A New Approach to Protect Confidentiality for Census Microdata with Missing Values

Yajuan Si[1], Jerome P. Reiter

Department of Statistical Science, Duke University, Box 90251, Durham NC, 27708. USA.

Abstract

Many statistical agencies release samples of census microdata, such as data on individual records, to the public. Protection of the sensitive values at high risk of disclosure or values of key identifiers is a crucial problem. Meanwhile, there usually exists missing data in the census. Multiple imputation has been a useful tool to protect data confidentiality and handle missing data. We propose a new approach to census microdata dissemination: sampling simultaneously with synthesis and missing data, which generates multiple imputed datasets that simultaneously handle missing data and disclosure limitation. The basic idea is to fill in the missing data first in the census to generate multiple complete populations, then replace the identifying or sensitive values in each population with multiple imputed values, and finally release samples from these multiple imputed populations. If we release same samples across the imputed populations, the usual one-stage multiple imputation for missing data will lead to positive bias for estimation. We construct a two-stage multiple imputation to obtain the complete populations and then apply multiple imputation again to replace the sensitive information repeatedly resulting in multiply synthetic populations. We start with simple random simpling cases and then deal with stratified random sampling approach. This article develops methods to obtain valid inferences from the new three-stage imputation process based on a Bayesian prospective. New combining rules are derived due to the double duty of multiple imputation involving two additional sources of variability and the elimination of with-in variability in the census. We use simulations to check the validity of the new combining rules.

Keywords: Bayesian, Confidentiality, Missing, Synthetic, Stratified Sampling.

1. Introduction

Releasing census microdata is crucial due to the large file size and high risk of disclosure. Statistical agencies usually release a random sample from the census to solve the large file size issue. However, sampling alone is not enough for the protection of confidential data with sensitive or identifiable characteristics. In the literature, available standard disclosure limitation techniques, such as coarsening, perturbation, and swapping, are applied to protect confidentiality adequately. Moreover, nonresponse are common in census data.

Multiple imputation is a popular tool for handling missing data. This creates multiple, complete datasets that can be used for analysis or distributed to others for public files, such as the Fatality Analysis Reporting System [3], the Consumer Expenditures Survey [5], the National Health and

Email address: yajuan.si@stat.duke.edu (Yajuan Si)

Nutrition Examination Survey [10], and the National Health Interview Survey [11]. [9] and [1] provide other more examples of multiple imputation for missing data.

Nowadays it is increasingly applied to protect confidential data in public use files. Typical methodologies include full synthesis[8] replacing all values but releasing samples and partial synthesis[4] replacing part of values but releasing populations. Sampling with synthesis[2] was proposed to disseminate public use census microdata, which is motivated by the shortcomings of standard disclosure limitation techniques and expanded from full synthesis and partial synthesis. The format of released data looks exactly like the real data and the data structure and confidentiality have been preserved.

In this paper we generalize the approach of sampling with synthesis to census data with missing cases, henceforth called sampling simultaneously with synthesis and missing data. Basically we implement multiple imputation for both missing and synthesis.

We will consider same sampling approach during dissemination: selecting same records from each population. When only a fixed sample of records is released for analysis, whereas the estimation of imputation models is based on all records, the existing multiple imputation variance estimator[7] has positive bias. Hence we construct a different imputation procedure and different inferences for the same sample approach. We draw m_1 different values of the parameters of the imputation model for missing data and based on each parameter we generate m_2 multiple complete datasets. Therefore, we have $m_1 * m_2$ completed populations[7], Then for each completed population, we use multiple imputation to obtain synthetic populations, the total number of which is $m_1 * m_2 * m_3$. The whole process is referring as a three-stage multiple imputation approach.

We describe the imputation process of sampling simultaneously with synthesis and missing data in detail. Section 2 describes the three-level multiple imputation process. Section 3 provides derivations of statistical inferences and posterior distributions under the non-informative. prior. New combining rules for inference are constructed respectively in Section 4. Simulation results are shown in Section 5. We implement both simple random sampling and stratified random sampling. We also implement simulation studies to find suitable values of (m_1, m_2, m_3) in Section 5.3. Finally Section 6 conclude the whole article with some remarks and discussions.

