University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2003 *Paper* 6

Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata

Rod Little* Fang Liu†

*University of Michigan, rlittle@umich.edu
†University of Michigan, fangliu@excite.com

Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata

Rod Little and Fang Liu

Abstract

The fundamental tension in statistical disclosure control (SDC) of microdata is the trade-off between the protection of individual respondents and the release of enough information for statistical inferences. We consider microdata that include key variables that contain identifying information and target variables that include sensitive information. Releasing the original data may expose some individuals in the sample to high risk of disclosure; deleting key variables is a common approach, but this loses information for some statistical analysis. This paper proposes selective multiple imputation of key variables (SMIKe) as an alternative SDC technique between those two extremes, and applies SMIKe to categorical key variables and continuous nonkey variables in the context of the general location model. Keys of sensitive cases and a mixing set of selected nonsensitive cases are multiply imputed from their posterior predictive distributions, and each set of imputed keys is released to the public with the rest of the data. The size of mixing set can be used to control the trade-off between information loss and protection. Data analysis is conducted using multiple imputation methods with some necessary correction in the case of SMIKe. Simulation studies and an application of SMIKe to the 1995 Health and Ways of Living Survey in Alameda County are also presented.

Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata

Roderick J. A. Little* **and Fang Liu****

Department of Biostatistics, University of Michigan,

Ann Arbor, Michigan 48105-2029, U.S.A.

**email:* rlittle@umich.edu

***email:* liufang@umich.edu

January 27, 2003

ABSTRACT: The fundamental tension in statistical disclosure control (SDC) of microdata is the trade-off between the protection of individual respondents and the release of enough information for statistical inferences. We consider microdata that include key variables that contain identifying information and target variables that include sensitive information. Releasing the original data may expose some individuals in the sample to high risk of disclosure; deleting key variables is a common approach, but this loses information for some statistical analysis. This paper proposes selective multiple imputation of key variables (SMIKe) as an alternative SDC technique between those two extremes, and applies SMIKe to categorical key variables and continuous nonkey variables in the context of the general location model. Keys of sensitive cases and a mixing set of selected nonsensitive cases are multiply imputed from their posterior predictive distributions, and each set of imputed keys is released to the public with the rest of the data. The size of mixing set can be used to control the trade-off between information loss and protection. Data analysis is conducted using multiple imputation methods with some necessary correction in the case of SMIKe. Simulation studies and an application of SMIKe to the 1995 Health and Ways of Living Survey in Alameda County are also presented.

Key words: Statistical disclosure control (SDC); general location model; information loss; protection; disclosure risk; Mahalanobis distance.

1 INTRODUCTION

Statistical disclosure control in microdata is an important practical topic for statistical organizations that provide public-use data sets to the research community. Today's sophisticated computer technology allows data intruders to access information and identify individuals more easily. This fact and the increased access to data through the internet and electronic

media make respondents more concerned about protection of their privacy. Protection of confidentiality is also vital for the future co-operation of respondents, and collecting high-quality data. On the other hand, the release of information in public-use data sets is important for policy makers to make efficient and timely decisions. It is the statistical agencies' obligation to balance the requirements of protection of respondents and dissemination of information.

In a microdata set containing information on individual respondents, we assume that variables can be divided into two groups: key/identifying variables and nonkey variables. Key variables $\mathbf{X}$ are characteristics of individuals (such as "age" and "sex") that are assumed to be known to data intruders from publicly available sources. Key variables 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*. If a key cell contains just one case, we call the cell *unique cell* and the case a *unique case* or *uniqueness*; if a key cell has two cases, then this is a *two-case cell*, and so on. Cells containing $\leq s$ cases, where s is a pre-specified sensitivity threshold (e.g. three or five), are called *sensitive*, and the cases belonging to them are called *sensitive cases*; a cell with more than s cases is a *nonsensitive cell* with *nonsensitive cases*. Sensitive cases are the major concern in issues of confidentiality since they have a higher risk of disclosure than cases in nonsensitive cells. Nonkey variables $\mathbf{Y}$ are assumed not to be available to intruders from external public-use database, and may include sensitive information that needs to be protected before data release. These variables can be categorical such as "HIV status" or continuous such as "income". The important question of which variables are keys and which are not depends on the setting, and is not addressed by our research.

Sensitive cases in a data set are protected by techniques that restrict access or techniques that restrict information release. The former includes legal strategies, administrative and ethical controls, and computer strategies (password, user ID, output control, electronic gatekeepers and monitors for remote access to computer database). Restriction of information is called *statistical disclosure control (SDC)* and is the main focus of data protection in statistical agencies and research institutes. We assume that releasing the full data set puts sensitive cases at excessive disclosure risk; deleting $\mathbf{x}$ gives full protection but does not allow statistical analyses involving $\mathbf{X}$ that could have been performed on the original data set. Common SDC methodologies lying between these two extremes include *global recoding, local suppression, data swapping, micro-aggregation*, and *post randomization (PRAM)*. Global recoding merges sensitive cells with other cells with low sample counts. Local suppression deletes sensitive information in some records, replacing the deleted values by missing value codes. These two techniques are often used in combination, and are available in the software program μ-ARGUS (*http://www.cbs.nl/sdc/argus.htm*) created by Statistics Netherlands They provide protection but the information loss can be large, and the missing values created by local suppression impose extra burden on data analysts. Hurken and Tiourine (1998) constructed a mathematical model for minimizing information loss from global recoding and local suppression, but the heavy computation associated with that model may be impractical for real applications.

Dalenius and Reiss (1982) proposed data swapping of categorical key variables in microdata. Greenberg (1987) suggested a refinement for ordinal or continuous key variables, called rank-based proximity swapping. In this method, records are first sorted and ranked on a continuous key variable, then swapping occurs only in records in the same neighborhood

defined by the key variable with a pre-specified size. Micro-aggregation (Defays and Anwar, 1998) can be performed on either key or sensitive variables, whether they are categorical or continuous, and is aimed at masking extreme values of the variables. The three steps of micro-aggregation are: i) segment the variables; ii) divide the cases into groups with a pre-specified size for each segmentation; iii) derive surrogate values for the variables to characterize the groups in each segmentation. To reduce information loss, cases need to be as homogeneous as possible in each group. Obviously, after micro-aggregation, variances and covariances between the variables decrease. Since micro-aggregation acts more on outlying observations while leaving the majority of the data structure intact, disclosure risk may still exist for sensitive cases that are not outlying. PRAM (Gouweleeuw, Willenborg and de Wolf, 1998) transforms categorical data for a variable with C categories by creating a $(C \times C)$ Markov probability matrix A whose $(i,j)^{th}$ entry a_{ij} is the probability of category i being transformed into category j. PRAM is performed independently on each case, and can be applied to several variable individually or in combination. The information loss from PRAM can be large (Doyle, Lane, Theeuwes, and Zayatz, 2001).

All the above SDC techniques are somewhat model-free and ad-hoc. Model-based SDC techniques replace observed values of the data by predictions based on a statistical model. An extreme version of this approach suppresses all the real data records and releases simulated microdata generated from a model estimated using the real data. The added simulation uncertainty can be reflected by multiple imputation (MI) (Rubin, 1987), a method originally proposed to handle missing data. In the SDC setting, Rubin (1993) proposes to build an imputation model from the sample data, and predict nonsampled values of survey variables for cases in the population where the sample is drawn; this imputation is repeated independently $D > 1$ times, and a sample from each of the D imputed populations is released to the public (the released sample may or may not overlap the original one). We refer to this methodology as *full MI*. Full MI has several strengths: first, valid inferences under well-specified models can be generated from the MI data sets using simple combining rules; second, information loss due to the nonrelease of the original data set can be easily assessed by the fraction of missing information from MI theory; third, no real data need to be disseminated if we restrict the release to imputed records not in the actual sample; finally, if there are already missing data in the original data set, MI solves two problems, namely, missing data and disclosure control. Raghunathan, Reiter and Rubin (2002) further develop the idea of full MI and presents analytical tools that data users can use when analyzing released multiple imputed data sets. Application of MI as an SDC tool can be found in Kennickell (1997, 1998, 2000). Other approaches to simulating pseudo-microdata include the Bayesian approach of Franconi and Seri (2001), the adjusted prediction approach of Stander (2001), and the bootstrap-like approach based on smoothed cumulative distribution function of Fienberg, Markov, and Steele (1998).

A drawback of full MI is that it requires building a statistical model for the whole population, which may be a formidable task for survey data sets. Quality of inferences from the synthetic MI data sets depends on how well this large model is specified. The imputation task is simplified by limiting imputations to a subset of the variables and/or cases ((Little, 1993)). In particular, Abowd and Woodcock (2001) discuss multiple imputation of the sensitive variables, and Reiter (2003) develops an adjusted combination rule for variance estimation when a subset of variables are masked by multiply imputation. An alternative

approach is to impede identification by applying MI to a subset of values of the key variables, since these are the variables that allow data intruders to identify their targets ((Little, 1993)). 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 reduces sensitivity of inferences to misspecification of imputation model. Reiter (2003)'s adjusted MI method can be applied to statistical inferences from SMIKed data. SMIKe can be thought of as a form of probabilistic "generalized" data swapping on categorical key variables. However, SMIKe is model-based and less ad-hoc than data swapping. SMIKe bears some similarity to PRAM as well, since both methods change the values of key variables probabilistically. However, the probability matrix that PRAM uses is the same for all cases and choices of the entries in that matrix are chosen by the user. SMIKe employs *empirically*-based probabilities that differ from case to case.

The article is organized as follows. Section 2 describes the steps of SMIKe, including selection of nonsensitive cases, construction of an imputation model, adjusted MI method for statistical inferences in SMIKe, measurement of information loss, identification of disclosure and assessment of disclosure risk and protection. Section 3 presents how SMIKe can be applied for the special case of categorical key variables and continuous nonkey variables. In Section 4, SMIKe and two model-free SDC techniques are assessed in simulation studies. A limited application of SMIKe to the 1995 Alameda County Health and Ways of Living Survey is discussed in Section 5. The paper concludes with some final remarks and topics for future research.

2 STEPS OF SMIKe

Suppose in a data set with n cases, there are a set of key variables $\mathbf{X}$, the cross-tabulation of which forms a scalar variable x with K categories; $\mathbf{Y}$ is a vector containing q nonkey variables ($\mathbf{Y}$ could be continuous, categorical or a mixture of both) and s is a chosen sensitivity threshold. n_{sen} and n_{non} are respectively the numbers of sensitive and nonsensitive cases in the data set. SMIKe consists of the following steps:

2.1 Selection of Nonsensitive Cases

Let $\mathbf{y}$ denote the set of nonkey variables to be used in model construction process with dimension $p \leq q$, 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. Specifically, for each sensitive case $i = 1, \ldots, n_{\text{sen}}$, select a mixing set $\mathcal{M}_i$ (of pre-specified size $n^i{}_{\text{mix}}$) of cases from nonsensitive cell(s) that is(are) close to the sensitive case with respect to $\mathbf{y}$. We call the nonsensitive cells – whose cases are selected into mixing sets – donor cells. The mixing sets for different sensitive cases may overlap. Various measures of closeness might be used.

