Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata

Fang Liu and Roderick J.A. Little

Biostatistics Department, School Of Public Health, University of Michigan, Ann Arbor, Michigan, 48109-2029

Key Words: Statistical Disclosure Control, General Location Model, Information Loss, Protection, Disclosure Risk, Selection

1. INTRODUCTION

Statistical disclosure control requires techniques that balance the requirements of protection of respondents' privacy and dissemination of information. In microdata set containing information on individual respondents, we assume that variables can be divided into two groups: key variables that are assumed to be known to data intruders from publicly available sources, and nonkey variables that are not publically available and include potentially sensitive information about respondents. Key variables $\mathbf{X}$ are treated as categorical and form a multi-way contingency table with K cells. A particular set of values of key variables defines a *key cell.* Cells containing $\leq s$ cases, where s is a pre-specified sensitivity threshold (e.g. three or five), are considered *sensitive cells*, Cases belonging to sensitive cells are called *sensitive cases*, and are considered to have a significant risk of disclosure.

Our focus here is on *statistical disclosure control (SDC)* techniques that provide protection to sensitive cases. Existing SDC methodologies include *global recoding, local suppression, data swapping, micro-aggregation*, and *post randomization (PRAM).* These techniques are model-free and somewhat adhoc. Model-based SDC techniques replace observed values of the data by predictions based on a statistical model. The added statistical uncertainty from these modifications can be reflected by releasing multiple independently modified data sets and using multiple imputation methods of inference (Rubin, 1987). In particular, Rubin (1993) proposes to build an imputation model from the sample data, and predict nonsampled values of survey variables in the population; this imputation is repeated independently $D > 1$ times, and a sample from each of the D imputed populations is released to the public. This methodology has several strengths , such as valid inferences under well-specified imputation models; a convenient measure of information loss due to modification of the original data; and a high protection of protection of respondents. A drawback of the method is the need to build a statistical model for the survey variable for the whole sample – a formidable task for large survey data sets. Quality of inferences from the synthetic MI data sets depends on how well this large model is specified. Since data intruders use keys to identify their targets, to impede their identification process and fulfill our purpose of protection, it may be sufficient to restrict MI to a subset of values of the key variables (Little, 1993). This simplifies the imputation task, and reduces the sensitivity of inferences to model misspecification. This article develops this idea through a method we call *Selective Multiple Imputation of Keys (SMIKe).* In SMIKe, only the values of key variables in a subset of cases – namely, sensitive cases mixed with a subset of nonsensitive cases – are imputed. Thus, instead of releasing samples of the imputed population data set, we release the sample data with values of key variables for some cases replaced by multiply imputations. The selective aspect of SMIKe limits information loss and allows satisfactory inferences without the need for a large and highly accurate model.

2. AN OVERVIEW OF SMIKe

Suppose in a data set, x is the key with K categories/cells formed by key variables $\mathbf{X}$, $\mathbf{Y}$ is a vector containing q nonkey variables, s is a chosen sensitivity threshold and n_{sen} is the total number of sensitive cases SMIKe consists of the following steps:

1. *Selection of nonsensitive cases.* For each sensitive case $i = 1, \ldots, n_{sen}$ ($i \in$ cell $\mathcal{S}_i$), select a *mixing set* $\mathcal{M}_i$ (of pre-specified size n^i_{mix}) of cases from nonsensitive cell(s) $\mathcal{C}_i$ that are close to the sensitive case with respect to $\mathbf{y}$. The mixing sets for sensitive cases may overlap. The value of n^i_{mix} may vary from case to case according to case sensitivity, but for simplicity, we choose the same value n_{mix} for all sensitive cases. n_{mix} serves as a tuning parameter to balance gains in protection against information loss. We define the following pooled sets of cases:

$$\mathcal{M} = \bigcup_{i=1}^{n_{sen}} (\mathcal{M}_i \cup i) \text{ and } \mathcal{C} = \bigcup_{i=1}^{n_{sen}} (\mathcal{C}_i \cup \mathcal{S}_i),$$

This research was supported by grant DMS9803720 from the National Science Foundation and grant UR6/CCU517481 from the Centers for Disease Control. We also thank the ICPSR at University of Michigan for access to the Alameda County survey data.

where $\mathcal{C}$ consists of n cases in K^* cells, and $\mathcal{M}$ is a subset of $n^*(< n)$ sensitive and mixing cases from $\mathcal{C}$ that are subject to imputation of keys.

