac{M_{kk}(1-\pi M_{kk})}{\sum_j F_j M_{kj} /(1-\pi M_{kj})} \le \frac{1}{F_k} \tag{8}$$

Under assumptions of small sampling fractions and small misclassification errors, the measure can be approximated by: $M_{kk} / \sum_j F_j M_{kj}$ or $M_{kk} / \tilde{F}_k$ where $\tilde{F}_k$ is the population count with $\tilde{X} = k$.

Aggregating the per-record disclosure risk measures, the global risk measure is:

$$\tau_2 = \sum_k I(f_k = 1) M_{kk} / \tilde{F}_k \tag{9}$$

Note that to calculate the measure only the diagonal of the misclassification matrix needs to be known, i.e. the probabilities of not being perturbed. Population counts are generally not known so the estimate in (9) can be obtained by probabilistic modeling on the misclassified sample:

$$\hat{\tau}_2 = \sum_k I(\tilde{f}_k = 1) M_{kk} \hat{E}\left(1/\tilde{F}_k \mid \tilde{f}_k\right) \tag{10}$$

3. Statistical Disclosure Limitation Methods for Sample Microdata

Depending on the outcome of the individual and global risk measures, SDL methods may need to be applied. Thresholds are set for releasing the microdata depending on the mode of access. SDL techniques for microdata include perturbative methods which alter the data and non-perturbative methods which limit the amount of information released without actually altering the data. Examples of non-perturbative SDC techniques are global recoding, suppression of values or variables and sub-sampling records (see Willenborg and De Waal, 2001). Perturbative methods for continuous variables include adding random noise (Kim, 1986, Fuller, 1993, Brand, 2002, Yancey, Winkler and Creecy, 2002), micro-aggregation (replacing values with their average within groups of records) (Defays and Nanopoulos, 1992, Anwar 1993, Domingo-Ferrer and Mateo-Sanz, 2002), rounding to a pre-selected rounding base, and rank swapping (swapping values between pairs of records within small groups) (Dalenius and Reiss, 1982, Fienberg and McIntyre, 2005). Perturbative methods for categorical variables include record swapping (typically swapping geography variables) and a more general post-randomization probability mechanism (PRAM) where categories of variables are changed or not changed according to a prescribed probability matrix and a stochastic selection process (Gouweleeuw, et al. 1998). For more information on these methods see also: Willenborg and De Waal, 2001, Gomatam and Karr, 2003, Domingo-Ferrer, Mateo-Sanz, and Torra, 2001, and references therein.

With non-perturbative SDL methods, the logical consistency of the records remain unchanged. Perturbative methods, however, alter the data, and therefore we can expect consistent records to start failing edit rules due to the perturbation. Edit rules, or edits for short, describe either logical relationships that have to hold true, such as “a two-year old person cannot be married” or “the profit and the costs of an enterprise should sum up to its turnover”, or relationships that have to hold true in most cases, such as “a 12-year old girl cannot be a mother”. Shlomo and De Waal, 2008 discuss methods for perturbing sample microdata which preserve the logical consistencies and minimize information loss. The following is a brief summary of some of the methods:

3.1 Additive noise

Additive noise is an SDL method that is carried out on continuous variables. In its basic form random noise is generated independently and identically distributed with a positive variance and a mean of zero. The random noise is then added to the original variable. Adding random noise will not change the mean of the variable for large datasets but will introduce more variance. This will impact on the ability to make statistical inferences. Researchers may have suitable methodology to correct for this type of measurement error but it is good practice to minimize these errors through better implementation of the method.