2. Imputation Process

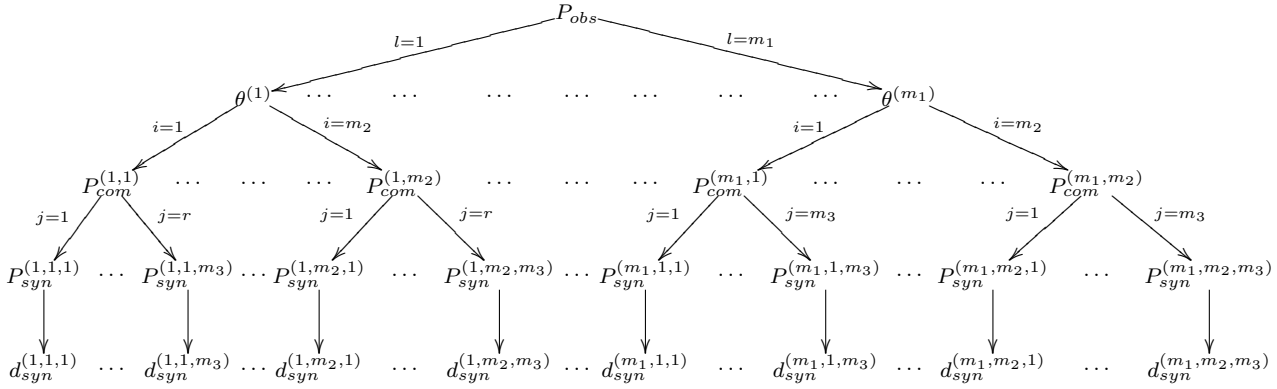
Suppose the population size of the census data is N. Let P_{obs} denote the observed data and P_{mis} denote the missing part for census data. Let P_{com} be the completed dataset after filling up missing data. P_{rep} represents the data that needs to be replaced and P_{syn} is the synthetic population after the confidential information has been replaced. d_{syn} denote the multiple samples for release.

For the same sampling approach with missing data, the released records are only part of those used to estimate the imputation model. The usual multiple imputation process for missing data has positive bias(Reiter,2008). We will also illustrate the positive bias here through simulation. We implement the two-stage imputation approach for missing data that takes all possible variances into account and adjust the whole imputation process to be specific as:

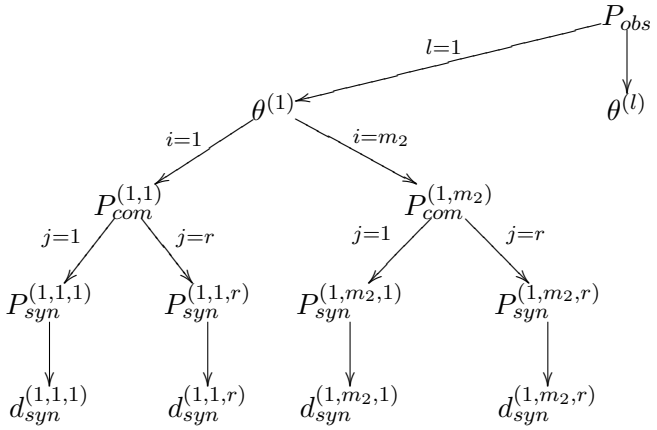
1. Estimate parameters θ for imputation model based on the posterior $(\theta|P_{obs})$, and obtain m_1 estimated values: $\theta^{(1)}, \ldots, \theta^{(m_1)}$;
2. Based on each $\theta^{(l)}$, fill in P_{mis} from the distribution $(P_{mis}|P_{obs}, \theta^{(l)})$ resulting in m_2 datasets $P_{com}^{(l,1)}, \ldots, P_{com}^{(l,m_2)}$. Totally, $m_1 m_2$ datasets are restored;