2. *Construction of an imputation model for keys.* A necessary condition for valid inferences from the multiply imputed data sets is that the masking mechanism in SMIKe is "masked at random (MAR)" in the sense defined by Little (1993). Suppose $\mathbf{y}$ is a subset of $\mathbf{Y}$ with dimension $p \leq q$ and $\tilde{x}$ is the predictive value of x. If the imputation model $p(\tilde{x}|\mathbf{y}, \mathcal{M})$ is built on $\mathcal{M}$, the requirement of MAR is fulfilled. However, since $\mathcal{M}$ is usually a small set, it is more efficient to build the imputation model using a larger data set $\mathcal{C}$ (the imputation model could be based on the full data set, but it seems more direct to restrict the model to key cells involved in the imputation). By Bayes' rule, we have

$$\begin{aligned}\frac{p(\tilde{x}=k_1|\mathbf{y},\mathcal{M})}{p(\tilde{x}=k_2|\mathbf{y},\mathcal{M})} &\propto \frac{p(\tilde{x}=k_1,\mathbf{y}|\mathcal{M})}{p(\tilde{x}=k_2,\mathbf{y}|\mathcal{M})}\\ &\propto \frac{p(\mathbf{y}|\tilde{x}=k_1,\mathcal{M})}{p(\mathbf{y}|\tilde{x}=k_2,\mathcal{M})}\frac{p(\tilde{x}=k_1|\mathcal{M})}{p(\tilde{x}=k_2|\mathcal{M})}\end{aligned}$$

where $\frac{p(\mathbf{y}|\tilde{x}=k_1,\mathcal{M})}{p(\mathbf{y}|\tilde{x}=k_2,\mathcal{M})}$ is the likelihood and $\frac{p(\tilde{x}=k_1|\mathcal{M})}{p(\tilde{x}=k_2|\mathcal{M})}$ is the prior odds. If the likelihood is based on $\mathcal{C}$, that is, $\frac{p(y|\tilde{x}=k_1,\mathcal{M})}{p(y|\tilde{x}=k_2,\mathcal{M})}$ in above equation is replaced by $\frac{p(\mathbf{y}|\tilde{x}=k_1,\mathcal{C})}{p(\mathbf{y}|\tilde{x}=k_2,\mathcal{C})}$, then efficiency of inferences from the imputed data is increased without destroying the MAR mechanism. $p(\tilde{x}|\mathbf{y})$ can be re-written as $\int p(\tilde{x}|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$. Thus to draw $\tilde{x}$ from its posterior predictive distribution, we first draw $\boldsymbol{\theta}$ from its posterior distribution given the data in $\mathcal{C}$ or $\mathcal{M}$ (components of $\boldsymbol{\theta}$ appearing in the likelihood part of are drawn from their posterior distribution given data $\mathcal{C}$, and that in the prior odds part are drawn from their posterior distribution given data $\mathcal{M}$). A noninformative prior distribution for $\boldsymbol{\theta}$ such as the Jeffeys' prior is suggested in the absence of strong prior information. Since x is treated as categorical, a natural imputation model for $p(\tilde{x}|\mathbf{y})$ (ignoring refinements to reflect clustering in the sample design) is the multinomial logit model. A computationally less onerous alternative when $\mathbf{y}$ is continuous is to fit the general location (GL) model (Olkin and Tate, 1961), as discussed in the next section.

3. *Multiple Imputation of Keys.* Independently draw D sets of $\tilde{x}$'s for each of the n^* cases in $\mathcal{M}$ from the model.

4. *Assessment of information loss and protection and release of MI data.* Combine each imputed set of $\tilde{x}$ with other nonimputed variables, and assess protection and information loss in these D SMIKe data sets. If the requirements of information loss and protection are satisfied, release these D data sets; Otherwise, go back to Step 1 with a modified selection plan and an adjustment of n_{mix} if necessary.