Additive noise should be generated within small homogenous sub-groups (for example, percentiles of the continuous variable) in order to use different initiating perturbation variance for each sub-group. Generating noise in sub-groups also causes less edit failures with respect to relationships in the data. A better technique is to add correlated random noise to the continuous variable thereby ensuring that not only means are preserved but also the exact variance. A simple method for generating correlated random noise for a continuous variable *z* is as follows:

Procedure 1 (univariate): Define a parameter δ which takes a value greater than 0 and less than equal to 1. When $\delta = 1$ we obtain the case of fully modeled synthetic data. The parameter δ controls the amount of random noise added to the variable *z*. After selecting a δ, calculate: $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$. Now, generate random noise ε independently for each record with a mean of $\mu' = \frac{1-d_1}{d_2}\mu$ and the original variance of the variable σ^2 . Typically, a Normal Distribution is used to generate the random noise. Calculate the perturbed variable z_i' for each record *i* in the sample microdata (*i=1,..,n*) as a linear combination: $z_i' = d_1 \times z_i + d_2 \times \varepsilon_i$. Note that $E(z') = d_1 E(z) + d_2[\frac{1-d_1}{d_2} E(z)] = E(z)$ and $Var(z') = (1-\delta^2)Var(z) + \delta^2 Var(z) = Var(z)$ since the random noise is generated independently to the original variable *z*.

An additional problem when adding random noise is that there may be several variables to perturb at once, and these variables may be connected through an edit constraint of additivity. If we were to perturb each variable separately, this edit constraint would not be guaranteed. One procedure to preserve additivity would be to perturb two of the variables and obtain the third from aggregating the perturbed variables. However, this method will not preserve the total, mean and variance of the aggregated variable and in general, it is not good practice to compound effects of perturbation (i.e., aggregate perturbed variables) since this causes unnecessary information loss.

We propose Procedure 1 in a multivariate setting where we add correlated noise to the variables simultaneously. The method not only preserves the means of each of the three variables and their co-variance matrix, but also preserves the edit constraint of additivity.

Procedure 1 (multivariate): Consider three variables x, y and z where $x + y = z$. This procedure generates random noise that a priori preserves additivity and therefore combining the random noise to the original variables will also ensure additivity. In addition, means and the covariance structure are preserved. The technique is as follows:

Generate multivariate random noise: $(\varepsilon_x, \varepsilon_y, \varepsilon_z)^T \sim N(\mu', \Sigma)$, where the superscript T denotes the transpose. In order to preserve sub-totals and limit the amount of noise, the random noise should be generated within percentiles (note that we drop the index for percentiles). The vector $\boldsymbol{\mu}'$ contains the corrected means of each of the three variables x, y and z based on the noise parameter δ: $\mu'^{\mathrm{T}} = (\mu_x', \mu_y', \mu_z') = (\frac{1-d_1}{d_2}\mu_x, \frac{1-d_1}{d_2}\mu_y, \frac{1-d_1}{d_2}\mu_z)$. The matrix $\boldsymbol{\Sigma}$ is the original covariance matrix. For each separate variable, calculate the linear combination of the original variable and the random noise as previously described. For example, for record *i*: $z_i' = d_1 \times z_i + d_2 \times \varepsilon_{zi}$. The mean vector and the covariance matrix remain the same before and after the perturbation, and the additivity is exactly preserved.

3.2 Micro-aggregation

Micro-aggregation is another SDL technique for continuous variables. Records are grouped together in small groupings of size *p*. For each individual in a group *k*, the value of the variable is replaced with the group average. This method can be carried out for both a univariate or multivariate setting where the latter can be implemented through sophisticated computer algorithms. Replacing values of variables with their average in a small group will not generally initiate inconsistencies in the data, such as the relationship between variables, although there may be problems at the boundaries of such edits. When carrying out micro-aggregation simultaneously on several variables within a group, additivity constraints will also be preserved since the sum of the means of two variables will equal the mean of the total variable in a grouping. The focus therefore for minimizing information loss is on the preservation of variances.

Micro-aggregation preserves the mean (and the overall total) of a variable *z* but will lead to a decrease in the variance. This is because the total variance can be decomposed into a "within" group variance and a "between" group variance. When implementing micro-aggregation and replacing values by the average of their group, only the "between" variance remains. In practice, there may be little decrease in the variance since the size of the groups is small. In order to minimize information loss due to a decrease in the variance, we generate random noise according to the magnitude of the difference between the total variance and the "between" variance, and add it to the micro-aggregated variable. Besides raising the variance back to its original level, this method will also result in extra protection against the risk of re-identification since micro-aggregation in some cases can easily be deciphered (see Winkler, 2002). The combination of micro-aggregation and additive random noise is discussed in Oganian and Karr, 2006. When adding random noise to several micro-aggregated variables that are connected through an additivity constraint, we can apply a straight-forward linear programming technique to preserve the additivity.