3. For each $P_{com}^{(l,i)}$, generate m_3 sets of $P_{syn}^{(l,i,j)}$ from the distribution $(P_{rep}|P_{com}^{(l,i)}, Z)$, where Z is an indicator where $Z_r^{(l,i)} = 1$ if record r in $P_{com}^{(l,i)}$ is selected to have any of its observed data replaced with synthetic values and $Z_r^{(l,i)} = 0$ for those units with unchanged data. In case of releasing any genuine or sensitive values for the selected units, the agency usually impute values for the same units in all $P_{com}^{(l)}$'s. We assume this is the case throughout and therefore drop the superscript l from Z. Total $M = m_1 m_2 m_3$ datasets are restored;
4. Release the same records $d_{syn}^{(l,i,j)}$ from $P_{syn}^{(l,i,j)}$, for $l = 1, \ldots, m_1$, $i = 1, \ldots, m_2$, and $j = 1, \ldots, m_3$.

The whole process is shown as:



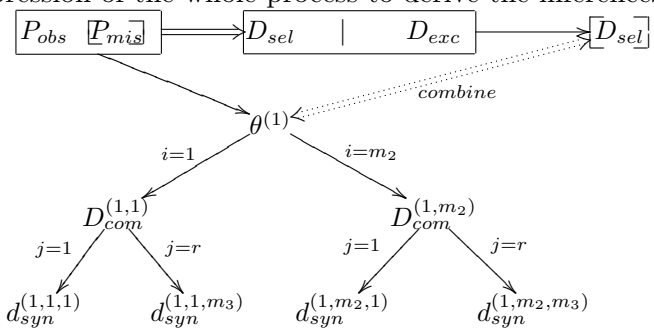
Dim out one part, it will become as following



3. Derivations of inference

Suppose Q is the quantity of interest and the analyst seeks for $f(Q|d_{syn})$. Assume the analyst's distributions are identical to those used to generate d_{syn}. We assume diffuse priors for all parameters using Bayesian arguments to derive the distributions. Suppose the sample sizes are large enough to allow normal approximations of the distributions, and then we just require the first two components for each distribution. The analyst seeks for $f(Q|d_{syn})$. However, when the same records comprise each $d_{syn}^{(l,i,j)}$, the correlations make the derivations different from those of

the different sample approach. We estimate the parameters for imputation model based on all the census data, while the estimate of Q is calculated on the sampled records. We propose an equivalent expression of the whole process to derive the inferences.



The records, or called individuals, or called units in the finally released samples are sampled from all the records in the population, including the observed part P_{obs} and the missing part P_{mis}, Even though this sampling happens in the last step when releasing, we can assume that we sample these part of records before the whole imputation process. Note that we assume knowing the records instead of their realized values beforehand. As shown in the chart above, these sampled records can be divided into two parts: D_{sel}, selected for release and D_{exc}, excluded for release. It is quite possible that records in both D_{sel} and D_{exc} have missing values. Then parameters θ of the imputation model are estimated based on all the population. Apply the imputation models and parameters to the selected records D_{sel}, and we obtain the complete dataset D_{com} with respect to the selected records D_{sel}. We perform multiple imputation on the estimation of θ and prediction of the missing values in D_{sel}. The second-level imputation is nested under the first level. Hence we will get $m_1 * m_2$ completed datasets D_{com}. Furthermore, we replace the values of confidential records repeatedly in D_{com} and get the synthetic datasets d_{syn}, which are the samples released. The parameters of the imputation model for replacement are also estimated from the population records.

The imputation process suggests

$$f(Q|d_{syn}) = \int f(Q|\theta, D_{com}, d_{syn}) f(\theta|D_{com}, d_{syn}) f(D_{com}|d_{syn}) d\theta dD_{com} \tag{1}$$

Notations defined:

- $Q^{(\theta)}$: the estimate of Q if the true parameter θ is known,for selected records, D_{sel};
- $V^{(\theta)}$: the estimate of $var(Q|Q^{(\theta)}, D_{sel})$;
- $Q^{(l)}$, for $l = 1, \ldots, m_1$: the estimate of Q if the parameter $\theta^{(l)}$ is known for D_{sel}, $\bar{Q}_{m_1} = \sum_{l=1}^{m_1} Q^{(l)}/m_1$ and denote $Q^* = \{Q^{(1)}, \ldots, Q^{(m_1)}\}$;
- $Q^{(l,i)}$, for $l = 1, \ldots, m_1$ and $i = 1, \ldots, m_2$: the estimate of Q for the selected sample after filling up missing data, called D_{com}; define $\bar{Q}_{\infty}^{(l)} = lim_{m_2 \to \infty} \sum_{i=1}^{m_2} Q^{(l,i)}/m_2$ and assume $\bar{Q}_{\infty}^{(l)} \simeq Q^{(l)}$. Denote $Q^{(**)}$={ $Q^{(l,i)}$, for $l = 1, \ldots, m_1$ and $i = 1, \ldots, m_2$ };
- $q^{(l,i,j)}$, for $j = 1, \ldots, m_3$ of each l and each i: the calculated quantity measuring Q in each released sample d_{syn}; define $\bar{q}_{\infty}^{(l,i)} = lim_{m_3 \to \infty} \frac{\sum_{j=1}^{m_3} q^{(l,i,j)}}{m_3}$ and assume $\bar{q}_{\infty}^{(l,i)} \simeq Q^{(l,i)}$;

- $B_\infty = lim_{m_1\to\infty}\sum_{l=1}^{m_1}(Q^{(l)}-\bar{Q}_{m_1})^2/(m_1-1)$
- $W1_\infty^{(l)} = lim_{m_2\to\infty}\frac{\sum_{i=1}^{m_2}(Q^{(l,i)}-Q^{(l)})^2}{m_2-1}$, $W1 = \sum_{l=1}^{m_1} W1_\infty^{(l)}/m_1$ and $W1^* = \{W1_\infty^{(1)},\ldots,W1_\infty^{(m_1)}\}$;
- $W2_\infty^{(l,i)} = lim_{m_3\to\infty}\frac{\sum_{j=1}^{m_3}(q^{(l,i,j)}-Q^{(l,i)})^2}{m_3-1}$, $W2 = \sum_{l=1}^{m_1}\sum_{i=1}^{m_2} W2_\infty^{(l,i)}/m_1m_2$ and $W2^{**} = \{W2_\infty^{(1,1)},\ldots,W2_\infty^{(m_1,m_2)}$

The derivation proceeds by using $Q^{(\theta)}$ to estimate Q, $\bar{Q}_{m_1}$ to estimate $Q^{(\theta)}$, $\bar{Q}_{m_1m_2} = \sum_{l,i} Q^{(l,i)}/m_1m_2 = \sum_{l=1}^{m_1}\bar{Q}^{(l)}/m_1$ to estimate $\bar{Q}_{m_1}$, and $\bar{q}_M = \sum_{l,i,j} q^{(l,i,j)}/(m_1m_2m_3)$. Specifically,

$$f(Q|d_{syn}) = \int f(Q|d_{syn},Q^{(\theta)},V^{(\theta)},Q^*,Q^{**},B_\infty,W1^*,W2^{**}) \tag{2}$$

$$f(Q^{(\theta)}|d_{syn},V^{(\theta)},Q^*,Q^{**},B_\infty,W1^*,W2^{**}) \tag{3}$$

$$f(\bar{Q}_{m_1}|d_{syn},W1^*,W2^{**})f(\bar{Q}^{(l)}|d_{syn},W2^{**}) \tag{4}$$

$$f(V^{(\theta)},B_\infty,W1^*,W2^{**}|d_{syn})dB_\infty dW1^* dW2^{**} \tag{5}$$

