Journal of Privacy and Confidentiality (2010) **2**, Number 1, pp. 73–91

Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility

Natalie Shlomo*

1 Introduction

Statistical agencies release sample microdata from social surveys under different modes of access ranging 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. In addition, statistical agencies often have on-site datalabs where registered researchers can access unperturbed statistical data. Statistical agencies will generally set up a panel of experts to form a Microdata Review Panel (MRP) who will then have the authority to release microdata. To make informed decisions about the release of microdata, the MRP needs objective disclosure risk measures to determine tolerable risk thresholds according to the access mode. They also need to monitor the application of data masking techniques and to ensure the quality and utility of the released microdata.

This paper provides a review of some recent developments in disclosure risk assessment and discusses how these may be integrated with established methods of data masking and some recent methods of utility assessment. It is only through a holistic approach of a disclosure risk-data utility assessment that microdata can safely be released while ensuring high quality and utility in the data.

In any released microdata set, direct identifying key variables, such as name, address or identification numbers, are removed. Disclosure risk typically arises from attribute disclosure where small counts on cross-classified indirectly identifying key variables (such as: age, sex, place of residence, marital status, occupation, etc.) can be used to identify an individual and confidential information may be learnt. Generally, identifying key variables are categorical. Sensitive variables are often continuous, but can also be categorical. In order to protect a data set, one can either apply a Statistical Disclosure Limitation (SDL) method on the identifying key variables or the sensitive variables. In the first case, identification of a unit is rendered more difficult, and the probability that a unit is identified is hence reduced. In the second case, even if an 'intruder' succeeds in identifying a unit by using the values of the identifying key variables, the sensitive variables would hardly disclose any useful information on the particular record. One can also apply SDL techniques on both the identifying and sensitive variables simultaneously. This offers more protection, but also leads to more information loss.

Based on the literature, methods for assessing disclosure risk for sample microdata arising from social surveys can be classified into three types:

*Southampton Statistical Sciences Research Institute, University of Southampton, SO17 1BJ, UK, mailto:N.Shlomo@soton.ac.uk

- Heuristics that identify special uniques on a set of cross-classified key variables, i.e., sample uniques that are likely to be population uniques (see, Skinner and Elliot, 2002; Elliot et al., 2005; and references therein),
- Probabilistic record linkage on a set of key (matching) variables that can be used to link the microdata to an external population file (see Yancey, Winkler, and Creecy, 2002; Domingo-Ferrer and Torra, 2003; and references therein),
- Probabilistic modeling of disclosure risk which was developed under two approaches: a full model-based framework taking into account all of the information available to 'intruders' and modeling their behavior (see Duncan and Lambert, 1989; Lambert, 1993; and later Reiter, 2005; and references therein), and a more simplified approach that restricts the information that would be known to 'intruders' (see Bethlehem, Keller, and Pannekoek, 1990; Benedetti, Capobianchi, and Franconi, 1998; Fienberg and Makov, 1998; Skinner and Holmes, 1998; Elamir and Skinner, 2006; and references therein).

Heuristics and record linkage suffer from the drawback that there is no framework for obtaining consistent record-level and global-level disclosure risk measures. Record-level disclosure risk measures can be used to target high-risk records in the microdata for SDL methods. Global disclosure risk measures are aggregated from record-level risk measures and are essential for MRPs to inform decisions when releasing microdata. In addition, these types, as well as the full model-based probabilistic approach, do not take into account the protection afforded by the sampling. In Section 2, we review the simplified probabilistic modeling approach to disclosure risk assessment as the optimal approach. It provides consistent global- and record-level disclosure risk measures, takes into account the sampling mechanism, and is simple to implement. Skinner and Shlomo (2007) have further developed this approach to take into account the realistic case where key variables may be misclassified or purposely perturbed as an SDL method.

Based on the disclosure risk assessment, statistical agencies must choose appropriate SDL methods either by perturbing, modifying, or summarizing the data. The choice 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. Examples of non-perturbative SDL methods that are often applied at statistical agencies are global recoding and suppression of values or whole key variables. Sub-sampling records is also a non-perturbative method and is often used for producing Census microdata. Perturbative methods for masking continuous sensitive variables include: adding random noise (see Kim, 1986; Fuller, 1993; Brand, 2002; Yancey, Winkler, and Creecy, 2002); micro-aggregation where records are grouped and their values replaced by their average (Defays and Nanopoulos, 1992; Anwar, 1993; Domingo-Ferrer and Mateo-Sanz, 2002); rounding to a pre-selected rounding base; and rank swapping where values between pairs of record

within a small group are swapped (Dalenius and Reiss, 1982; Fienberg and McIntyre, 2005). Perturbative methods for categorical key 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 perturbative and non-perturbative methods see also: Willenborg and De Waal, 2001; Domingo-Ferrer, Mateo-Sanz, and Torra, 2001; and references therein.