3. SMIKe FOR CONTINUOUS y

We now discuss the steps of SMIKe in more detail. We focus on the special case where the nonkey variables $\mathbf{y}$ are continuous, while outlining extensions to other situations.

3.1 Selection of nonsensitive cases

We propose to choose mixing cases for a sensitive case i that are as similar as possible to i in terms of the nonkey variables $\mathbf{y}$. Intuitively, imputing keys within relatively homogeneous sets of cases has the virtue of tending to distribute the multiply imputed cases over the set of sensitive and nonsensitive key cells in the mixing set, thus promoting the mixing of sensitive and nonsensitive cases, and increasing protection.

If the nonkey variables are continuous and approximately normal, a natural measure of closeness between cases i and j is the Mahalanobis distance (MD) $(\mathbf{y}_i - \mathbf{y}_j)^T S^{-1}(\mathbf{y}_i - \mathbf{y}_j)$, where S^{-1} is the pooled sample covariance matrix. There is considerable flexibility in how mixing sets might be chosen; we consider two variants of selection based the closeness measure, global selection (GS) and local selection (LS).

GS places no restriction on the set of nonsensitive key cells that contribute to the mixing set. It computes the distance between sensitive case i and each nonsensitive case j and then chooses the n_{mix} closest nonsensitive cases in terms of the closeness measure. LS restricts the set of key cells that contribute to the mixing set. It first picks $Q(\geq 1)$ nonsensitive cell(s) that are closest to a sensitive case i as measured by the distance between $\mathbf{y}_i$ and the cell means. The cells are chosen so that they contain at least n_{mix} cases. LS then selects n_{mix} closest cases within these Q cells to be the mixing set for case i. The mixing sets can be further constrained to avoid information loss for particular analyses, for example, by preserving the margins of tables formed by a subset of key variables. LS may involve less computation than GS and also involves less information loss for some analyses by restricting the mixing set to a smaller set of cells. On the other hand, GS may provide better protection, since sensitive cases are mixed with cases from a wider range of key cells.

3.2 Construction of an imputation model for keys

With continuous $\mathbf{y}$, the GL model is defined in terms of the marginal distribution of x and conditional distribution of $\mathbf{y}$ given x:

$$p(x_i = k|\mathcal{M}) = \pi_k, \text{ where } k = 1, \ldots, K^*, \sum_k \pi_k = 1$$

$$p(\mathbf{y}_i|x_i, \mathcal{C}) \overset{indep}{\sim} N_p(\boldsymbol{\mu}_{x_i}, \Sigma) \text{ for } i = 1, \ldots, n.$$

A transformation of $\mathbf{y}$ might improve the fit of the model, which assumes $\mathbf{y}$ is normal with a constant within-cell covariance matrix. Another possible model is the extended general location (EGL) model (Liu and Rubin, 1998), where covariance matrix does not have to be constant and normal distribution may be replaced by other distributions. If the cases $i = 1, \ldots, n$ are not independent, as in multistage samples, we may need to modify the GL model to incorporate correlation among cases. Denote the parameters in the model by $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_{K^*-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K^*}, \Sigma\}$. The log-likelihood for the GL model is $L(\boldsymbol{\theta}) =$

$$-\frac{1}{2}|\Sigma|^n + \sum_{k=1}^{K^*} n_k^* log(\pi_k) - \frac{1}{2}\sum_{k=1}^{K^*}\sum_{i=1}^{n_k}(\mathbf{y}_i - \boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k),$$