3.3 Unbiased Random Rounding

Rounding to a predefined base is a form of adding noise, although in this case the exact value of the noise is known a priori and is controlled via the rounding base. As in micro-aggregation, it is unlikely that inconsistencies will result when rounding the data. However, rounding continuous variables separately may cause additivity edit failures since the sum of rounded variables will not necessarily equal their rounded total. In addition, summing rounded values will not equal their rounded total and large discrepancies can occur. We demonstrate a method for preserving totals when carrying out an unbiased random rounding procedure on a continuous variable.

Rounding procedures are relatively easy to implement. In this example, we describe a one-dimensional random rounding procedure for a variable which not only has the property that it is stochastic and unbiased, but also preserves the overall total (and hence the mean) of the variable being rounded. Moreover, the strategy that we propose reduces the extra variance induced by the rounding. The algorithm is as follows:
Let *m* be the value to be rounded and let $Floor(m)$ be the largest multiple *k* of the base *b* such that $bk < m$. In addition, define the residual of *m* according to the rounding base *b* by $res(m) = m - Floor(m)$. For an unbiased random rounding procedure, *m* is rounded up to $(Floor(m) + b)$ with probability $res(m)/b$ and rounded down to $Floor(m)$ with probability $(1 - res(m)/b)$. If *m* is already a multiple of *b*, it remains unchanged. The expected value of the rounded value is the original value. The rounding is usually implemented "with replacement" in the sense that each value is rounded independently, i.e. a random uniform number *u* between 0 and 1 is generated for each value. If $u < res(m)/b$ then the entry is rounded up, otherwise it is rounded down. In order to preserve the exact total of the variable being rounded, we define a simple algorithm for selecting "without replacement" the values that are rounded up and the values that are rounded

down: for those entries having $res(m)$, randomly select a fraction of $res(m)/b$ of the values and round upwards, the rest of the values round downwards. Repeat this process for all $res(m)$. As mentioned, similar to the case of simple random sampling with and without replacement, this selection strategy reduces the additional variance caused by the rounding.

The rounding procedure should be carried out within sub-groups in order to benchmark important totals. This may, however, distort the overall total across the entire dataset. Users are typically more interested in smaller sub-groups for analysis and therefore preserving totals for sub-groups is generally more desirable than the overall total. Reshuffling algorithms can be applied for changing the direction of the rounding for some of the values across the records in order to preserve additivity constraints and the overall totals.

3.4 Protecting Categorical Variables by PRAM

As presented in Shlomo and De Waal (2008), we examine the use of a technique called the Post-randomization Method (PRAM) (Gouweleeuw, et al., 1998) to perturb categorical variables. This can be seen as a general case of a more common technique based on record swapping. Willenborg and De Waal (2001) describe the process as follows:
Let $\mathbf{P}$ be a $L \times L$ transition matrix containing conditional probabilities $p_{ij} = p(\text{perturbed category is } j \mid \text{original category is } i)$ for a categorical variable with L categories, $\mathbf{t}$ the vector of frequencies and $\mathbf{v}$ the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/n$, where *n* is the number of records in the micro-data set. In 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. $\mathbf{t}^*$ is a random variable and $\mathrm{E}(\mathbf{t}^* \mid \mathbf{t}) = \mathbf{tP}$. Assuming that the transition probability matrix $\mathbf{P}$ has an inverse $\mathbf{P}^{-1}$, this can be used to obtain an unbiased moment estimator of the original data: $\hat{\mathbf{t}} = \mathbf{t}^*\mathbf{P}^{-1}$. 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.