Each SDL method impacts differently on the level of protection obtained in the microdata and information loss. Oganian and Karr (2006) discuss combining SDL methods in order to obtain more effective protection in the microdata. Shlomo and De Waal (2008) discuss optimizing SDL methods to preserve sufficient statistics as well as the logical consistencies in the microdata. In Section 3, we provide a review of some standard SDL methods for microdata which can be adapted to increase the utility in the data under the same levels of protection.

Information loss measures have been developed in Domingo-Ferrer, Mateo-Sanz, and Torra, 2001; Gomatam and Karr, 2003; Karr et al., 2006; Shlomo and Young, 2006; and Shlomo, 2007. In Section 4, we review some useful information loss measures that quantify the effects of SDL methods on statistical analysis.

In Section 5, we illustrate the Disclosure Risk-Data Utility assessment on samples drawn from a Census where the population is known and can be used to investigate sample-based inference and validate results. The aim is to provide an example of how a statistical agency might carry out a holistic assessment of microdata with respect to managing disclosure risk while ensuring high quality data to users. 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 would enable matching the released microdata to a publicly available file containing names and addresses. Under a probabilistic approach, 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. The assumption is 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.

Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global

risk measures for the entire file. Denoting F_k the population size in cell k of a table spanned by key variables having K cells and f_k the sample size and $\sum_{k=1}^{K} F_k = N$ and $\sum_{k=1}^{K} f_k = n$. The set of sample uniques is defined: $SU = \{k : f_k = 1\}$ 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. The population frequencies F_k are unknown and need to be estimated from the probabilistic model the risk measures by:

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(F_k = 1|f_k = 1)$$

and

$$\hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(1/F_k|f_k = 1) \tag{1}$$

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: $f_k|F_k \sim Bin(F_k, \pi_k)$. It follows that:

$$f_k \sim Pois(\pi_k\lambda_k)$$

and

$$F_k|f_k \sim Poisson(\lambda_k(1 - \pi_k)) \tag{2}$$

where $F_k|f_k$ 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:

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

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:

$$\begin{aligned} &P(F_k = 1|f_k = 1) = \exp(\lambda_k(1-\pi_k)), \\ &E(1/F_k|f_k = 1) = [1 - \exp(\lambda_k(1-\pi_k))]/[\lambda_k(1-\pi_k)] \end{aligned} \tag{4}$$

Plugging $\hat{\lambda}_k$for λ_k in (**??**) leads to the estimates $\hat{P}(F_k = 1|f_k = 1)$ and $\hat{E}[1/F_k|f_k = 1]$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (**??**). 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 $h(\lambda_k) = P(F_k = 1|f_k = 1)$ for τ_1 and $h(\lambda_k) = E(1/F_k|f_k = 1)$ 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

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

For example, for τ_1 they obtain:

$$\hat{B}_1 \approx \sum_k \hat{\lambda}_k \exp(-\hat{\lambda}_k)(1-\pi_k)\{(f_k - \pi_k\hat{\lambda}_k) + (1-\pi_k)[(f_k - \pi_k\hat{\lambda}_k)^2 - f_k]/(2\pi_k)]\} \tag{6}$$

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

Skinner and Shlomo (2008) address the estimation of disclosure risk measures under complex survey designs with stratification, clustering, and survey weights. While the

method described assumes that all individuals within cell k are selected independently using Bernoulli sampling, i.e., $P(f_k = 1|F_k) = F_k\pi_k(1-\pi_k)^{F_k-1}$, this may not be the case when sampling clusters (households). In practice, key variables typically include variables such as age, sex, and occupation, that 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 geography. Strata indicators should always be included in the key variables to take into account differential inclusion probabilities in the model. Under complex sampling, the $\{\lambda_k\}$ can be estimated consistently using pseudo-maximum likelihood estimation (Rao and Thomas, 2003), where the estimating equation in (**??**) is modified as:

$$\sum_k [\hat{F}_k - \exp(x_k'\beta)]x_k = 0 \tag{7}$$

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 (**??**) and π_k is replaced by the estimate $\hat{\pi}_k = f_k/\hat{F}_k$. Note that the risk measures in (**??**) 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. Denoting the cross-classified key variables in the population and the microdata as X and assuming that X in the microdata have undergone some misclassification or perturbation error denoted by the value $\tilde{X}$ and determined independently by a misclassification matrix M,

$$M_{kj} = P(\tilde{X} = k|X = j), \tag{8}$$

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

$$\frac{M_{kk}(1-\pi M_{kk})}{\sum_j F_j M_{kj}/(1-\pi M_{kj})} \leq \frac{1}{F_k} \tag{9}$$