For continuous $\mathbf{y}$ we use the Mahalanobis distance. The value of n^i_{mix} may vary from case to case, say, according to case sensitivity, but here for simplicity we fix n_{mix} for all sensitive cases. The value of n_{mix} serves as a tuning parameter to balance gains in protection against information loss, as discussed in Section 4. We define $\mathcal{M}$ as the union of sensitive cases and selected nonsensitive cases. $\mathcal{M}$ is a subset of $\mathcal{C}$, which is the union of sensitive cases and all the nonsensitive cases in donor cells. The number of cells (K^*) is the same in $\mathcal{M}$ and $\mathcal{C}$. The cases in $\mathcal{M}$ is denoted by n^* and these are the cases subject to imputation of keys (the percentage of cases with modified keys is thus $n^*/n \times 100\%$).

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 may involve less computation than GS in that distances are only computed for each nonsensitive *cell* and cases in the selected cells, rather than for each nonsensitive *case*. LS also restricts the mixing set to a smaller set of cells and hence may involve less information loss for some analyses. On the other hand, GS may provide better protection, since sensitive cases are mixed with cases from a wider range of key cells. We provide a limited comparison of the two approaches in simulation studies in Section 4.

The mixing sets can be further constrained to avoid information loss for particular analyses. For example, if we want to preserve the row margins of a table formed by a subset of key variables, we may only allow sensitive cases to be mixed only with cases from the same row in that table, and separate imputation model should be built for each row involved in imputation.

2.2 Construction of an Imputation Model for Keys

The basic idea behind SMIKe is to build a Bayesian model on the original data $\mathcal{D}$, then pretend x_i for cases in $\mathcal{M}$ are missing and impute these "missing" values from the posterior predictive distribution of $p(\tilde{x}_i|\mathbf{y}_i, i \in \mathcal{M}, \mathcal{D})$. For valid inferences, this model needs to condition on the mixing set, $\mathcal{M}$, which differs from the full sample by its method of selection. The predictions thus require a model for $p(x, \mathbf{y}|\mathcal{M})$. One approach is to model this distribution directly using only the data in $\mathcal{M}$, but since this set of cases is likely to be small, this could be impractical or inefficient. If selection of the mixing set is based *only* on x, then we can write

$$p(x, \mathbf{y}|\mathcal{M}) = p(x|\mathcal{M})p(\mathbf{y}|x, \mathcal{M}) = p(x|\mathcal{M})p(\mathbf{y}|x, \mathcal{D}), \tag{1}$$

since the distribution of $\mathbf{y}$ given x is not subject to bias from selection on x. This conditional distribution can be modeled using the data in $\mathcal{D}$ or the data in $\mathcal{C}$, which avoids modeling cells of x not involved in the mixing sets. The gain in information in moving to these larger sets of cases is particularly important if $\mathbf{y}$ is high-dimensional. The LS method described in Section 2.1, with mixing sets chosen randomly within the selected cells, is an example of a selection method that depends only on x. On the other hand, the GS method involves

selection of mixing sets based on both x and $\mathbf{y}$, so for this selection method models estimated on the larger data sets are potentially biased. This bias does not appear to be serious in our simulations, but does argue for LS rather than GS as the method of selection. More technical details on this issue are provided in Appendix 1.

We consider here parametric imputation models $p(x, \mathbf{y}|\boldsymbol{\theta}, \mathcal{M})$, with a noninformative prior distribution for the parameters $\boldsymbol{\theta}$. Multiple imputation is then achieved by drawing $\boldsymbol{\theta}$ from its posterior distribution, and then drawing $\tilde{x}$ for cases in $\mathcal{M}$ from its posterior predictive distribution given $\mathbf{y}$ and the drawn $\boldsymbol{\theta}$. Repeating this procedure D times and combining each set of predictions $\tilde{x}$ with the nonimputed data yields D imputed datasets. Rubin (1987) calls this a proper MI procedure since uncertainty in estimating $\boldsymbol{\theta}$ is propagated. A simpler but improper MI method is to based all D sets of draws of $\tilde{x}$ from one draw of $\boldsymbol{\theta}$. Improper SMIKe does not fully propagate imputation uncertainty and results in underestimated variances of parameters of interest. More specifics on the choice of imputation model are provided in Section 3.

2.3 Statistical Inferences for SMIKed Data

Statistical inferences for SMIKed data are easy for the data user, involving simple manipulations of complete-data analyses applied to each imputed data set. The combining rules are similar to those for missing data MI (Rubin, 1987), but as we shall see, differ in one important respect. Suppose ϕ is a scalar parameter of interest. For completed data set d $(d = 1, \ldots, D)$, let $\hat{\phi}_d$ denote an estimate of ϕ and V_d an estimate of the variance of $\hat{\phi}_d$. The MI estimate of ϕ is given by

$$\bar{\phi} = \sum_{d=1}^{D} \phi_d / D \tag{2}$$

and the estimated variance of $\bar{\phi}$ is given by T

$$\begin{aligned} T &= W + \frac{1}{D}B, \text{ where} \\ W &= \sum_{d=1}^{D} V_d/D, \text{ and } B = \sum_{d=1}^{D} (\hat{\phi}_d - \bar{\phi})^2/(D-1). \end{aligned} \tag{3}$$

T, W and B are called the total variance, within-variance and between-variance of $\hat{\phi}$, and $1/D$ is a correction factor for small D. Note that the combining rule for T in Eqn. (3) differs from the rule for missing data, namely $T = W + (1 + \frac{1}{D})B$ (Rubin, 1987). The reason for the difference is that in SMIKe the parameters are drawn from the complete data prior to masking rather than the incomplete data with values of x masked. Since the posterior distribution of the parameters is based on more information than in standard missing-data application of MI, the standard combination rule overestimates the variance of $\bar{\phi}$ and results in conservative inferences. Reiter (2003) derived Eqn. (3) for the situation where all the values of a subset of variables are multiply imputed. The method remains valid in our case, providing the MI predictive distribution takes into account the selection of mixing sets, as discussed in Section 2.2 (See Appendix II for technical details).

2.4 Assessment of Information Loss and Disclosure Risk

The information loss for inferences is found by analogous arguments to those in Rubin (1987) in the missing data case. For scalar ϕ, it is given by:

$$\gamma = \frac{B/D}{T}. \tag{4}$$

Note that this measure of information loss is smaller than for the case of missing data, where $\gamma = \frac{(1+1/D)B}{T}$, and unlike that case tends to zero as the number of MI data sets D increases. Information loss varies for different analyses, and hence might be computed for a range of analyses of interest.

The assessment of disclosure risk is difficult since it requires conjectures about the behavior of data intruder. In SMIKe the difficulties are compounded by the release of multiple imputed data sets. We first discuss the measure of disclosure risk in the original data set (R(orig)), and then present two measures of disclosure risk in SMIKed data. Other measures of disclosure risk might also be developed based on alternative assumptions about intruder behavior. We assume that (a) the data intruder's target is from the population being sampled; (b) the data intruder holds the correct values of key variables for his target; and (c) the data intruder identifies his target by matching the target's key with those of the cases in the sample. Since non-sampled cases have no risk of disclosure, we define the overall disclosure risk as

$$R = \sum_{i=1}^{n} r_i,$$

where r_i is the disclosure risk for case i in the sample, $i = 1, \ldots, n$. If an intruder knows that a particular target individual is in the sample, then simple measure of risk is

$$r_i = \begin{cases} \frac{1}{n_i} & \text{if } n_i \leq s \\ 0 & \text{if } n_i > s \end{cases} \tag{5}$$

where n_i is the number of cases in the sample in the same key cell as the target, and s is the threshold for sensitive cells. In another intruder model, the intruder does not have a specific person in mind, but has information about key variables for the whole population in a database. A simple measure of risk for sampled individuals i is then

$$r_i = 1/N_i, \tag{6}$$

where N_i is the number of cases in the population in the same key cell as individual i. In this case, risk is determined by the distribution of population counts over the key cells rather than sample counts. Assessment of this risk requires a method of estimating the population sizes N_i from the sample and any auxiliary information about the population. For research on this issue, mainly focused on the issue of assessing the probability of population unique ($N_i = 1$) given sample unique ($n_i = 1$), see Bethlehem, Keller and Pannekoek (1980); Chen and Keller-McNulty (1998); Fienberg and Markov (1998); Samuels (1998); Skinner and Holmes (1998); Skinner, Marsh, Openshaw and Wymer (1994). In our empirical work we focus on the sample measure Eqn. (5), but if estimates of the population counts are available, the second measure 6 can be applied, and would determine the choice

of sensitive cases.

Our measures of R(smik), disclosure risk of the population caused by releasing SMIKe-modified MI data, are also based on Eqn. (5). Two empirical approaches are presented. The simpler approach is to measure disclosure risk R_1^d separately in each of the D SMIKed data sets, and take an average. That is,

$$R_1(\text{smik}) = \sum_{d=1}^{D} R_1^d / D \tag{7}$$

In imputed data set d, suppose that a case i in key cell k is imputed in a cell $\tilde{k}$ containing $m_{\tilde{k}}$ cases, that is, $x_i = \tilde{k}$, then r_i^d for imputed data set d is

$$\begin{cases} 0 & \text{if } \tilde{k} \neq k \\ 0 & \text{if } \tilde{k} = k \text{ and } m_{\tilde{k}} > s \\ \frac{1}{m_{\tilde{k}}} & \text{if } \tilde{k} = k \text{ and } m_{\tilde{k}} \leq s \end{cases} \tag{8}$$

The aggregated risk R_1^d is the sum of r_i^d over $i = 1, \ldots, n$.

The measure R_1(smik) is simple and intuitive, but does not account for the fact that additional information is available if the D SMIKed data sets are considered in aggregate rather than one at a time. Our second measure R_2(smik) attempts to model the search for a target by an intruder with more sophisticated statistical understanding or tools. To define R_2(smik), consider the two-way cross-tabulation of the keys and cases in Table 1, where

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_{n1}	e_{n2}	...	e_{nk}	...	e_{nK}	$e_{n+} = D$
column total	e_{+1}	e_{+2}	...	e_{+k}	...	e_{+K}	$D \cdot n$

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 case in set $\mathcal{M}$, the collection of D MI data sets yields a sample from an independent multinomial distribution (the posterior predictive distributions of its key) with row margins fixed at D; In column k, $p_{i|k} = e_{ik}/e_{+k}$ estimates the conditional probability that case i is a unique case in target cell k, given that there is a single 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 identification rule, we measure r_i^d for a case i in true key cell k as follows:

$$\begin{cases} 0 & \text{if } p_{i|k} \neq \max_j\{p_{j|k}\} \text{ or} u_k = 0; \\ 0 & \text{if } p_{i|k} = \max_j\{p_{j|k}\} \text{ or } u_k > s; \\ \frac{1}{u_k} & \text{if } p_{i|k} = \max_j\{p_{jk}\} \text{ and } 0 < u_k \leq s. \end{cases} \tag{9}$$

Summing over all n cases, we get the second measure of disclosure risk, R_2(smik).