where n_k is the number of observations in cell k and n_k^* is the number of selected observations in cell k. If Jeffreys' prior is used, $p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K^*} \pi_k^{-\frac{1}{2}} |\Sigma|^{-\frac{p+1}{2}}$,
then the posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{array}{rcl} [\boldsymbol{\pi}|Data] & \sim & Dirichlet(n_1^* + \frac{1}{2}, \ldots, n_{K^*}^* + \frac{1}{2}) \\ [\Sigma|\boldsymbol{\pi}, Data] & \sim & Inv - Wishart(S, n - K^*) \qquad (1) \\ [\boldsymbol{\mu}_k|\boldsymbol{\pi}, \Sigma, Data] & \sim & N_p(\bar{\mathbf{y}}_k, \Sigma/n_k) \text{ for } k = 1, \ldots, K^*, \end{array}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K^*})^T, \boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{pk})^T$, S is the pooled sample covariance matrix of n cases, and $\bar{\mathbf{y}}_k$ is the sample mean of $\mathbf{y}$ in cell k. The full conditional posterior predictive distribution of $\tilde{x}_i$ for case $i = 1, \ldots, n^*$ is given by

$$p(\tilde{x}_i = k|\boldsymbol{\theta}, Data) = \frac{\pi_k exp(\omega_{ik})}{\sum_{k'=1}^{K^*} \pi_{k'} exp(\omega'_{ik'})} \text{ for } k = 1, \ldots, K^*, \tag{2}$$

where

$$\omega_{ik} = \mathbf{y}_i^T\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k \text{ (similarly for } \omega'_{ik'}). \tag{3}$$

3.3 Multiple imputation of keys

Given $\mathbf{y}$ and drawn parameters $\boldsymbol{\theta}$, we can calculate ω_{ik} by Eqn. [3] and independently draw $D > 1$ time of $\tilde{x}_i$ from the set $(1, \ldots, K^*)$ with the probabilities given in Eqn. [2] for $i = 1, \ldots, n^*$. Inferences, procedures for hypothesis testing and other problems from SMIKe data are based on standard MI methods, as discussed in Rubin (1987), Little and Rubin (2002) or Schafer (1997). In particular, suppose θ is a scalar parameter of interest. In each data set d ($d = 1, \ldots, D$), we can calculate the estimate $\hat{\theta}_d$ of θ and the estimated variance V_d of $\hat{\theta}_d$. Based on these quantities, $W = \sum_{d=1}^{D} V_d/D$ (within variance), $B = \sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta})^2/(D-1)$ (between variance) and $T = W + (1+\frac{1}{D})B$ (total variance) can can calculated, where $\bar{\theta} = \sum_{d=1}^{D}\theta_d$ and $(1 + \frac{1}{D})$ is a correction factor for small D.

Strictly speaking, standard MI theory does not apply to these inferences, since the parameters are drawn from the complete data prior to masking, rather than the incomplete data with values of $\mathbf{x}$ masked. Since the posterior distribution of the parameters is based on more information than in standard missing-data application of MI, and the amount of imputation is modest, we expect the MI inferences to be slightly conservative. This conjecture is consistent with results from simulation studies, to be reported elsewhere in the interests of space.

3.4 Assessment of information loss and protection

In order to assess the performance of SMIKe, we need measures of information loss and of reduction of disclosure risk from the masking procedure. Information loss varies for different analyses, and hence might be computed for a range of analyses of interest. Our measure of information loss for inference about a particular parameter θ is the fraction of missing information from MI theory (Rubin 1987), which is given by the expression

$$\gamma = (1 + \frac{1}{D})\frac{B}{T}, \; (\gamma \in [0, 1]) \tag{4}$$

Measurement of disclosure risk is difficult, since it requires conjectures about the behavior of the intruder. In SMIKe the difficulties are compounded by the release of MI data sets. We first discuss the measure of disclosure risk in the original data set (R(orig)), then present two empirical approaches to measure disclosure risk in SMIKe data (R(smike)). In the original data, disclosure risk associated with a case i in a cell k, R_{ki}, is defined as

$$\begin{cases} 1/n_k & \text{if } n_k \leq s; \\ 0 & \text{if } n_k > s, \end{cases} \tag{5}$$

where n_k is the cell size of k. R_{ki} can be interpreted as the probability of case i being correctly identified. Disclosure risk of cell k is defined as

$$R_k = \sum_{i=1}^{n_k} R_{ki}, \tag{6}$$

the sum of the disclosure risk of cases in that cell. (Another possibility is $R_k = 1/n_k$ and $R_{ki} = 1/n_k^2$, which assigns more risk to unique cases, but has a less direct interpretation). If $C_1, C_2, \ldots, C_s$ are respectively the numbers of unique, 2-case, ..., s-case cells, then

$$\begin{array}{rcl} R(\text{orig}) & = & \sum_{k=1}^{K^*}\sum_{i=1}^{n_k} R_{ki} = \sum_{k=1}^{K^*} R_k \\ & = & C_1 + C_2 + \ldots + C_s, \qquad (7) \end{array}$$

which is the number of sensitive cells.