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:

$$\tau_2 = \sum_k I(f_k = 1) M_{kk} / \tilde{F}_k \tag{10}$$

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 (**??**) can be obtained by probabilistic modeling on the misclassified sample:

$$\hat{\tau}_2 = \sum_k I(\tilde{f}_k = 1) M_{kk} \hat{E}\left(1/\tilde{F}_k | \tilde{f}_k\right) \tag{11}$$

3 Statistical Disclosure Limitation Methods for Sample Microdata

Depending on the outcome of the disclosure risk measures, the risk thresholds set by MRPs and the mode of access, SDL methods may need to be applied. The standard procedures when releasing microdata are to recode and collapse the identifying categorical key variables in order to reduce the risk of identification. This is a non-perturbative method which limits the amount of information released. Categorical key variables can also be protected using a perturbative method, such as record swapping or more generally the post-randomization method (PRAM) (see Gouweleeuw et al., 1998). As a perturbative method, PRAM alters the data, and therefore we can expect consistent records to start failing edit rules. Edit rules 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".

Willenborg and De Waal (2001) describe the process of applying PRAM as follows: Let $\mathbf{P}$ be a $L \times L$ transition matrix containing conditional probabilities $p_{ij} = p$ perturbed category is j|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 $\mathbf{u}$ 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 $\mathrm{E}(\mathbf{t}^*|\mathbf{t}) = \mathbf{tP}$. 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.

The condition of invariance can be placed 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, a matrix $\mathbf{Q}$ is calculated by transposing matrix $\mathbf{P}$, multiplying each column j by v_j and then normalizing its rows so that the sum of each row equals one. The invariant matrix is obtained by $\mathbf{R} = \mathbf{PQ}$. The invariant matrix $\mathbf{R}$ may distort the desired probabilities on the diagonal, so Shlomo and De Waal (2008) 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, a "without" replacement selection strategy for choosing values to perturb can be implemented based on the expectations calculated from the transition probabilities. 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. The perturbation can be controlled 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.

In addition to the categorical key variables, sensitive continuous variables can also be perturbed so that even if an identification is made based on the key variables, the information is protected from the 'intruder'. The following are some common perturbative methods for masking sensitive continuous variables that have been adapted to preserve sufficient statistics and logical consistencies in the microdata:

3.1 Additive noise

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 (see Brand, 2002 and references therein for a summary and discussion of additive random noise). 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. Following Kim (1986) and Fuller (1993), correlated random noise can be added 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 as described in Shlomo and De Waal (2008) 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. 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 by aggregating perturbed variables since this causes unnecessary information loss.

Shlomo and De Waal (2008) propose implementing Procedure 1 in a multivariate setting where correlated Gaussian noise is added 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: $(\varepsilon_x, \varepsilon_y, \varepsilon_z)^T \sim N(\mu', \mathbf{\Sigma})$, 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 μ 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 $\mathbf{\Sigma}$ 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 (see Defays and Nanopoulos, 1992; Anwar, 1993; Domingo-Ferrer and Mateo-Sanz, 2002; and references therein). 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, random noise can be generated according to the magnitude of the difference between the total variance and the "between" variance, and added 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 simultaneously that are connected through an additivity constraint, a linear programming technique can be applied 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.

Fellegi (1975) proposed a technique for implementing unbiased random rounding on a one-dimensional table that preserves the overall total (and hence the mean) of the variable being rounded. The technique can be carried out 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, a 'without replacement' strategy can be used for selecting values to round up: for 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 $res(m)$. 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 overall totals.

4 Information Loss Measures

The utility of microdata that has undergone SDL methods 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, proxy measures have been developed that assess the utility based on measuring distortions to distributions and the impact on bias, variance and other statistical analysis tools (Chi-squared statistic, R^2 goodness of fit, rankings, etc.).