1. $Q^{(\theta)}\to Q$
   All quantities associated with imputations are irrelevant for inference about Q given $Q^{(\theta)}$ and $V^{(\theta)}$. Assume
   $$Bayes: (Q|D_{sel},Q^{(\theta)},V^{(\theta)}) \sim N(Q^{(\theta)},V^{(\theta)}) \tag{6}$$
   If the released sample size n is not small enough with respect the the population size N, we need to add the finite population correction factor $(1-n/N)$. Then we will have
   $$Bayes: (Q|D_{sel},Q^{(\theta)},V^{(\theta)}) \sim N(Q^{(\theta)},(1-n/N)*V^{(\theta)}) \tag{7}$$
   If $n \ll N$, we can ignore the finite population correction factor. To be clear for the whole derivation, we will ignore it first and include it later.
2. $Q^{(l)}\to Q^{(\theta)}$
   The selected values of D_{sel} and $W1^*$ are irrelevant for inference about $Q^{(\theta)}$ given Q^* and B_∞. Assume
   $$Bayes: (Q^{(\theta)}|D_{sel},B_\infty,Q^*) \sim N(\bar{Q}_{m_1},(1+1/m_1)B_\infty) \tag{8}$$
3. $Q^{(l,i)}\to Q^{(l)}$
   Assume
   $$Freq: (Q^{(l,i)}|D_{sel},Q^{(l)},W1_\infty^{(l)}) \sim N(Q^{(l)},W1_\infty^{(l)}) \tag{9}$$
   then the posterior distribution is
   $$Bayes: (Q^{(l)}|D_{com},W1_\infty^{(l)}) \sim N(\bar{Q}^{(l)},W1_\infty^{(l)}/m_2) \tag{10}$$
   where $\bar{Q}^{(l)} = \sum_{i=1}^{m_2} Q^{(l,i)}/m_2$. Consider $Q^{(l)}$'s are independent,
   $$Bayes: (\bar{Q}_{m_1}|D_{com},W1^*) \sim N(\bar{Q}_{m_1m_2},W1/m_1m_2) \tag{11}$$
4. $q^{(l,i,j)}\to Q^{(l,i)}$
   The inference of $Q^{(l,i)}$ is only related to $q^{(l,i,j)}$, for $j=1,\ldots,m_3$ of each l and each i and

$W2_\infty^{(l,i)}$. Assume the sampling distribution

$$Freq : (q^{(l,i,j)}|D_{com}, Q^{(l,i)}, W2_\infty^{(l,i)}) \sim N(Q^{(l,i)}, W2_\infty^{(l,i)}) \tag{12}$$

so that

$$Bayes : (Q^{(l,i)}|d_{syn}, W2_\infty^{(l,i)}) \sim N(\frac{1}{m_3}\sum_{j=1}^{m_3} q^{(l,i,j)}, W2_\infty^{(l,i)}/m_3) \tag{13}$$

Based on the independence of $Q^{(l,i)}$'s,

$$Bayes : (\bar{Q}^{(l)} = \frac{1}{n}\sum_{i=1}^{n} Q^{(l,i)}|d_{syn}, W2_\infty^{(l,i)}, i = 1, \ldots, m_2) \sim N(\frac{1}{m_2 m_3}\sum_{i=1}^{m_2}\sum_{j=1}^{m_3} q^{(l,i,j)}, \frac{1}{m_2 m_3}\bar{W2}_\infty^{(l)}) \tag{14}$$

where $\bar{W2}_\infty^{(l)} = \sum_{i=1}^{m_2} W2_\infty^{(l,i)}/m_2$.

5. $q^{(l,i,j)} \to Q^{(l)}$

$$(\bar{Q}_{m_1}|d_{syn}, W1^*, W2^{(**)}) \sim N(\bar{q}_M, \frac{W1}{m_1 m_2} + \frac{W2}{m_1 m_2 m_3}) \tag{15}$$

6. $f(V^{(\theta)}|d_{syn}, B_\infty, W1^*, W2^{(**)})$
First define $V^{(l,i,j)} = var(Q|d_{syn}^{(l,i,j)}, Q^{(\theta)} = Q^{(l)}, W1_\infty^{(l)}, W2_\infty^{(l,i)})$, and use $u^{(l,i,j)}$ to estimate $var(Q|d_{syn}^{(l,i,j)}, W1_\infty^{(l)}, W2_\infty^{(l,i)})$. Based on an iterated variance computation to relate these quantities,

$$u^{(l,i,j)} = E(V^{(l,i,j)}|d_{syn}^{(l,i,j)}, W1_\infty^{(l)}, W2_\infty^{(l,i)}) + W1_\infty^{(l)} + W2_\infty^{(l,i)} \tag{16}$$