We consider two measures of R(smike). For both measures, we assume that (a) the data intruder's target is in the released data set; (b) the data intruder holds the correct key for his target; and (c) the data intruder identifies his target by matching the target's key with those of the cases in the data set. Our simpler approach is to first calculate disclosure risk $R_1^d(d = 1, \ldots, D)$ in each of the D data sets, then average R_1^d over the data sets. That is,

$$R_1(\text{smike}) = \sum_d R_1^d / D \tag{8}$$

To obtain a measure of R_1^d, suppose the intruder knows the value k of the key for a particular target case i. In imputed data set d, case i gets imputed into cell $\tilde{k}$ with $m_{\tilde{k}}$ cases, that is, $x_i = \tilde{k}$. Thus the disclosure risk on case i in imputed data set d is

$$R_{i\tilde{k}}^d = \begin{cases} 0 & \text{if } \tilde{k} \neq k; \\ 0 & \text{if } \tilde{k} = k \text{ and } m_{\tilde{k}} > s. \\ \frac{1}{m_{\tilde{k}}} & \text{if } \tilde{k} = k \text{ and } m_{\tilde{k}} \leq s; \end{cases}$$

That is, there is assumed to be no disclosure risk if case i is imputed out of its original cell k; if it remains in its original cell, then the disclosure risk is computed in the same way as for the original data set, based on the number of cases in cell k in the imputed data set. The resulting measures of risk for cell k is

$$R_{k1}^d = \begin{cases} 0 & \text{if } m_k = 0 \text{ or } m_k > s; \\ \frac{m_{nat,k}}{m_k} & \text{if } 0 < m_k \leq s, \end{cases} \tag{9}$$

where $m_{nat,k}$ is the number of cases in cell k whose original keys are k. Thus, $m_k - m_{nat,k}$ represents the inflow to cell k from other cells in the d^{th} imputed data sets. The summation of disclosure risk over all the key cells gives the measure R_1^d. Averaging over data sets as in Eqn. [8] gives the final measure $R_1(\text{smike})$.

The measure $R_1(\text{smike})$ is easily understood and calculated, but averaging the risk over the multiply-imputed data sets may not reflect how a sophisticated data intruder picks a case as the target from the D data sets, taken in aggregate. Our second measure of disclosure risk $R_2(\text{smike})$ is intended to model the selection of the possible target by an intruder with more sophisticated statistical understanding or tools. To define $R_2(\text{smike})$, consider the two-way cross-tabulation of the keys and cases in Table 1, where $e_{ik}(i = 1, \ldots, n^*, k = 1, \ldots, K^*)$ is the number of MI data sets ($\leq D$) with $x_i = k$. Thus, for each sensitive case, the collection of D MI data sets yields a sample from an independent multinomial distribution with row margins fixed at D. Hence $p_{i|k} = e_{ik}/e_{+k}$ estimates the conditional probability that case i is a unique sensitive case in target cell k, given that there is a single sensitive case in that cell. We assume that the data intruder picks as the target the case with the maximum $p_{i|k}$ among all cases. If there are u_k cases sharing the maximum, he randomly picks one from them with the probability $1/u_k$. This rule is optimal in that it minimizes the expected loss with the binary loss function (loss=0 if the decision is right, loss=1 if it is wrong). Based on this selection rule, we measure the disclosure risk associated with sensitive case i with true key k as follows:

Table 1: *Cross-tabulation of Keys and Cases in SMIKe data*

	key	1	2	...	k	...	K^*	Row total
case								
1		e_{11}	e_{12}	...	e_{1k}	...	e_{1K^*}	$e_{1+} = D$
2		e_{21}	e_{22}	...	e_{2k}	...	e_{2K^*}	$e_{2+} = D$
.		.	.	.	.	.	.	
.		.	.	.	.	.	.	
.		.	.	.	.	.	.	
i		e_{i1}	e_{i2}	...	e_{ik}	...	e_{iK^*}	$e_{i+} = D$
.		.	.	.	.	.	.	
.		.	.	.	.	.	.	
.		.	.	.	.	.	.	
n^*		e_{n^*1}	e_{n^*2}	...	e_{n^*k}	...	$e_{n^*K^*}$	$e_{n^*+} = D$
column	total	e_{+1}	e_{+2}	...	e_{+k}	...	e_{+K^*}	$D \cdot n^*$

$$R_{ik2} = \begin{cases} 0 & \text{if } p_{i|k} \neq max_j\{p_{j|k}\} \text{ or } u_k = 0; \\ 0 & \text{if } p_{i|k} = max_j\{p_{jk}\} \text{ and } u_k > s; \\ \frac{1}{u_k} & \text{if } p_{i|k} = max_j\{p_{jk}\} \text{ and } u_k \leq s. \end{cases} \tag{10}$$

This leads to the following measure of risk for sensitive cell k:

$$R_{k2} = \begin{cases} 0 & \text{if } u_k > s \text{ or } u_k = 0; \\ \frac{u_{nat,k}}{u_k} & \text{if } 0 < u_k \leq s, \end{cases} \tag{11}$$

where as above, $u_{nat,k}$ is the number of cases among u_k whose original keys are k. Summation of R_{k2} over all the sensitive cells gives our second measure of disclosure risk, $R_2(\text{smike})$.

For either measure of disclosure risk, $R_1(\text{smike})$ or $R_2(\text{smike})$, we measure the protection from SMIKe by the reduction of disclosure risk

$$\begin{aligned} P_t &= \frac{R(\text{orig}) - R_t(\text{smike})}{R(\text{orig})} \\ &= 1 - \frac{R_t(\text{smike})}{R(\text{orig})}, \; (P_t \in [0,1]) \text{ for } t = 1, 2, \end{aligned} \tag{12}$$

with the interpretation that disclosure risk is reduced by $(100 \cdot P_t)\%$ by SMIKe. The absolute disclosure risk is also of interest in applications, but the relative reduction is more useful for measuring the trade-off between information loss and protection.

The above measures of disclosure risk would not be applicable to a data intruder interested in aggregate information, such as the mean of an outcome in a cell. Other measures of disclosure risk might also be developed based on alternative assumptions about intruder behavior.

4. A SMALL APPLICATION

Simulation studies of the proposed method will be reported elsewhere, in the interests of space. We illustrate the method on a subset of data from the 1995 panel of the Alameda County Health and Ways of Living Survey (AC-hawls) (ref). We select a subset of

$n = 1349$ cases with positive responses to two variables, "volunteer?" and "currently employed at paid job?', We divide the cases in into 3 groups – group 1 with data on "volhrs" only ($n_1 = 361$), group 2 data on both 'volhrs" and "emphrs"($n_2 = 828$), and group three with data on "emphrs" only ($n_3 = 160$). Key cells and imputation models are constructed within each group. We designate five variables as keys: "age" (recoded into 6 categories), "sex"(2), "race"(9), "retired?"(2) and "student?"(2); and choose two continuous nonkey variables – y_1="hours working as volunteer per week (volhrs)" and y_2="hours working as employee per week (emphrs)". The sensitivity threshold s is set at three and LS is used as the selection plan with $n_{mix} = 5$ for all sensitive cases. The number of sensitive cells and their mixing cases and cells are given in Table 2. Note that even when the number of sensitive cases is large, the number of mixing cells in $\mathcal{C}$ can be small due overlap in the mixing sets of sensitive cases. Before constructing the GL models, a logarithm transformation is applied to y_1 to correct for right skewness. For simplicity, we assume independence of the respondents though they are stratified into three cities). Therefore, in group 1, $\mathbf{y} = log(y_1)$, $\mathbf{y} = (log(y_1), y_2)^T$ in group 2 and $\mathbf{y} = y_2$ in group 3. Thus, three GL models are fitted to cases in $\mathcal{C}$ respectively in three groups and keys of cases in $\mathcal{M}$ are imputed independently for $D = 10$ times.