We can place 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 $\mathbf{t}$. To obtain an invariant transition matrix, we calculate a matrix $\mathbf{Q}$ 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. The invariant matrix is obtained by $\mathbf{R} = \mathbf{PQ}$. The property of invariance means that the vector of the original frequencies $\mathbf{v}$ is an eigenvector of $\mathbf{R}$. The invariant matrix $\mathbf{R}$ may distort the desired probabilities on the diagonal, so we define a parameter α and calculate $\mathbf{R}^* = \alpha\mathbf{R} + (1-\alpha)\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. $\mathbf{R}^*$ will also be invariant and the amount of perturbation is controlled by the value of α. The property of invariance means that the expected values of the marginal distribution of the variable being perturbed are preserved. In order to obtain the exact marginal distribution and reduce the additional variance caused by the perturbation, we propose using a “without” replacement selection strategy to choose values to perturb based on the expectations calculated from the transition probabilities (see Section III.C for the case of random rounding). This method was used to perturb the Sample of Anonymized Records (SARs) of the 2001 UK Census (Gross, Guiblin and Merrett, 2004).

As in most perturbative SDL methods, joint distributions between perturbed and unperturbed variables are distorted, in particular for variables that are highly correlated with each other. If no controls are taken into account in the perturbation process, edit failures may occur resulting in inconsistent and “silly” combinations. Controlling the perturbation can be carried as follows:

1. Before applying PRAM, the variable to be perturbed is divided into subgroups, $g = 1,...,G$. The transition (and invariant) probability matrix is developed for each subgroup *g*, R_g. The transition matrices for each subgroup are placed on the main diagonal of the overall final transition matrix

where the off diagonal probabilities are all zero, i.e. the variable is only perturbed within the subgroup and the difference in the variable between the original value and the perturbed value will not exceed a specified level. An example of this is perturbing *age* within broad age bands.

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

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

4. Information Loss Measures

The utility of microdata that has undergone SDL techniques is based on whether statistical inference can be carried out and the same analysis and conclusions drawn on the perturbed data compared to the original data. This depends on user requirements and the types of analysis. In general, microdata is multi-purposed and used by many different users. Therefore, we use proxy measures to assess the utility based on assessing distortions to distributions and the impact on bias, variance and other statistical analysis tools (Chi-squared statistic, R^2 goodness of fit, rankings, etc.). Shlomo, 2007 and Shlomo and Young, 2006 describe the use of such measures for assessing information loss in perturbed statistical data. A brief summary of some useful proxy measures are the following:

4.1 Distance Metrics

Distance metrics are used to measure distortions to distributions as a result of applying SDL methods. We apply these measures on distributions calculated from the perturbed microdata. Some useful metrics for aggregated data were presented in Gomatam and Karr, 2003.

Let *D* represent a frequency distribution produced from the microdata and let $D(c)$ be the frequency in cell *c*. Two useful distance metrics are:

- Average Absolute Distance per Cell:

$$AAD(D_{orig}, D_{pert}) = \sum_{c} | D_{pert}(c) - D_{orig}(c) | / n_c \qquad (11)$$

where n_c is the number of cells in the distribution

- Kolmogorov-Smirnov Two- Sample Test Statistic:

For unweighted data, the empirical distribution of the original values is defined as: $D_{orig}(t) = \sum_{c} I(c \le t) / n_c$ and similarly $D_{pert}(t)$ where *I* is the indicator function. The *KS* statistic is defined as:

$$KS(D_{orig}, D_{pert}) = \max_{j}(| D_{pert}(t_j) - D_{orig}(t_j) |) \qquad (12)$$

where the $\{t_j\}$ values are the n_c jointly ordered original and perturbed values of *D*.

The *AAD* is intuitive and describes the average absolute difference per cell of the distribution. The *KS* two-sample test assumes independence of the two samples and therefore the test itself is invalid. However, it is still useful to use the *KS* statistic as a relevant distance metric.

.