Domingo-Ferrer, Mateo-Sanz, and Torra, 2001; Gomatam and Karr, 2003; Shlomo and Young, 2006; and Shlomo, 2007 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 in the microdata as a result of applying SDL methods. Some useful metrics for aggregated data are presented in Gomatam and Karr, 2003. The AAD is a distance metric based on the average absolute difference per cell in the distribution. Let D represent a frequency distribution produced from the microdata and let $D(c)$ be the frequency in cell c. The Average Absolute Distance per Cell is defined as:

$$AAD(D_{orig}, D_{pert}) = \sum_{c} |D_{pert}(c) - D_{orig}(c)|/n_c \tag{12}$$

where n_c is the number of cells in the distribution.

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. Typically, the Cramer's V is used, which is a measure of association between two categorical variables: $CV = \sqrt{\frac{\chi^2/n}{\min(R-1),(C-1)}}$. The information loss measure is the percent relative difference between the original and perturbed table:

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

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

4.3 Impact on a Regression Analysis

For continuous variables, it is useful to assess the impact on the correlation and in particular the R^2 of a regression (or ANOVA) analysis. For example, in an ANOVA,

the test involves whether a continuous dependent variable has the same means across groups defined by a categorical explanatory variable. 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 distribution.

The information loss is based on assessing differences in the means across categories of an explanatory variable. Let $\bar{x}_k$ be the mean in category k and define the 'between' variance of this mean by: $BV(\bar{x}_{orig}) = \frac{1}{|k|-1}\sum_k(\bar{x}_k - \bar{x})^2$ where $\bar{x}$ is the overall mean in the sample and $|k|$ is the number of categories of the explanatory variable. Information loss is measured by:

$$BVR(\bar{x}_{pert}, \bar{x}_{orig}) = 100 \times \frac{BV(\bar{x}_{pert}) - BV(\bar{x}_{orig})}{BV(\bar{x}_{orig})} \tag{14}$$

In addition, another analysis of information loss involves comparing coefficient estimates based on applying a regression model on both the original and perturbed microdata.

5 Example

We present an example of how a statistical agency might assess disclosure limitation strategies through a disclosure risk-data utility analysis. We use the 1995 Israel 20% Census sample composed of N=753,711 individuals aged 15 and over living in households in Israel at the time of the Census. This large sample serves as a 'population' from which we draw samples. Since the population is known, we can investigate the properties of sample-based methods and verify results. We draw simple random samples of individuals with a sampling fraction of $\pi = 1/100$ (n=7,537). The key variables in the microdata 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, we focus on one sensitive variable in the microdata: income from earnings.

The statistical agency needs to assess disclosure risk of the sample microdata and considers SDL methods. Since disclosure risk is defined as the risk of identification based on the categorical key variables, we consider two SDL methods which reduce the risk by masking the Locality Code:

- Recoding and collapsing categories of the Locality Code (from 85 to 30 categories),
- PRAM on the large Locality Codes with 0.70 on the diagonal of the misclassifi-

cation matrix. We implement an invariant PRAM to preserve expected marginal frequencies of the Locality Codes.

After applying SDL methods, the disclosure risk needs to be reassessed and compared to tolerable risk thresholds set by the MRP at the statistical agency. In addition, information loss measures need to be calculated in order to compare and understand the impact of the methods on statistical inference. In this example we use:

- AAD calculated by differencing the marginal frequencies of the original Locality Codes to the perturbed Locality Codes. For the recoded collapsed Locality Codes, we imputed the average frequency, for example, if 10 localities were recoded into a single code, each locality would receive 1/10 of the total,
- RCV on a table defined by original and collapsed or perturbed Locality Code and Occupation,
- BVR where the dependent variable is average income and the independent variable the original and collapsed or perturbed Locality Code.

Table 1 presents a comparison of these two SDL methods with respect to disclosure risk and data utility. The 'true' risk measure based on $\tau_2 = \sum_k I(f_k = 1)1/F_k$ is 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 (**??**) is given in parenthesis. Note that to estimate the disclosure risk for PRAM we used the formula in (**??**).

As can be seen in Table 1, recoding and collapsing Locality Code causes more information loss compared to PRAM, even with 30% of the Locality Codes perturbed. The AAD had an average difference of 7.2 per code for the recoded Locality Codes while PRAM had an average difference of 3.9 per code. This result is not surprising since we used the invariance property for PRAM which preserves expected marginal frequencies. The other information loss measures based on the original Locality Codes compared to the recoded or perturbed Locality Codes were significantly worse under the method of recoding. 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. 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 SDL methods, for example, by identifying those records that remain unique after the recoding and subsequently implementing PRAM on the high-risk records only.