Absolute disclosure risk is important in applications, but the relative reduction in disclosure risk is useful for measuring the trade-off between information loss and protection. For risk measure R_j(smik) (j=1, 2), the percent reduction in disclosure risk from SMIKe is $(100 \cdot P_j)\%$, where

$$P_j = 1 - \frac{R_j(\text{smik})}{R(\text{orig})}, \ (P_j \in [0,1]) \tag{10}$$

3 IMPLEMENTATION OF SMIKe FOR CONTINUOUS y

Since x is treated as categorical, a natural choice for $p(x|\mathbf{y}, \mathcal{M})$ (ignoring refinements for 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 model (Olkin and Tate, 1961) for the joint distribution of x and $\mathbf{y}$. We describe SMIKe for that model, while outlining extensions to other situations.

According the steps outlined in Section 2, we should first select nonsensitive cases. If the nonkey variables are approximately normal, a natural measure of closeness between two cases i and j is the Mahalanobis distance $(\mathbf{y}_i - \mathbf{y}_j)^T S^{-1} (\mathbf{y}_i - \mathbf{y}_j)$, where S^{-1} is the pooled sample covariance matrix, and that from case i to a cell k is measured by $(\bar{\mathbf{y}}_k - \mathbf{y}_i)^T S^{-1} (\bar{\mathbf{y}}_k - \mathbf{y}_i)$. With the measure of closeness, both GS and LS can be applied in the selection step. With continuous and categorical nonkey variables, the Mahalanobis distance might be combined with a measure of closeness between the cells formed by the categorical nonkeys.

With continuous $\mathbf{y}$, the general location model can be used to construct an imputation model for keys, which is defined in terms of the marginal distribution of x and conditional distribution of $\mathbf{y}$ given x (for now, assume that the model is built on $\mathcal{M}$)

$$\begin{aligned} p(x_i = k) &= \pi_k, \text{ where } k = 1, \ldots, K^*; \sum_k \pi_k = 1 \\ p(\mathbf{y}_i | x_i) &\overset{indep}{\sim} N_p(\boldsymbol{\mu}_{x_i}, \Sigma) \text{ for } i = 1, \ldots, n^*. \end{aligned}$$

Transformations of $\mathbf{y}$ can be considered to improve the fit of the model. Another possible model is the extended general location model (Liu and Rubin, 1998), where covariance matrix does not have to be constant across cells 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 above 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

general location 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 size of cell k in $\mathcal{M}$. If Jeffreys' priors are used,

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K^*} \pi_k^{-\frac{1}{2}}|\Sigma|^{-\frac{p+1}{2}},$$

then posterior distributions of $\boldsymbol{\theta}$ is

$$\begin{aligned} [\boldsymbol{\pi}|x, \mathbf{y}] &\sim Dirichlet(n_1^* + \frac{1}{2}, \ldots, n_{K^*}^* + \frac{1}{2}) \\ [\Sigma|\boldsymbol{\pi}, x, \mathbf{y}] &\sim Inv - Wishart(S, n^* - K^*) \\ [\boldsymbol{\mu}_k|\boldsymbol{\pi}, \Sigma, x, \mathbf{y}] &\sim N_p(\bar{\mathbf{y}}_k, \Sigma/n_k^*) \text{ for } k = 1, \ldots, K^*, \end{aligned} \tag{11}$$

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

$$p(\tilde{x}_i = k|\boldsymbol{\theta}, x, \mathbf{y}) = \frac{\pi_k exp(\psi_{ik})}{\sum_{k'=1}^{K^*} \pi_{k'} exp(\psi_{ik'})} \text{ for } k = 1, \ldots, K^*, \tag{12}$$

where

$$\psi_{ik} = \mathbf{y}_i^T\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k \text{ (similar for } \psi_{ik'}). \tag{13}$$

The imputation process involves drawing $\boldsymbol{\theta}$, imputing $\tilde{x}_i$ given drawn $\boldsymbol{\theta}$. Based on independently imputed D data sets, information loss and disclosure risk can be assessed by the formulas presented in Section 2.

4 SIMULATION STUDIES

Simulation studies in this section evaluate the performance of SMIKe in terms of information loss and protection, examine validity of statistical inferences based on SMIKed data, and compare SMIKe with two model-free SDC techniques – data swapping and post randomization (PRAM). The simulation studies involve continuous $\mathbf{y}$, and base imputations of x on the general location model estimated on $\mathcal{M}$.

4.1 Simulation Study I

A very simple scenario with one binary key variable $x = \{1, 2\}$ and one sensitive variable y is used to illustrate how selection in SMIKe affects the trade-off between information loss and protection. One thousand data sets of size $n = 20$ are generated from the following general

location model:

$$\begin{aligned} x_i &\sim Bern(\pi) \\ y_i|x_i &\sim N(\mu_{x_i}, \sigma^2), \end{aligned}$$

where $\pi = P(x = 2) = 0.90$, $\mu_1 = 0, \mu_2 = \{0, 3\}$, $\sigma^2 = 1$. SMIKe is applied in data sets where cell $x = 1$ is sensitive ($s = 3$) while cell $x = 2$ is not. Suppose size of cell $x = 1$ is n_1 and that of cell $x = 2$ is n_2 ($n = n_1 + n_2$). Selection of mixing cases is based on Mahalanobis distance and the size of $\mathcal{M}$ is $(n_1 + n_{\text{mix}})$. Take $D = 10$ and vary n_{mix} in the range of $[2, n_2]$ to investigate how it affects the trade-off between information loss and protection in SMIKed data. Note when $n_{\text{mix}} = n_2$, SMIKe becomes full MI. The parameters of interest are μ_2 and σ^2. Examples of the imputed data sets are given in Appendix III. Figure 1 shows the effect of different n_{mix} on the trade-off between information loss and protection in data sets with $\mu_1 = 0, \mu_2 = 0$ and $\mu_1 = 0, \mu_2 = 3$ respectively. First, the levels of information loss associated with μ_2 and σ^2 are about the same in $\mu_1 = 0, \mu_2 = 0$ and $\mu_1 = 0, \mu_2 = 3$, though the former increases with n_{mix} and the later is rather stable. Second, protections in the case of $\mu_1 = 0, \mu_2 = 0$ is much higher than those in the case of $\mu_1 = 0, \mu_2 = 3$ and this tells the correctness of our conjectures about selecting nonsensitive cases that are closest to sensitive ones in terms of $\mathbf{y}$ for the sake of good protection. Third, in the case of $\mu_1 = 0, \mu_2 = 0$, when n_{mix} goes up, protection increases, as does information loss, but the change in the former is more dramatic than that in the latter. If we are satisfied with protection in range of $[0.70, 0.80]$, then n_{mix} as small as 5 or 6 is enough while information loss can still be kept low. Fourth, in the case of $\mu_1 = 0, \mu_2 = 3$, where mixing of the sensitive and nonsensitive cells is inhibited by the lack of overlap of the distribution of y in the two cells, both protection and information loss are limited and insensitive to the value of n_{mix}.

4.2 Simulation Study II

This more complex simulation study examines the properties of SMIKe and two existing SDC methods in more detail. The distribution of the key variables is based on a cross-tabulation of four categorical variables $\mathbf{X} = \{X_1, X_3, X_3, X_4\}$ on page 160 of Agresti (1990). Probabilities $(\pi_1, \ldots, \pi_{84})$ for the 84 cells were created from the fit of a loglinear model including four-way associations to this data set, and are shown in Table 2. Five hundred samples of $n = 750$ cases were then generated from the following general location model:

$$\begin{aligned} x_i &\sim \text{multinomial}\,(\pi_1, \ldots, \pi_{84}) \\ \mathbf{y}_i|x_i &\sim N_{(2)}(\boldsymbol{\mu}_{x_i}, \Sigma), \end{aligned}$$

where $x_i = \{1, \ldots, 84\}$, $i = 1, \ldots, n$ and Σ is the covariance matrix with $\sigma_1^2 = 1.0$, $\sigma_2^2 = 1.44$ and $\sigma_{12} = 1.02$ (the values of 84 sets of mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ are not presented here).

In each of the simulated samples, we vary the sensitivity threshold s from 3 to 10 to obtain different proportions of sensitive cases. The sensitivity level of a sample is summarized by $s_1 =$ the proportion of sensitive cells among all key cells, and $s_2 =$ the proportion of sensitive cases in the sample data. Variation of s allows us to vary the amount of data modification by the SDC techniques. The parameters from two multiple linear regressions – y_1 on $(\mathbf{X}, y_2)$ and y_2 on $(\mathbf{X}, y_1)$ – are used to assess information loss and quality of statistical inferences

Table 2: *Population Key Cell Structure* ($\boldsymbol{\pi} \cdot 100$)

		X_3	1		2		3	
X_1	X_2	X_4	1	2	1	2	1	2
1	1		2.64	1.38	2.17	1.64	1.61	0.93
	2		2.37	2.43	0.86	1.33	0.44	0.83
2	1		1.33	0.85	0.43	0.13	0.64	0.07
	2		0.36	0.56	0.20	0.15	0.14	0.07
3	1		3.26	3.23	1.15	1.03	0.80	0.60
	2		2.10	2.56	0.33	0.38	0.11	0.17
4	1		2.95	4.53	1.41	1.69	1.35	0.97
	2		1.64	4.08	0.70	1.41	0.34	0.65
5	1		2.55	1.75	0.73	0.68	1.02	0.39
	2		1.40	1.14	0.40	0.45	0.32	0.24
6	1		4.76	3.28	1.06	0.49	1.02	0.11
	2		2.01	2.44	0.42	0.36	0.32	0.10
7	1		3.03	1.70	1.08	0.69	0.69	0.30
	2		1.27	0.91	0.69	0.58	0.30	0.27

(in all 24 parameters, including 2 intercepts, 20 regression coefficients associated with $\mathbf{X}$, 1 associated with y_2 and 1 associated with y_1). We try both LS and GS to select cases for $\mathcal{M}$ with n_{mix} set at 5. Both improper and proper versions of SMIKe are tried, and statistical inferences based on the standard and adjusted MI methods are also compared.

SMIKe is compared with three existing SDC methods, two versions of data swapping and PRAM. In random data swapping (RDS), values for a random $r\%$ of cases are swapped. Sanil, Gomatam and Karr (2002) suggest r between $1 \sim 10\%$. In our setting w choose $r = s_2$, so that the fraction of swapped cases goes up with the sensitivity threshold. Obviously, RDS leaves some sensitive cases unprotected due to random selection of cases for swapping. To give full protection on sensitive cases, we also consider deterministic data swapping (DDS), where all sensitive cases are swapped with cases randomly chosen from other cells. For PRAM, we employ the method proposed in Gouweleeuw, Willenborg and de Wolf (1998) where the Markov matrix A is chosen so that the joint distribution of transformed key variables is invariant with respective to A (invariant PRAM). The entries in a_{kl} in A is defined as

$$a_{kl} = \begin{cases} 1 - (\theta\, T(K)/T(k)) & \text{if } l = k \\ \theta\, T(K)/((K-1)T(k)) & \text{if } l \neq k\,, \end{cases}$$

where k is the original key cell of a case in the data set and l is the candidate cell after transformation. Without loss of generality, we let $T(K)$ be the minimum of key cell counts, which is 1 in our case. With $T(K) = 1$, $\theta \in (0, 1)$ means the probability (a_{kk}) that cases in cell k remain in their original cell is $1 - \theta/T(k)$. If $T(k) = 1$, then $a_{kk} = 1 - \theta$. Therefore,

if we want to obtain good protection for unique cases, then the likelihood of the uniqueness being transformed out of their original cells should be high and θ should be close to 1. The probability that cases remain in original cell k increases with $T(k)$. For an instance, with $\theta = 0.99$ for cases in a 3-case cell, the probability of remaining in original cell is 0.67, which is unsatisfactorily high. Hence, PRAM focuses emphasizes the protection of unique cases over those will small counts greater than one. With our measure of disclosure risk (Eqn. (5)), we expect the protection given by PRAM to be unsatisfactory since *each* case bears risk of divulgence. In this study, we fix $\theta = 0.99$ for all eight sensitivity levels; a value close to its upper limit of 1 seems needed to provide adequate protection under this method.