4.2 Impact on Measures of Association

Tests for independence are often carried out on joint frequency distributions between categorical variables that span a table calculated from the microdata. The test for independence for a two-way table is based on a Pearson Chi-Squared Statistic $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ where o_{ij} is the observed count and $e_{ij} = (n_{i.} \times n_{.j}) / n$ is the expected count for row *i* and column *j*. If the row and column are independent then χ^2 has an asymptotic chi-square distribution with (R-1)(C-1)and for large values the test rejects the null hypothesis in favor of the alternative hypothesis of association. We use the measure of association, Cramer's V: $CV = \sqrt{\frac{\chi^2 / n}{\min(R-1),(C-1)}}$ and define the information loss measure by the percent relative difference between the original and perturbed table:

$$RCV(D_{pert}, D_{orig}) = 100 \times \frac{CV(D_{pert}) - CV(D_{orig})}{CV(D_{orig})} \qquad (13)$$

For multiple dimensions, log-linear modeling is often used to examine associations. A similar measure to (13) can be calculated taking the relative difference in the deviance obtained from a model based on the original and perturbed microdata.

4.3 Impact on a Regression Analysis

For continuous variables, we assess the impact on the correlation and in particular the R^2 of a regression (or ANOVA) analysis. For example, in an ANOVA, we test whether a continuous dependent variable has the same means within groupings defined by categorical explanatory variables. The goodness of fit criterion R^2 is based on a decomposition of the variance of the mean of the dependent variable. By perturbing the statistical data, the groupings may lose their homogeneity, the "between" variance becomes smaller, and the "within" variance becomes larger. In other words, the proportions within each of the groupings shrink towards the overall mean. On the other hand, the "between" variance may become artificially larger showing more association than in the original distributions.

We define information loss based on the "between" variance of a proportion: Let $P^k_{orig}(c)$ be a target proportion k for a cell c, i.e. $P^k_{orig}(c) = \frac{D^k_{orig}(c)}{D_{orig}(c)}$ and let $P^k_{orig} = \frac{\sum_c D^k_{orig}(c)}{\sum_c D_{orig}(c)}$ be the overall proportion. The "between" variance is defined as: $BV(P^k_{orig}) = \frac{1}{n_c - 1} \sum_c (P^k_{orig}(c) - P^k_{orig})^2$

and the information loss measure is:

$$BVR(P^k_{pert}, P^k_{orig}) = 100 \times \frac{BV(P^k_{pert}) - BV(P^k_{orig})}{BV(P^k_{orig})} \qquad (14)$$

In addition, we can assess the impact on the regression coefficient for a model based on a continuous variable where the independent variable is also continuous and has undergone different methods of perturbation such as additive noise, micro-aggregation and rounding.

3. Example

We present an example of how a Statistical Agency might assess disclosure limitation strategies through a disclosure risk-data utility analysis. We use a population from the 1995 Israel Census sample composed of all individuals aged 15 and over living in 20% of the households in Israel at the time of the Census, *N*=753,711. We draw a $\pi = 1/100$ sample of individuals, *n*=7,537.

The MRP needs to assess the disclosure risk and consider SDL techniques. Initial recoding of key variables is carried out. The key variables for assessing disclosure risk are the following:
Locality Code (single codes for large localities above 10,000 inhabitants and single combined code for smaller localities) – 85 categories; Sex – 2 categories; Age groups - 15 categories; Occupation -11 categories, Income groups - 17 categories (*K*=476,850).

In addition to the initial key, the MRP might consider further perturbation of the geography variable. We examine 2 techniques: recoding and collapsing the large locality codes according to a larger geographical area and locality size (30 categories) and invariant PRAM (a general case of record swapping) on the large locality codes with 0.70 on the diagonal of the misclassification matrix. Table 1 presents a comparison of these two techniques. The true risk based on $\tau_2 = \sum_k I(f_k = 1) 1/F_k$ are given in the column headings in parenthesis. The true disclosure risk for PRAM is calculated by summing $1/F_k$ across sample uniques that were not perturbed. The estimates $\hat{\tau}_2$ in Table 1 are similar to the true values. The asymptotically normal test statistic based on (5) is given in parenthesis. Note that to estimate the disclosure risk for PRAM we used the formula in (10). The distance metrics *AAD* and *KS* for the recoded localities are calculated by imputing the average across the recoded cells. For example, if 10 localities were recoded into a single cell, each locality would receive 1/10 of the total in the cell.