After protecting key variables, statistical agencies might consider taking further action by perturbing sensitive variables, such as income. In our example, income was

Table 1: Comparison of SDL techniques: Recoding and PRAM

	Original Locality Codes ($\tau_2 = 1,025.7$)	Recorded locality Codes (30 categories) ($\tau_2 = 571.5$)	PRAM on localities (85 categories with 70% perturbation) ($\tau_2 = 714.7$)
Disclosure Risk			
$\hat{\tau}_2$ (test statistic)	1015.5 (194)	599.9 (1.32)	729.5 (1.42)
Sample uniques	4,005	3,376	3,479
$\hat{\tau}_2/SU$	25.3%	17.8%	20.9%
$\hat{\tau}/n$	13.5%	8.0%	9.7%
Utility			
AAD across 85 localities with mean imputation for recoded cells	0	7.22	3.88
RCV for lacalities $\times$ occumpation (11) (true=0.1370)	0	-32.7%	-7.5%
BVR for average income between localities (true=$3.082 * 10^9$)	0	-44.4%	-8.9%

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 basic techniques for perturbing income from earnings for those records with non-zero income (3,249 out of the 7,537 individuals in the sample): correlated and uncorrelated additive noise, controlled and uncontrolled random rounding to base 100 and micro-aggregation (size of groups=10) with and without additive noise. Results across 50 simulated samples are given in Table 2.

Table 2 shows that adding noise to the variables causes the greatest average absolute distance between original and perturbed cells of income groups which is also reflected in the high percentages of records that are switching income groups. There is not much difference between controlled and uncontrolled rounding to base 100 because of the large sample size (3,729 individuals with non-zero income) and hence carrying out a with or without replacement strategy for selecting values to round provides the same results. The percent difference in the variance as well as the BVR have negative numbers for the microaggregation showing a decrease in the overall and between variance of average income. Adding noise to the microaggregated variable should have corrected the variance but this seems to have an adverse effect on the BVR. Adding correlated noise also improved the variance and the BVR of average income although it introduced more association between the income groups and occupations resulting in a higher RCV. Rounding and microaggregation also increased the association. Overall, while the frequencies of the income groups may have changed significantly, the impact on statistical inference is minimal.

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

	Additive Noise		Rounding to Base 100		Microaggregation	
	Uncorrelated	Correlated	Uncontrolled	Controlled	Without noise	With noise
AAD across 16 income groups	27.8	24.8	5.9	5.9	4.8	20.8
Percent Difference in Variance	7.0%	0.0%	0.0%	0.0%	-1.5%	-0.7%
RCV for income groups (16) $\times$ occupation (11) (true=0.1736)	0.2%	1.9%	0.9%	0.7%	1.0%	1.5%
BVR average income between localities (85) (true=$3.082*10^9$)	1.0%	0.0%	0.0%	0.1%	-0.9%	-1.4%
Percentage of records switching income groups	26.6%	17.4%	5.0%	5.1%	1.8%	13.9%

6 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: (**??**) the need for a reliable method for objectively assessing disclosure risk; (**??**) SDL methods 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.

Duncan, G. and Lambert, D. (1989) The Risk of Disclosure for Microdata. *Journal of Business and Economic Statistics 7*, 207–217.

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.

Fellegi, I.P. (1975) Controlled random rounding. *Survey Methodology I*, 123–33.

Fienberg, S.E. and Makov, U.E. (1998) Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data. *Journal of Official Statistics 14*, 385–397.

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 Ran-

domisation 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`.

Karr F., Kohnen, A., Oganian, A., Reiter, J. And Sanil, A.(2006) A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *National Institute of Statistical Sciences. Technical Report Number 153.*

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.

Lambert, D. (1993) Measures of Disclosure Risk and Harm. *Journal of Official Statistics 9,* 313–331.

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.

Reiter, J.P. (2005) Estimating Risks of Identification Disclosure in Microdata. *Journal of the American Statistical Association 100*, 1103–1112.

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. (2007) Assessing the Disclosure Protection Provided by Misclassification and Record Swapping. *56th Session of the International Statistical Institute Invited Paper*, Lisbon 2007.

Skinner, C.J. and Shlomo, N. (2008) Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of American Statistical Association, Vol. 103, Number 483*, 989–1001.

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.