Measures of information loss are restricted to SMIKe, since measures of information loss are not available from data swapping and PRAM, which release only one modified data set. All the SDC methods can be compared with respect to the measures of protection and the quality of the statistical inferences, assessed by the change in nominal 95% confidence interval (CI) width measured by CI(SDC)/CI(ori)-1, the coverage probability (CP) of nominal 95% CI and estimation bias of the parameters from the two linear regressions. Measurements of protection are presented in Table 3. In DDS, RDS and data swapping, there is only one measure of protection P_1, since the other measure P_2 is only relevant for multiply imputed data sets. In DDS, $P_1 = 1$, since all sensitive cases are swapped to other cells. The two rows labeled as "Sensitivity Index" present the average of s_1 and s_2 over the 500 simulated data sets. Differences in the estimates of protection between GS and LS and the proper and improper SMIKe methods are relatively small, with P_1 giving higher estimates of protection than P_2. As s_1 and s_2 increase, P_2 increases while P_1 is quite stable in SMIKe with LS but goes up in GS. RDS and PRAM are somewhat inferior to SMIKe, particularly at low sensitivity levels.

Figure 2 displays the information loss of the four versions of SMIKe for the 24 parameters, for each of the 8 sensitivity levels. We conclude that 1) as expected, the information loss increases with the sensitivity threshold; 2) information loss of intercepts and regression coefficients associated with $\mathbf{x}$ are greater and change more dramatically with (s_1, s_2) than the information loss of regression coefficients associated with $\mathbf{y}$; 3) proper SMIKe entails more estimated information loss than its improper counterparts; 4) GS causes slightly more information loss than LS. Figure 3 displays the change of CI width of the modified data relative to the CI based on the original data. CI width increases after SMIKe, which propagates the information loss in the inferences, but not consistently higher or lower after data swapping or PRAM, since these methods do not propagate the loss of information from the SDC modifications. The trends of inflation of CI width in SMIKe are similar to those of information loss, as would be expected. Estimation biases of the 24 parameters are plotted in Figure 4. Data swapping (both RDS and DDS) have the most serious biases, and PRAM also has large biases. Biases are much smaller for the SMIKe methods. For SMIKe and data swapping, estimation biases go up with (s_1, s_2) Biases for PRAM are insensitive to s_1 and s_2, since the degree of switching is determined by A, which is fixed at all sensitivity levels. Biases in the estimates of the intercepts and x-coefficients are more affected by changes in sensitivity than other parameters; Biases of the y-coefficients are small and not strongly affected by choice of sensitive level.

Since data swapping and PRAM yield biased estimates and do not propagate modification uncertainty, we would not expect them to yield confidence intervals with the nominal

Table 3: *Protection in Data Sets Modified by SMIKe, RDS, DDS and PRAM*

		Sensitivity Index							
SDC	$\bar{s}_1$:	0.253	0.325	0.388	0.446	0.497	0.543	0.584	0.619
Technique	$\bar{s}_2$:	0.054	0.088	0.123	0.162	0.202	0.243	0.284	0.324
SMIKe (GS,proper)	P_1 :	0.885	0.855	0.865	0.883	0.902	0.916	0.928	0.936
	P_2 :	0.879	0.835	0.719	0.660	0.609	0.615	0.620	0.656
SMIKe (GS,improper)	P_1 :	0.878	0.870	0.882	0.898	0.914	0.925	0.933	0.938
	P_2 :	0.876	0.801	0.715	0.639	0.606	0.597	0.623	0.652
SMIKe (LS,proper)	P_1 :	0.920	0.906	0.894	0.894	0.900	0.905	0.911	0.916
	P_2 :	0.880	0.840	0.759	0.731	0.714	0.704	0.701	0.699
SMIKe (LS,improper)	P_1 :	0.913	0.910	0.903	0.904	0.908	0.913	0.917	0.920
	P_2 :	0.876	0.856	0.756	0.732	0.713	0.701	0.700	0.696
DDS	P_1 :	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
RDS ($r = s_2$)	P_1 :	0.755	0.708	0.677	0.664	0.661	0.669	0.681	0.697
PRAM	P_1 :	0.629	0.640	0.638	0.628	0.630	0.625	0.599	0.577

coverage. The arguments in Section 2 suggest that proper SMIKe methods with the adjusted MI estimate of variance should give asymptotically valid CIs, while those based on standard MI estimates of variance should be conservative. To test these conjectures, Figure 5 presents the coverage rate (CP) of nominal 95% CIs from SMIKe, data swapping, PRAM and the original data. The anti-conservatism of inferences based on data swapping and PRAM is clear. In fact for data swapping, CP can be as low as ~ 0 as a consequence of bias and underestimated variance. The expected conservatism of SMIKe with *standard* MI variance formulas is apparent as well – almost all the CP's are above those from original data and some of them even give $\sim 100\%$ coverage. This conservatism is greatly reduced by using adjusted MI inference method, as can be seen in the first plot. CP from SMIKe with adjusted MI method is about at the same level as that from original data. The anti-conservatism expected for improper SMIKe relative to proper SMIKe is not obvious in the simulation study, perhaps because the fraction of missing information is not large.

4.3 Simulation Study III

Study III provides a simple illustration of how the selection of the mixing sets can affect SMIKe statistical inferences if it is not reflected in the prediction model for x. One thousand data sets of $n = 50$ independent cases on two variables (x, y) are independently simulated from the following model:

$$\begin{aligned} x_i &\sim Bern(0.7), \text{ where } x_i = 1, 2 \text{ and } P(x_i = 2) = 0.7 \\ y_i|x_i &\sim N(0, 1.0). \end{aligned}$$

The following five scenarios for selection and imputation are implemented:

I. randomly select $n_1 + n_{\text{mix}}$ ($n_{\text{mix}} = 1, \ldots, n_2$) cases from all n cases into set $\mathcal{M}$; build the imputation model on set $\mathcal{C}$ (which is the whole data set) to impute x for cases in $\mathcal{M}$.

II. put all n_1 cases in cell $x = 1$ into $\mathcal{M}$, at the same time, randomly select n_{mix} cases from cell $x = 2$ into $\mathcal{M}$; build imputation model on $\mathcal{M}$.

III. selection of cases is the same as in scenario II, but build imputation model on $\mathcal{C}$.

IV. put all n_1 cases in cell $x = 1$ into $\mathcal{M}$, at the same time, select n_{mix} cases from cell $x = 2$ into $\mathcal{M}$ based on Mahalanobis distance; build imputation model on $\mathcal{M}$.

V. selection of cases is the same as in scenario IV, but build imputation model on $\mathcal{C}$.

In all cases $D = 10$ and adjusted MI method is applied for inferences. According to the development presented in Appendix I, inferences based on SMIKed data should be valid in Scenario (SC) I, II, III, IV but invalid in SC V. Figure 6 confirms this conclusion visually, presenting the CP of nominal 95% CIs of μ_1, μ_2, σ^2 based on adjusted MI method. Results from SC V are extremely poor, but CP for the other scenarios are close to nominal levels.

5 APPLICATION

5.1 SMIKe Applied to Alameda County Health and Ways of Living Survey

The Alameda County Health and Ways of Living Survey is a longitudinal sample survey that explores the influence of health practice and social relationships on the physical and mental practice health of a typical sample of the population. We illustrate the application of SMIKe on a small subset of the data from the 1995 panel. Among the 112 variables, there are two continuous variables: y_1="hours working as volunteer per week (volhrs)" and y_2="hours working as employee per week (emphrs)", both of which have data only for cases with values of other two variables –"volunteer?" and "currently employed at paid job?"– being "yes". As an application SMIKe to continuous $\mathbf{y}$, we consider only the subset of cases have data for either "volhrs" or "emphrs". This subset of data has sample size $n = 1349$ and is denoted by $\mathcal{D}$. We divide the cases in $\mathcal{D}$ 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 key variables: "age" (measured in years), "sex", "race", "retired?(rtrd)" and "student?(stdnt)". Prior to application of SMIKe, we recoded "age" into 6 categories to reduce the number of key cells. Table 4 summarizes the extracted data for this illustrative example.

The sensitivity threshold s is set at three and LS is used as the selection plan with $n_{\text{mix}} = 5$ for all sensitive cases. The number of sensitive cells and their mixing cases and cells are given in Table 5. Note that even when the number of sensitive cases is large, the number of mixing cells in $\mathcal{M}$ can be small, because of overlap in the mixing sets of sensitive cases. Before constructing the general location models, a logarithm transformation is applied

Table 4: *Key and Continuous nonkey Variables in $\mathcal{D}$*

	Key Variables ($\mathbf{x}$)					Sensitive Variables ($\mathbf{y}$)	
	age	sex	race	rtrd	stdnt	volhrs (y_1)	emphrs (y_2)
# of categories	6	2	9	2	2	continuous	continuous

Table 5: *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

to y_1 to correct for right skewness. 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 imputation models are fitted to cases in $\mathcal{M}$ respectively in three groups and keys of cases in $\mathcal{M}$ are imputed independently for 10 times.

The parameters we choose for assessing information loss are means of $\mathbf{y}$ in the mixing cells 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 the data. Both Z_1 and Z_2 have 4 categories in the order of "1=excellent, 2=good, 3=fair, 4=poor", and the 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 6 and Figure 7. Table 6 shows the effects of SMIKe on $\mathcal{D}$ in terms of both information loss of the means of the mixed cells and protection; Figure 7 shows the information loss of the parameters from the two logistic regressions (all the measures are Monte-Carlo estimates based on 500 repetitions).

5.2 Results and Conclusions

The last two columns in Table 6 show that SMIKe provides a high degree of increased protection in this example for both P_1 and P_2 (disclosure risk is reduced by $\sim 98\%$). There are at least two reasons for this favorable result. 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 as well. 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

Table 6: *Information Loss vs. Protection in SMIKed Data*

group	InfoLoss		Mixing Cell[a]		Protection	
	μ_1	μ_2	n_k^*	n_k	P_1	P_2
	0.203(0.085)	-	6	6	0.987(0.004)	0.978(0.016)
	0.084(0.033)	-	5	13		
1	0.044(0.022)	-	5	18		
	0.048(0.021)	-	5	19		
	0.010(0.005)	-	5	57		
	0.066(0.028)	-	10	17		
	0.160(0.068)	0.162(0.062)	5	5		
	0.183(0.090)	0.242(0.121)	5	6		
2	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		
	-	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		
3	-	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] Size of mixing cells in $\mathcal{M}$ and $\mathcal{C}$. Some sensitive cases choose mixing cases from the same cell with or without overlapping, this is why n_k^*'s in some cells are greater than the pre-specified "$n_{\text{mix}} = 5$".

for the intercepts and coefficients of the nonkey variables is negligible. One explanation of this is that 4 out of 5 key variables 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, with minor losses in information for statistical analysis.