Table 2: *Counts of Sensitive and Mixing Cases and Cells in Three Groups*

group	sensitive cell			mixing	# of cells	# of cases
	unique	2-case	3-case	cell	in $\mathcal{M}$	in $\mathcal{M}$
1	16	10	1	6	33	75
2	15	4	2	5	26	59
3	32	20	3	11	66	131

The parameters we choose for assessing information loss are from three sources: means of $\mathbf{y}$ in mixing cells (separately in each group) used in SMIKe, coefficients from a logistic regression (on data set $\mathcal{D}$) of Z_1 ="health in general" on 24 independent variables (19 variables selected by back-elimination procedure plus five key variables), and coefficients from a logistic regression of Z_2 ="mental health" on 21 variables (16 variables selected by backward elimination procedure plus five key variables). Z_1 and Z_2 are two health indexes of broad interest to analysts of AC-hawls data which are not used in the imputation model. Both Z_1 and Z_2 have 4 categories in the order of "1=excellent, 2=good, 3=fair, 4=poor", and proportional odds logit model is fitted to each of them. There are 57 parameters in the regression of Z_1, and 46 in the regression of Z_2 (we recoded "race" to five categories and "age" to 4 categories to solve the multicollinearity problem among some of independent variables). When measuring protection, $\mathcal{D}$ is treated as an entity and the overall protection given by SMIKe is assessed by both P_1 and P_2. The results are presented in Table 3 and Figure 1, where Table 3 shows the effects of SMIKe on $\mathcal{D}$ in terms of both information loss of the means of the mixed cells and protection and Figure 1 shows the information loss of the parameters from the two logistic regressions (all the measures are Monte-Carlo estimates based on 500 repetitions).

Table 3: *Information Loss vs. Protection in SMIKe-treated $\mathcal{D}$*

group	InfoLoss		Cell in $\mathcal{C}$	
	μ_1	μ_2	n(use)[a]	original size
1	0.203(0.085)	-	6	6
	0.084(0.033)	-	5	13
	0.044(0.022)	-	5	18
	0.048(0.021)	-	5	19
	0.010(0.005)	-	5	57
	0.066(0.028)	-	10	17
2	0.160(0.068)	0.162(0.062)	5	5
	0.183(0.090)	0.242(0.121)	5	6
	0.042(0.020)	0.037(0.016)	5	23
	0.030(0.015)	0.033(0.016)	7	50
	0.071(0.033)	0.085(0.030)	8	16
3	-	0.118(0.042)	5	6
	-	0.066(0.029)	5	8
	-	0.076(0.034)	5	8
	-	0.059(0.025)	5	9
	-	0.039(0.017)	5	10
	-	0.017(0.008)	5	33
	-	0.010(0.005)	5	42
	-	0.001(0.001)	5	228
	-	0.066(0.040)	7	8
	-	0.084(0.032)	8	11
	-	0.004(0.002)	15	211

[a] Number of mixing cases in $\mathcal{C}$. Some sensitive cases select the same mixing cell, but choose mixing cases from the cell with or without overlap. This is why n(use) in some cells are greater than the pre-specified "$n_{mix} = 5$" after the combination step.

The protection provided by SMIKe is 0.987(0.004) and 0.978(0.016) respectively in P_1 and P_2; that is disclosure risk is reduced by $\sim$ 98%. There are at least two reasons for this favorable results. One is the large number of categories in the keys, which results in dispersal of the sensitive cases over a large set of cells of x; The second reason lies in matching of each sensitive case to cases in its mixing set, which promotes mixing. Given the large reduction in disclosure risk, the performance of SMIKe in terms of information loss is impressive. In Table 3, when the selected nonsensitive cell is large in size, information loss of the cell mean is modest ($< 9\%$). In both proportional odds logistic regressions, there are three parameters ($1 \sim 3$) for intercepts, 10 parameter ($4 \sim 13$) associated with key variables; the others are coefficients for other variables. The figures shows that information loss for the parameters associated with key variables are high ($\sim$ 20% to $\sim$ 40%) and that for the intercepts and coefficients of the nonkey variables is negligible. One explanation of this is that 4 out of 5 key variables