Table 1: Comparison of SDL techniques: Recoding and PRAM

	Original Key ($\tau_2 = 1025.7$)	Recoded localities (30 categories) ($\tau_2 = 571.5$)	PRAM (70% on the diagonal) ($\tau_2 = 714.7$)
Disclosure Risk			
$\hat{\tau}_2$ (test statistic) Sample uniques $\hat{\tau}_2 / SU$	1015.5 (1.94) 4005 25.3%	599.9 (1.32) 3376 17.8%	729.5 (1.42) 3479 20.9%
Utility			
AAD across 85 localities	0	7.22	3.88
KS across 85 localities	0	1.53	0.46
RCV for localities (85)×occupation (11) (true=0.1370)	0	-0.33	-0.08
BVR for average income between localities (85) (true=3.082*10^9)	0	-0.44	-0.09

As can be seen in Table 1, recoding causes significantly more information loss compared to PRAM, even with 30% of the localities misclassified. The disclosure risk however is more effectively reduced with recoding than with PRAM. The MRP might consider reducing the disclosure risk further by combining the techniques, for example, by identifying those records that remain unique after the recoding and implementing PRAM on the high-risk records only. Note that both methods give negative values for *RCV* and *BVR* which reflect a loss of association and more heterogeneity as a result of the SDL techniques.

After deciding on key variables, MRPs might consider taking further action by perturbing sensitive variables, such as income. In our example, income was also used as a key variable so disclosure risk would need to be reassessed if perturbation is carried out on the income variable. We carried out three improved techniques for perturbing income from wages for those records with non-zero income (3,249 out of the 7,537 individuals in the sample): correlated additive noise, controlled random rounding to base 10 and micro-aggregation (size of groups=10) with additive noise. All three techniques preserve the mean and its variance of the original variable. Results are given in Table 2.

Table 2 shows conflicting results for the two distance metrics. While micro-aggregation with additive noise has more perturbation per cell compared to the other methods, correlated noise has more distance between the empirical distributions based on the original and perturbed variable. The controlled random rounding has the smallest distance metrics and not surprisingly the lowest amount of records that switch out of their original income group. Table 2 also shows that the utility in the data with respect to some common statistical analysis tools is preserved and this is due to the improvements in the implementation of the SDL techniques which aim to preserve sufficient statistics.

Table 2: Information loss measures for income from wages after perturbation for individuals with non-zero income

	Correlated Noise	Controlled Random Rounding Base 10	Micro-aggregation and Additive Noise
Utility			
AAD across 17 income groups	17.50	2.00	23.25
KS across 17 income groups	1.05	0.66	0.87
RCV for income groups (17)×occupation (11) (true=0.1736)	0.01	0	0
BVR average income between localities (85) (true=$3.082*10^9$)	-0.01	0	0.01
Percentage of records switching income groups	17.4%	0.8%	12.5%

4. Discussion

In this paper, we focus on how a Statistical Agency might carry out a disclosure risk-data utility analysis to inform decisions about the release of sample microdata. The main conclusions of the paper are: (1) the need for a reliable method for objectively assessing disclosure risk; (2) SDL techniques should be optimized and combined to ensure utility in the perturbed microdata.

Statistical Agencies generally release same sets of microdata on a yearly basis but the disclosure risk-data utility analysis need not be repeated every year if no significant changes are applied to the microdata. Therefore, it is recommended that time and resources be spent at least once on an in-depth analysis for ensuring high quality microdata with tolerable risk thresholds for each mode of access.

Distributing different sets of the same microdata may be a cause for concern since different versions of the microdata can be linked and the original data disclosed. MRPs must ensure strict licensing rules and guidelines to ensure that this does not occur. In the future, it is likely that microdata will be distributed via remote access and Statistical Agencies will have more control of who receives the microdata.

References

Anwar, N. (1993). Micro-Aggregation – The Small Aggregates Method. *Informe Intern*, Luxembourg, Eurostat.

Benedetti, R., Capobianchi, A., and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design. *Contributi Istat.*

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure limitation of Microdata. *Journal of the American Statistical Association 85*, 38-45.