6 DISCUSSION

In this paper, we have discussed SMIKe as an SDC tool in microdata and presented its application to data with continuous nonkey variables. We propose two alternative selection plans, present an imputation model for keys, provide the measures of information loss and protection, describe confidence interval estimation from SMIKed data, and indicate how the size of mixing set can be varied to control the trade-off between protection and information loss. The simulation studies and real-data analysis suggest that SMIKe is a promising tool for SDC in microdata. Useful features of SMIKe are:

1) Practical feasibility. The increasing popularity of MI and associated development of computer software make it increasingly easy to implement SMIKe in practice. Examples of statistical software with implementation of MI include: SAS, Stata, SOLAS, IVEware (*http://www.isr.umich.edu/src/smp/ive/*). SMIKe limits the imputation to a subset of variables (the keys) and cases (the sensitive cases and their mixing sets). This feature of SMIKe limits computational cost, saves storage space, and limits the impact of misspecification of the imputation model.

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 adjusted MI methods of analysis. Other SDL methods, such as noise injection or swapping, do not measure or propagate this information loss.

3) SMIKe provides a tool for balancing the gain in disclosure protection against the information loss, namely the size of mixing sets. Our can deliver satisfactory protection without major information loss, and our theory suggests that information loss can be reduced to negligible levels by increasing the number of multiply imputed data sets.

4) 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.

SMIKe is still at an early stage of development, and more work is needed to implement the method in large-scale survey settings. We based SMIKe on a general location model for continuous outcomes, with transformations to accommodate lack of normality or a constant covariance matrix of the continuous variable. Extensions such as the extended general location model (Liu and Rubin, 1998) might improve the fit to the data and the robustness of inferences. We have outlined extensions to data involving mixtures of continuous and categorical variables, but details of these models need more development, such as measures of closeness between sensitive cases and cases in the mixing sets. Our imputation model assumes independence of respondents, which is violated in complex sample designs involving clustering of units. A useful extension would be to build an imputation model that accounts for clustering via random effects.

Measures of disclosure risk and protection are always a concern in SDC. We proposed several empirical measures for SMIKe, but refinements or better methods are an avenue for further research. 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.

Acknowledgments: 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.

REFERENCES

Abowd J. and Woodcock, S. (2001), "Disclosure Limitation in Longitudinal Linked Data", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies,* New York: North Holland.

Agresti, A. (1990), *Categorical Data Analysis,* New York: John Wiley, pp. 160.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer Verlag.

Bethlehem J.G., Keller,W.J., Pannekoek,J. (1990), "Disclosure Control of Microdata," *Journal of American Statistical Association*, 85, 38-45.

Chen, Guang and Keller-McNulty Sallie (1999), "Estimation of Identification Disclosure Risk in Microdata," *Journal of Official Statistics*, 14, 79-95.

Cox, L.H. (1980), "Suppression Methodology and Statistical Disclosure Control," *Journal of American Statistical Association*, 75, 377-385.

Dalenius, T. and Reiss, S.P. (1982) "Data-Swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inferences*, 6, 73-85.

Defays, D. and Anwar, M.N. (1998), "Masking Microdata Using Micro-aggregation," *Journal of Official Statistics*, 14, 449-461.

Domingo-Ferrer, J. and Mateo-Sanz, J.M.(2001), "An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss and Disclosure Risk," Working paper No. 5 at Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, March 2001, Skopje.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (2001), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science: North-Hollande.

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

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

Fienberg, S.E. Markov, U.E. and Steele R.J. (1998)," Disclosure Limitation Using Perturbation and Related Methods for Categorical Data," *Journal of Official Statistics*, 14, 485-502.

Fienberg, S.E. and Willenborg, L. (1994), "A Radical Proposal for the Provision of Microdata Samples and the Preservation of Confidentiality," Technical Report, Carnegie Mellon University, Department of Statistics.

Franconi, L. and Seri, G. (2001), "Experience on Model-based Disclosure Control Limitation," Working paper No. 15 at Joint ECE/Europe Work Session on Statistical Data Confidentiality, March 2001, Skopje.

Greenberg, B (1987), "Rank Swapping for Masking Ordinal Microdata," Unpublished manuscript, US Census of Bureau.

Gouweleeuw, P.K., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998), "Post Randomization for Statistical Disclosure Control: Theory and Implementation," *Journal of Official Statistics*, 14, 463-478.

Hurkens, C.A.J. and Tiourine, S.R. (1998), "Models and Methods for the Microdata Protection Problem," *Journal of Official Statistics*, 14, 437-447.

Gelman, A.B., Carlin, J.S, Stern, H.S.,and Rubin, D.B. (1995), *Bayesian Data Analysis*, Chapman & Hall/CRC.

Kennickell, A.B. (1997), "Multiple Imputation and Disclosure Protection: the Case of 1995 Survey of Consumer Finances," Working Paper, Survey of Consumer Finances.

Kennickell, A.B. (1998), "Multiple Imputation in the Survey of Consumer Finances," Working Paper, Survey of Consumer Finances.

Kennickell, A.B. (2000), "Wealth Measurement in the Survey of Consumer Finances: Methodology, and Direction for Future Research," Working Paper, Survey of Consumer Finances.

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

Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, New York: John Wiley and Sons.

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.

Oganian, A. and Domingo-Ferrer, J. (2001), "On the Complexity of Micro-aggregation," Working paper No. 6 at Joint ECE/Europe Work Session on Statistical Data Confidentiality, March 2001, Skopje.

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

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2002), "Multiple Imputation for Statistical Disclosure Limitation," Technical Report, Department of Biostatistics, University of Michigan, Ann Arbor.

Reiter, J.P. (2003), "Inferences for partially Synthetic, Public Use Microdata Sets," Unpublished manuscript.

Rubin, D.B.(1987), *Multiple Imputation for Nonresponse in Survey*, New York: John Wiley and Sons.

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

Samuels, S.M.(1998), "A Bayesian, Species-Sampling-Inspired Approach to the Unique Problem in Microdata Disclosure Risk Assessment," *Journal of Official Statistics*, 14(4), 373-383.

Sanil, A., Gomatam, S. and Karr, F. A.(2002), "NISS WebSwap: A Web Service for Data Swapping," Technical report for Digital Government Project as National Institute of Statistical Sciences.

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

Skinner, C.J. and Holmes, D.J.(1998), "Estimating the Re-identification Risk per Record in Microdata," *Journal of Official Statistics*, 14(4), 361-372.

Skinner, Marsh, Openshaw, and Wymer (1994), "Disclosure Control for Census Microdata," *Journal of Official Statistics*, 10(1), 31-51.

Stander, J. (2001), "A Model-based Disclosure Limitation Method for business Data," Working paper No. 22 at Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, March 2001, Skopje.

APPENDIX I: CONSTRUCTION OF IMPUTATION MODEL IN THE CASE OF CONTINUOUS y

This appendix provide more detail on model building issues. Denote keys and nonkey variables of the cases in set $\mathcal{M}$ by $\tilde{x}_1$ and $\mathbf{y}_1$. Imputation of $\tilde{x}_1$ in $\mathcal{M}$ is performed in two steps: draw parameters $\boldsymbol{\theta}$ from their posterior distributions, then draw $\tilde{x}_1$ given these drawn parameters from its full conditional posterior distribution. We consider three different ways to build the imputation model.

1. *Selection of cases is based on Mahalanobis distance and imputation model built on selected cases $\mathcal{M}$.* If the model estimation is based only on information in $\mathcal{M}$, valid predictions of x_1 result if the model is well specified. In symbols:

$$\begin{aligned} & p(\tilde{x}_1|\mathbf{y}_1, \mathcal{M}) \\ = & \int p(\tilde{x}_1|\boldsymbol{\theta}, x_1, \mathbf{y}_1, \mathcal{M})\, p(\boldsymbol{\theta}|x_1, \mathbf{y}_1, \mathcal{M})\, d\boldsymbol{\theta} \\ = & \int p(\tilde{x}_1|\boldsymbol{\theta}, x_1, \mathbf{y}_1, \mathcal{M})\, p(\boldsymbol{\pi}|x_1, \mathbf{y}_1, \mathcal{M})\, p(\boldsymbol{\mu}, \Sigma|x_1, \mathbf{y}_1, \mathcal{M})\, d\boldsymbol{\theta} \\ = & \int p(\tilde{x}_1|\boldsymbol{\theta}, x_1, \mathbf{y}_1, \mathcal{M})\, p(\boldsymbol{\pi}|x_1, \mathcal{M})\, p(\boldsymbol{\mu}, \Sigma|x_1, \mathbf{y}_1, \mathcal{M})\, d\boldsymbol{\theta} \qquad (14) \\ \propto & \int p(\tilde{x}_1|\boldsymbol{\theta}, x_1, \mathbf{y}_1, \mathcal{M})\, p(x_1|\boldsymbol{\pi}, \mathcal{M}) p(\boldsymbol{\pi})\, p(\mathbf{y}_1|x_1, \boldsymbol{\mu}, \Sigma, \mathcal{M}) p(\boldsymbol{\mu}, \Sigma)\, d\boldsymbol{\theta} \\ = & \int p(\tilde{x}_1|\boldsymbol{\theta}, x_1, \mathbf{y}_1, \mathcal{M})\, p(x_1|\boldsymbol{\pi}, \mathcal{M}) p(\mathbf{y}_1|x_1, \boldsymbol{\mu}, \Sigma, \mathcal{M})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}\,. \qquad (15) \end{aligned}$$

   However, as stated in Section 2, when $\mathcal{M}$ is a small set, imputation may be inefficient.

2. *Selection of cases is based on Mahalanobis distance and imputation model for distribution of $\mathbf{y}|x$ built on all cases in donor cells $\mathcal{C}$.* Note that the posterior distributions of the parameters of the distribution of x must be based on $\mathcal{M}$ since selection is based on x. Symbolically, this imputation scheme replaces Eqn. (14) with

$$\begin{aligned} & \int p(\tilde{x}_1|\boldsymbol{\theta}', x_1, \mathbf{y}_1, \mathcal{M})\, p(\boldsymbol{\pi}|x_1, \mathcal{M})\, p(\boldsymbol{\mu}', \Sigma'|x, \mathbf{y}, \mathcal{C})\, d\boldsymbol{\theta}' \\ \propto & \int p(\tilde{x}_1|\boldsymbol{\theta}', x_1, \mathbf{y}_1, \mathcal{M})\, p(x_1|\boldsymbol{\pi}, \mathcal{M}) p(\boldsymbol{\pi})\, p(\mathbf{y}|x, \boldsymbol{\mu}', \Sigma', \mathcal{C}) p(\boldsymbol{\mu}', \Sigma')\, d\boldsymbol{\theta}' \\ = & \int p(\tilde{x}_1|\boldsymbol{\theta}', x_1, \mathbf{y}_1, \mathcal{M})\, p(x_1|\boldsymbol{\pi}, \mathcal{M}) p(\mathbf{y}|x, \boldsymbol{\mu}', \Sigma', \mathcal{C})\, p(\boldsymbol{\pi}) p(\boldsymbol{\mu}', \Sigma')\, d\boldsymbol{\theta}'\,. \qquad (16) \end{aligned}$$

   Though this approach is attractive because $\mathcal{C}$ is generally much larger than $\mathcal{M}$, it is improper. The reason lies in the selection of $\mathcal{M}$ based on $\mathbf{y}$, which implies that the distribution $p(\mathbf{y}|x, \boldsymbol{\mu}', \Sigma', \mathcal{C})$ is different from $p(\mathbf{y}_1|x_1, \boldsymbol{\mu}, \Sigma, \mathcal{M})$ and $(\boldsymbol{\mu}, \Sigma) \neq (\boldsymbol{\mu}', \Sigma')$

3. *Selection of cases is random from a selected cell and imputation model is built on all cases in $\mathcal{C}$.* Imputation model built on $\mathcal{C}$ works this time since the selected cases in a cell comprise a **random sample** from that cell, that is

$$p(\mathbf{y}|x, \boldsymbol{\mu}, \Sigma, \mathcal{C}) \approx p(\mathbf{y}|x, \boldsymbol{\mu}, \Sigma, \mathcal{M})\,.$$

   In fact, it is better to use all the cases in the cell than random subsample to build the model, for improved efficiency. This approach is more natural for LS than for GS.