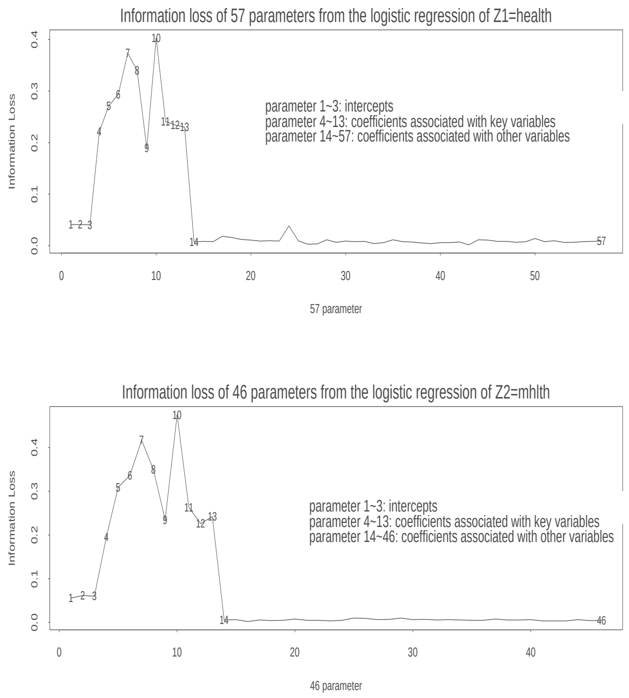

Figure 1: *Information Loss of the parameters from the logistic regression of Z_1 and Z_2 in SMIKe*

do not show statistically significant relationships with either Z_1 or Z_2 in the regressions, so adjustment for the keys does not have much impact on inferences for the regression coefficients of the nonkey variables.

This example, though illustrative, suggests that SMIKe can achieve major gains in disclosure protection and still preserve a large amount of information for statistical analysis.

5. DISCUSSION

SMIKe has the following attractive features: 1. Practical feasibility. Software for MI is becoming increasingly available, and SMIKe limits the degree of imputation to a subset of variables (the keys) and cases (the sensitive cases and their mixing sets). 2. Existing MI procedures for statistical analysis measure and propagate the loss of information from SDC using SMIKe. The user is provided with a set of imputed rectangular data sets that can be analyzed using standard statistical software, and inference combined using the comparatively simple MI methods of analysis. SMIKe is particular attractive if data collectors multiply impute missing values in the data set, since MI can be then applied simultaneously to deal with missing data and provide increased disclosure protection. 3. SMIKe The size of mixing sets can be chosen to balance the gain in disclosure protection against the information loss.

SMIKe is still at an early stage of development, and more work is needed to implement the method in large-scale survey settings, to develop models that accommodate mixed variable types and clustering in the sample design, and to develop more refined measures of disclosure risk. For the latter, we considered the decision-theoretical approach suggested by Duncan and Lambert (1986), which is based on specification of some loss function and its expectation over a probability distribution of possible target values. However, as the number of sensitive cases increases this approach becomes complicated quickly, and it is unclear whether it is a good model in practice for intruder behavior. Our initial simulation studies of the method are promising, but theoretical and empirical studies of the statistical properties of the method are needed and are currently in progress.

References

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

Little, R.J.A. (1993), “Statistical Analysis of Masked Data,” *Journal of Official Statistics*, 9(2), 407-426.

Liu, C.H. and Rubin, D.B. (1998), “Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data,” *Biometrika*, 85(3), 673-688.

Olkin, I. and Tate, R.F. (1961), “Multivariate Correlation Models with Mixture Discrete and Continuous Variables,” *Annals of Mathematical Statistics*, 32, 448-465

Rubin, D.B.(1993), “Discussion of Statistical Disclosure Limitation,” *Journal of Official Statistics*, 9(2), 461-468.

Schafer, J.L.(1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.