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

Dalenius, T. and Reiss, S.P. (1982) Data Swapping: A Technique for Disclosure limitation. *Journal of Statistical Planning and Inference, 7*, 73-85.

Defays, D. and Nanopoulos, P. (1992) Panels of Enterprises and Confidentiality: The Small Aggregates Method. *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys*, 195–204.

Domingo-Ferrer, J., Mateo-Sanz, J. and Torra, V. (2001) Comparing SDC Methods for Micro-Data on the Basis of Information Loss and Disclosure Risk. *ETK-NTTS Pre-Proceedings of the Conference*, Crete, June 2001.

Domingo-Ferrer, J. and Mateo-Sanz, J. (2002) Practical Data-Oriented Micro-aggregation for Statistical Disclosure limitation. *IEEE Transactions on Knowledge and Data Engineering, Vol. 14, Issue 1*, 189-201.

Domingo-Ferrer, J. and Torra, V.(2003) Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage, *Statistics and Computing*, *Vol. 13, No. 4,* 343-354.

Elamir, E. and Skinner, C.J. (2006) Record-Level Measures of Disclosure Risk for Survey Micro-data. *Journal of Official Statistics, 22,* 525-539.

Elliot, M., Manning, A., Mayes, K., Gurd J. and Bane, M. (2005) SUDA: A Program for Detecting Special Uniques, In: *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Geneva, 353-362.

Fienberg, S.E. and McIntyre, J. (2005) Data Swapping: Variations on a Theme by Dalenius and Reiss. *Journal of Official Statistics, 9*, 383-406.

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

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

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

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

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

Oganian, A. and Karr, A. (2006) Combinations of SDC Methods for Micro-data Protection. Privacy. In: *Statistical Databases-PSD2006* (eds. J. Domingo-Ferrer and L. Franconi), Springer LNCS 4302, 102-113.

Rao, J.N.K. and Thomas, D.R. (2003) Analysis of Categorical Response Data from Complex Surveys: an Appraisal and Update. In: *Analysis of Survey Data* (eds. R.L. Chambers and C.J. Skinner), Chichester: Wiley, 85-108.

Rinott, Y. and Shlomo, N (2006) A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation. In *PSD'2006 Privacy in Statistical Databases*, (Eds. J. Domingo-Ferrer and L. Franconi), Springer LNCS 4302, 82-93.

Rinott, Y. and Shlomo, N. (2007a) A Smoothing Model for Sample Disclosure Risk Estimation. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond, IMS Lecture Notes Monograph Series*, Vol. 54, 161-171.

Rinott, Y. and Shlomo, N. (2007b) Variances and Confidence Intervals for Sample Disclosure Risk Measures. *56th Session of the International Statistical Institute Invited Paper, Lisbon 2007 (to appear).*

Shlomo, N. (2007) Statistical Disclosure Limitation Methods for Census Frequency Tables. *International Statistical Review, Vol. 75, Number 2, pp. 199-217.*

Shlomo, N. and De Waal T. (2008) Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. *Journal of Official Statistics, 24, No. 2*, 1-26.

Shlomo, N. and Young, C. (2006) Statistical Disclosure Limitation Methods Through a Risk-Utility Framework. In *PSD'2006 Privacy in Statistical Databases*, (Eds. J. Domingo-Ferrer and L. Franconi), Springer LNCS 4302, pp. 68-81.

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

Skinner, C.J. and Holmes, D. (1998) Estimating the Re-identification Risk Per Record in Microdata. *Journal of Official Statistics 14*, 361-372.

Skinner, C.J. and Shlomo, N. (2008) Assessing Identification Risk in Survey Micro-data Using Log-linear Models. *JASA Applications and Case Studies* (forthcoming) See: http://eprints.soton.ac.uk/41842/01/s3ri-workingpaper-m06-14.pdf.

Skinner, C.J. and Shlomo, N. (2007) Assessing the Disclosure Protection Provided by Misclassification and Record Swapping. *56th Session of the International Statistical Institute Invited Paper*, Lisbon 2007 (to appear).

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

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

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