APPENDIX II: ADJUSTED STATISTICAL INFERENCES IN SMIKED DATA

As given in Eqn. (3) in Section 2, T is expressed as

$$T = W + \frac{1}{D}B$$

If we let $D \to \infty$, then $T \to W$. That is, asymptotically, within variance is the posterior variance of $\bar{\theta}$. We review Reiter;s (2003) argument leading to the above equation and verify its validity in the case of selection. We define $\Omega_0 = (x, \mathbf{y})$ as the original data, $\Omega_D = \{\omega_1, \ldots, \omega_D\} = (\tilde{x}, \mathbf{y})$ as D sets of imputed data. The posterior mean of θ based on Ω_0 is $q_0 = E(\theta|\Omega_0)$ and posterior variance of θ is $v_0 = V(\theta|\Omega_0)$. With Ω_D, we should base our inferences about θ on $f(\theta|\Omega_D)$, or just $E(\theta|\Omega_D)$ and $V(\theta|\Omega_D)$ with standard large-sample-theory arguments. We can rewrite $f(\theta|\Omega_D)$ as

$$\begin{aligned} f(\theta|\Omega_D) &= \int f(\theta|\Omega_0, \Omega_D) f(\Omega_0|\Omega_D) d\Omega_0 \quad (17) \\ &= \int f(\theta|\Omega_0) f(\Omega_0|\Omega_D) d\Omega_0, \quad (18) \end{aligned}$$

Eqn. (18) follows since generation of Ω_D totally depends on Ω_0 with noninformative prior on θ. By standard large-sample Bayesian arguments (Gelman, Carlin, Stern and Rubin , 1995),

$$f(\theta|\Omega_0) \approx N(q_0, v_0), \quad (19)$$

which can be further simplified by replacing Ω_0 by (q_0, v_0) on the left-hand side, that is,

$$(\theta|q_0, v_0) \sim N(q_0, v_0). \quad (20)$$

The above transformation, together with Eqn. (18), implies that to obtain $f(\theta|\Omega_D)$, we can first draw q_0 and v_0 from their conditional sampling distributions given Ω_D, then draw θ from $N(q_0, v_0)$ given drawn q_0 and v_0. Repeating the drawing process, finally we will have a sample of θ from distribution $f(\theta|\Omega_D)$.

To obtain sampling distribution of (q_0, v_0) given Ω_D, we first examine the distribution (q_d, v_d) given Ω_0 where $q_d = E(\theta|\omega_d)$ and $v_d = V(\theta|\omega_d)$ for $d = 1, \ldots, D$. If the prediction model for $\tilde{x}$ is well-specified, then with the aid of standard large-sample Bayesian arguments it is reasonable to assume that

$$(q_d|\Omega_0) \sim N(q_0, B), \quad (21)$$

where B is the large-sample variance of q_d given Ω_0. Numerically, B can be attained by Monte-Carlo method as

$$\begin{aligned} B &= \lim_{D\to\infty} : \sum_{d=1}^{D} (q_d - \bar{q}_\infty)^2 / D, : \text{where} \quad (22) \\ \bar{q}_\infty &= \lim_{D\to\infty} : \sum_{d=1}^{D} q_d / D = q_0 \quad (23) \end{aligned}$$

A similar assumption is made for the distribution of v_d given Ω_0, that is,

$$(v_d|\Omega_0) \sim N(v_0, \ll B), \quad (24)$$

where $\ll B$ means variance of v_d given Ω_0 has lower-order variability than that of q_d (Rubin, 1987).

Assuming flat priors on (q_0, v_0), with sampling distributions of (q_d, v_d) presented in Eqn. (21) and (24) and the application of standard Bayesian theory, we can derive the conditional distributions of (q_0, v_0) given $\Omega_D = \{\omega_1, \ldots, \omega_D\}$:

$$(q_0|\Omega_D) \sim N(\bar{q}_D, B/D) \tag{25}$$
$$(v_0|\Omega_D) \sim N(\bar{v}_D, \ll B/D) \tag{26}$$

where $\bar{q}_D = \sum_{d=1}^{D} q_d/D$ and $\bar{v}_D = \sum_{d=1}^{D} v_d/D$. When D is finite, we have B approximated by $b_D = \sum_{d=1}^{D} (q_d - \bar{q}_D)^2/(D-1)$. That is,

$$(q_0|\Omega_D) \sim N(\bar{q}_D, b_D/D) \tag{27}$$
$$(v_0|\Omega_D) \sim N(\bar{v}_D, \ll b_D/D), \tag{28}$$

Based on $f(\theta|\Omega_D)$ presented in Eqn. (18), we have

$$\begin{aligned} E(\theta|\Omega_D) &= E(E(\theta|\Omega_0, \Omega_D)|\Omega_D) = E(E(\theta|\Omega_0)|\Omega_D) = E(q_0|\Omega_D) \\ V(\theta|\Omega_D) &= E(V(\theta|\Omega_0, \Omega_D)|\Omega_D) + V(E(\theta|\Omega_0, \Omega_D)|\Omega_D) \\ &= E(V(\theta|\Omega_0)|\Omega_D) + V(E(\theta|\Omega_0)|\Omega_D) \\ &= E(v_0|\Omega_D) + V(q_0|\Omega_D). \end{aligned}$$

With Eqn. (27) and Eqn. (28), the above two equations can be further reduced to

$$E(\theta|\Omega_D) = E(q_0|\Omega_D) = \bar{q}_D \tag{29}$$
$$V(\theta|\Omega_D) = E(v_0|\Omega_D) + V(q_0|\Omega_D) = \bar{v}_D + b_D/D \tag{30}$$

Reiter (2003) provides formulas for inferences with small D, which, "in a similar spirit to Rubin (1987)", approximates sampling distribution of θ given Ω_D by a t-distribution with degrees of freedom $\nu = (D-1)(1+\gamma^{-1})^2$, where

$$\gamma = \frac{b_D/D}{\bar{v}_D + b_D/D} : . \tag{31}$$

With selection, some cases have their x replaced by predicted values $\tilde{x}$, and others retain their original x in imputed data sets. As long as the imputation model of keys is validly constructed (for example, by building the model on $\mathcal{M}$ instead of $\mathcal{C}$ with nonrandom selection of nonsensitive cases), we would expect the assumptions about $p(q_d|\Omega_0)$ and $p(v_d|\Omega_0)$ in Eqn. (21) and (24) to still hold. Following the same steps as the above, we arrive at the same conclusion as given in Eqn. (29) and (30).

APPENDIX III: EXAMPLES OF SMIKed DATA

This appendix provides two examples of SMIKed data sets from simulation study I. There are 20 cases in the data set with the first one being a uniqueness. n_{mix} is set at 6. Cases in $\mathcal{M}$ are marked with $*$.

Table 7: Example 1: $\mu_1 = \mu_2 = 0, \sigma^2 = 1, \pi = 0.9$

original x	y	10 sets of imputed x
*0	-1.498256	1 1 1 1 0 1 1 1 1 1
1	2.196066	1 1 1 1 1 1 1 1 1 1
1	-0.308010	1 1 1 1 1 1 1 1 1 1
1	2.473768	1 1 1 1 1 1 1 1 1 1
*1	-1.703345	1 1 1 1 1 1 0 1 1 1
1	-0.039146	1 1 1 1 1 1 1 1 1 1
1	0.516418	1 1 1 1 1 1 1 1 1 1
1	-0.392493	1 1 1 1 1 1 1 1 1 1
1	1.503878	1 1 1 1 1 1 1 1 1 1
1	0.197190	1 1 1 1 1 1 1 1 1 1
1	-0.481482	1 1 1 1 1 1 1 1 1 1
1	1.122731	1 1 1 1 1 1 1 1 1 1
*1	-2.257808	1 1 0 1 1 1 0 0 1 1
1	-0.337457	1 1 1 1 1 1 1 1 1 1
*1	-0.578633	1 1 1 1 1 1 1 1 0 1
*1	-1.979207	1 1 1 1 1 1 0 1 1 1
1	1.507346	1 1 1 1 1 1 1 1 1 1
*1	-1.482834	0 1 1 1 1 1 1 1 1 1
*1	-0.510513	0 1 1 0 1 1 1 1 1 0
1	0.691961	1 1 1 1 1 1 1 1 1 1

Table 8: Example 2: $\mu_1 = 0, \mu_2 = 3, \sigma^2 = 1, \pi = 0.9$

original x	y	10 sets of imputed x
*0	0.137912	1 1 1 0 1 0 0 1 0 0
*1	0.165671	1 0 0 1 0 1 1 1 1 0
1	3.124605	1 1 1 1 1 1 1 1 1 1
1	3.846479	1 1 1 1 1 1 1 1 1 1
1	3.553457	1 1 1 1 1 1 1 1 1 1
*1	1.720408	1 1 1 1 1 1 1 1 1 1
1	2.778747	1 1 1 1 1 1 1 1 1 1
1	4.479235	1 1 1 1 1 1 1 1 1 1
*1	0.856713	1 1 1 0 1 0 1 1 0 1
*1	2.720610	1 1 1 1 1 1 1 1 1 1
1	3.816940	1 1 1 1 1 1 1 1 1 1
1	4.125836	1 1 1 1 1 1 1 1 1 1
1	3.295777	1 1 1 1 1 1 1 1 1 1
1	3.503293	1 1 1 1 1 1 1 1 1 1
*1	2.630290	1 1 1 1 1 1 1 1 1 1
1	3.235821	1 1 1 1 1 1 1 1 1 1
1	2.967702	1 1 1 1 1 1 1 1 1 1
1	2.906859	1 1 1 1 1 1 1 1 1 1
1	3.023004	1 1 1 1 1 1 1 1 1 1
*1	1.897888	1 1 1 1 1 1 1 1 1 1

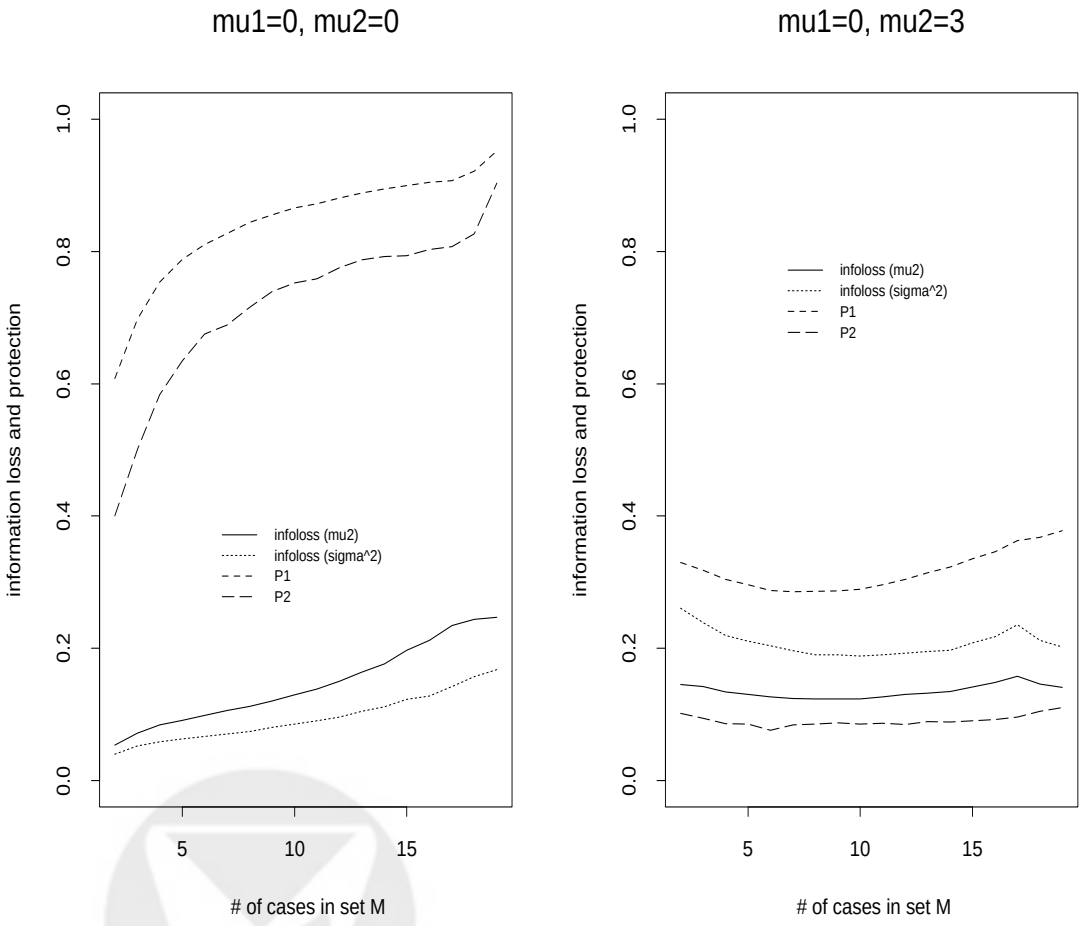
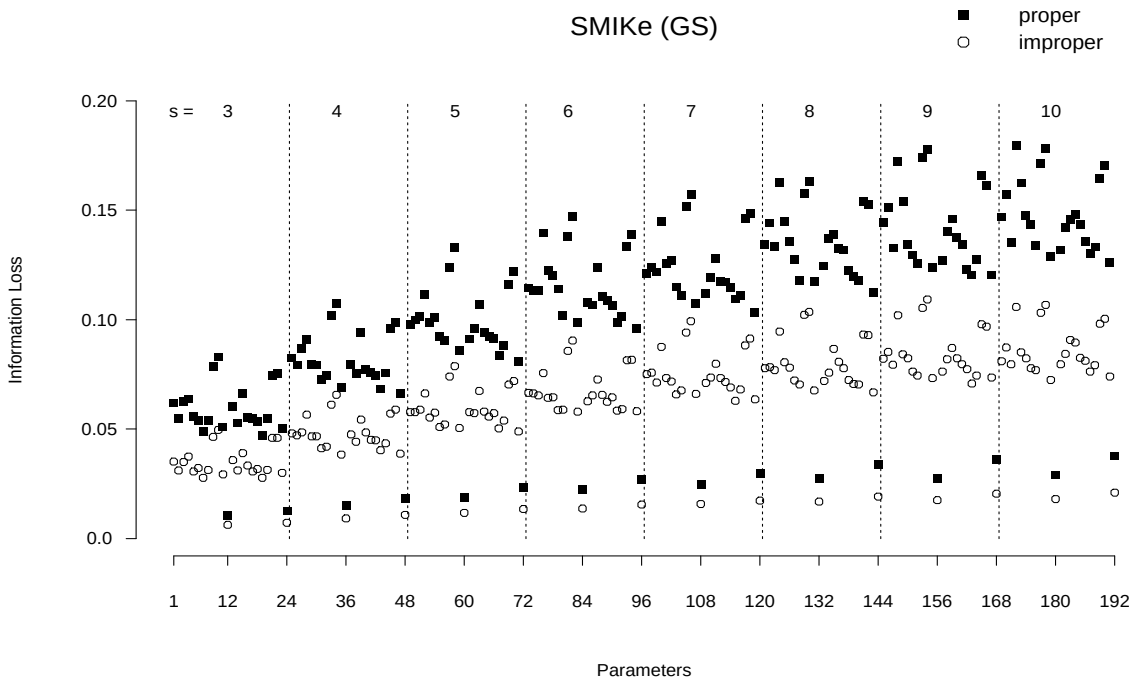


Figure 1: *Information Loss vs. Protection in SMIKe with Different* n_{mix}

Figure 2: *Information Loss of Parameters in Simulation Study II*



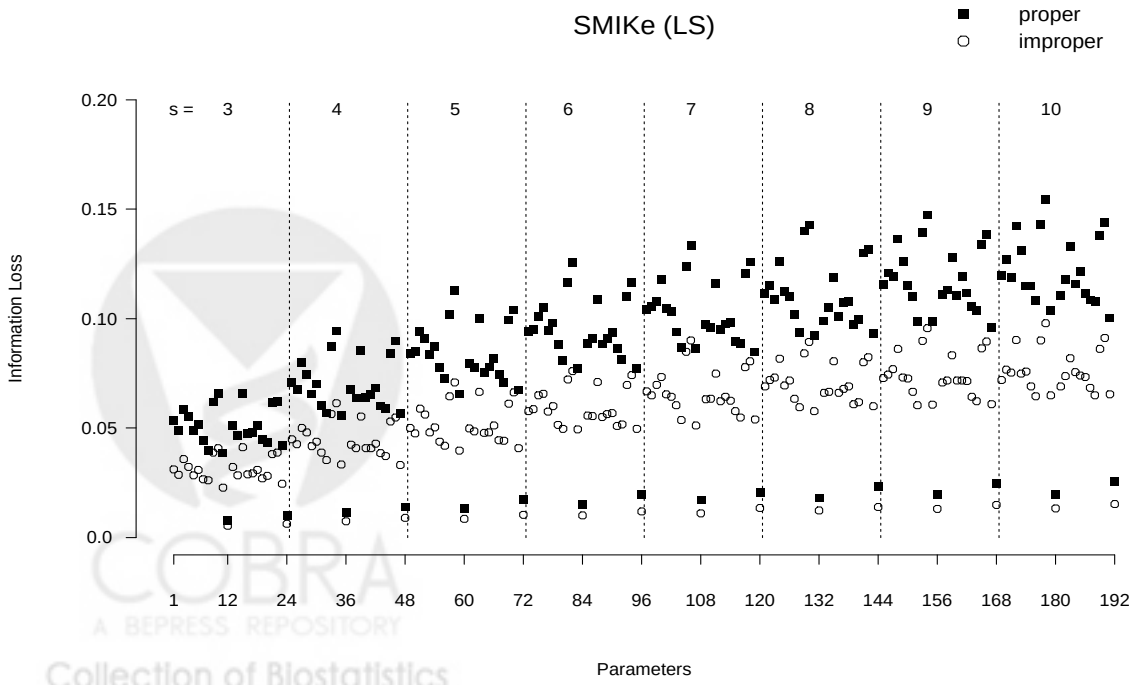


Figure 3: *Change of Confidence Interval in Simulation Study II*

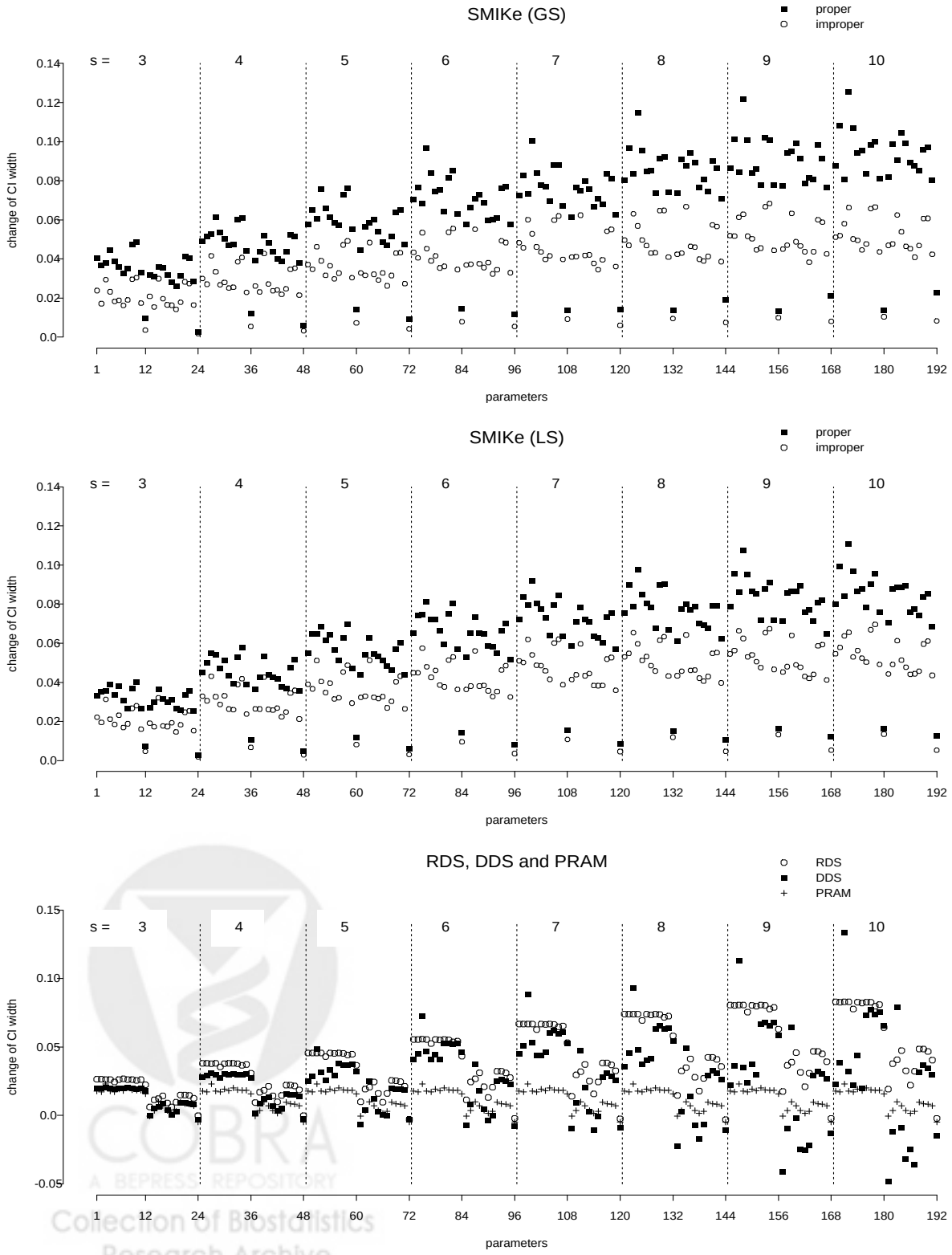


Figure 4: *Estimation Bias in Simulation Study II*

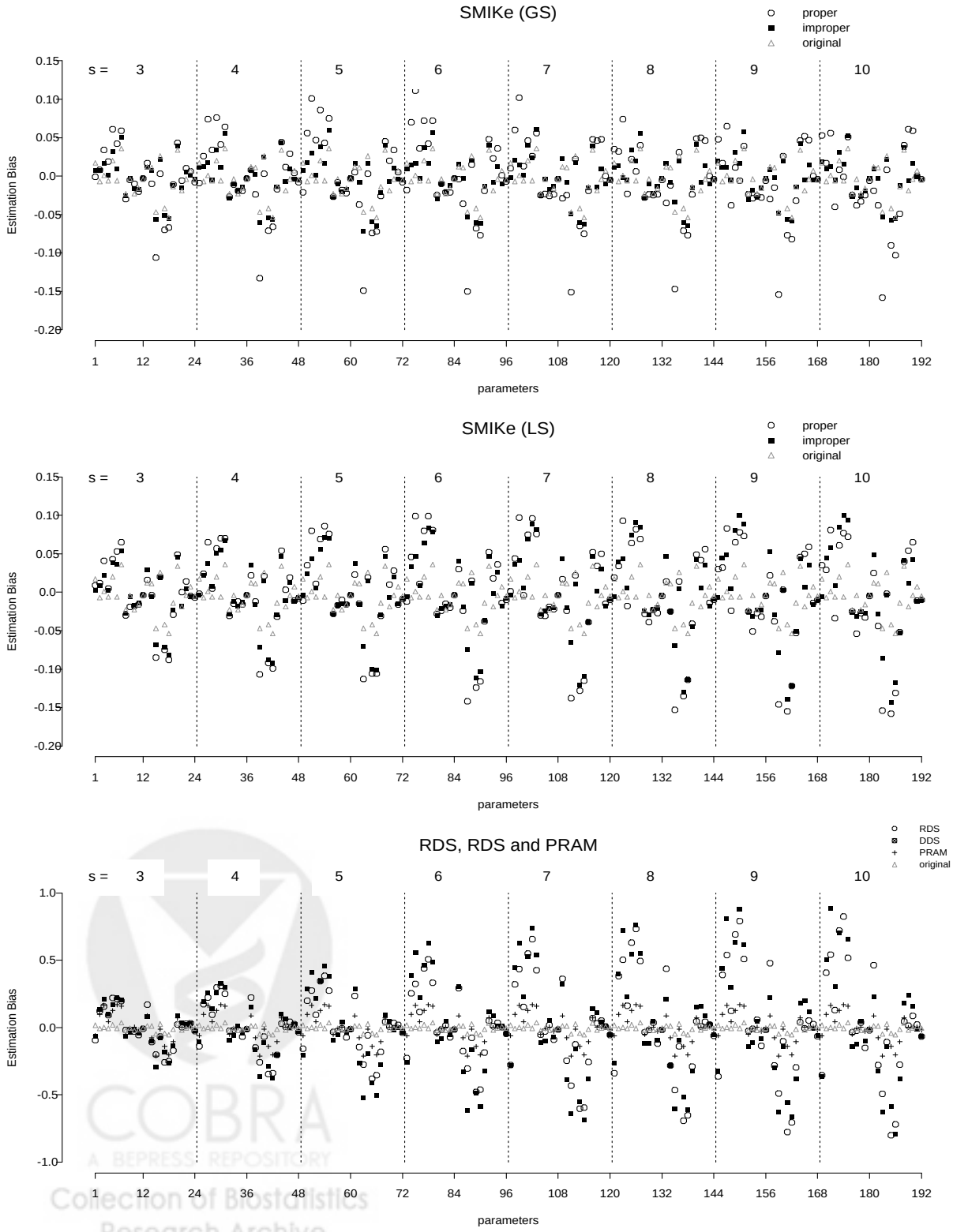


Figure 5: *Coverage Probability of nominal 95% CI in Simulation Study II*

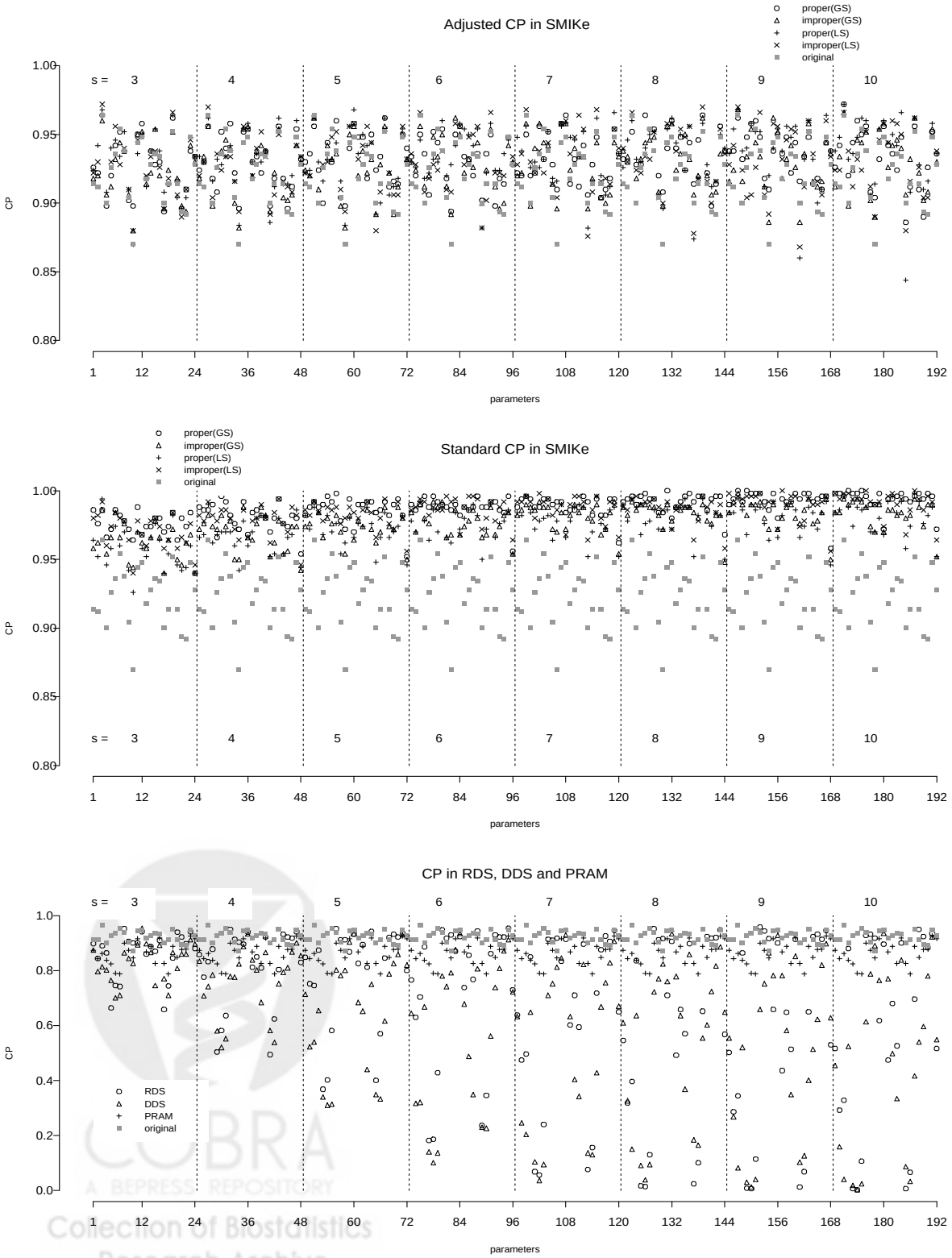


Figure 6: *Coverage Probability of Parameters from 5 Scenarios in Simulation Study III*

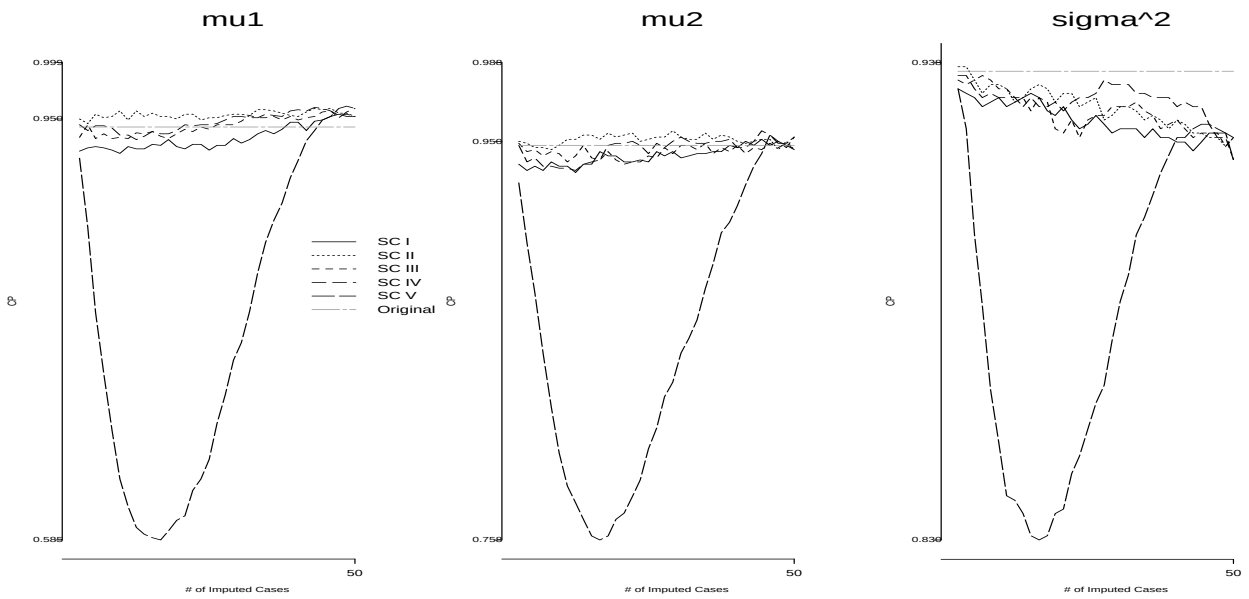


Figure 7: *Information Loss of the parameters from the logistic regression of Z_1 and Z_2 in SMIKe*