Rewrite this as an expression of $V^{(l,i,j)}$, we have

$$E(V^{(l,i,j)}|d_{syn}^{(l,i,j)}, W1_\infty^{(l)}, W2_\infty^{(l,i)}) = u^{(l,i,j)} - W1_\infty^{(l)} - W2_\infty^{(l,i)} \tag{17}$$

Assume the sampling distribution of $V^{(l,i,j)}$ has mean $V^{(\theta)}$, so that

$$E(V^{(\theta)}|d_{syn}, B_\infty, W1^*, W2^{**}) = E\{E(v^{(\theta)}|d_{syn}, B_\infty, W1^*, W2^{**}, Q^{**}, Q^*)|d_{syn}, B_\infty, W1^*, W2^{**}\} \tag{18}$$

$$= E\{\sum_{l,i,j} V^{(l,i,j)}/(m_1 m_2 m_3)|d_{syn}, B_\infty, W1^*, W2^{**}\} \tag{19}$$

$$= \bar{u}_M - W1 - W2 \tag{20}$$

Assume the sampling variance for $V^{(l,i,j)}$ is negligible, which also implies negligible variance of $\bar{u}_M$. Then $f(V^{(\theta)}|d_{syn}, B_\infty, W1^*, W2^{(**)})$ can be treated as a distribution concentrated at $\bar{u}_M - W1 - W2$ with negligible variance.

7. B_∞, $W1$ and $W2$
The analysis-of-variance analysis gives the posterior distribution of $(W2|d_{syn})$, $(W1|d_{syn}, W2)$ and $(B_\infty|d_{syn}, W1, W2)$,

$$\frac{m_1 m_2(m_3 - 1)\bar{w2}_M}{W2}|d_{syn} \sim \chi^2_{m_1 m_2(m_3-1)} \tag{21}$$

where $\bar{w}2_M = \frac{1}{m_1 m_2(m_3-1)} \sum_{l=1}^{m_1} \sum_{i=1}^{m_2} \sum_{j=1}^{m_3} (q^{(l,i,j)} - \bar{q}_{m_3}^{(l,i)})^2$ and $\bar{q}^{(l,i)} = \frac{\sum_{j=1}^{m_3} q^{(l,i,j)}}{m_3}$;

$$\frac{m_1(m_2-1)\bar{w}1_M}{W1 + W2/m_3} | d_{syn}, W2 \sim \chi^2_{m_1(m_2-1)} \tag{22}$$

where $\bar{w}1_M = \frac{1}{m_1(m_2-1)} \sum_{l=1}^{m_1} \sum_{i=1}^{m_2} (\bar{q}^{(l,i)} - \bar{q}_{m_2 r}^{(l)})^2$ and $\bar{q}_{m_2 r}^{(l)} = \frac{\sum_{i=1}^{m_2} \bar{q}^{(l,i)}}{m_2}$;

$$\frac{(m_1-1)b_M}{B_\infty + W1/m_2 + W2/m_2 m_3} | d_{syn}, W1, W2 \sim \chi^2_{m_1-1} \tag{23}$$

where $b_M = \frac{1}{m_1-1} \sum_{l=1}^{m_1} (\bar{q}_{m_2 m_3}^{(l)} - \bar{q}_M)^2$.

8. T_s

For large m_1, m_2 and m_3, $f(Q|d_{syn})$ can be approximated by a normal distribution with mean $E(Q|d_{syn}) = \bar{q}_M$ and the variance

$$\begin{aligned} var(Q|d_{syn}) &= E\{var(Q|Q^{(\theta)})|d_{syn}\} + var\{E(Q|Q^{(\theta)})|d_{syn}\} \\ &= E\{E(V^{(\theta)})|d_{syn}\} + E\{var(Q^{(\theta)}|Q^*)|d_{syn}\} + var\{E(Q^{(\theta)}|Q^*)|d_{syn}\} \\ &= \bar{u}_M - E(W1|d_{syn}) - E(W2|d_{syn}) + E\{(1+1/m)B_\infty|d_{syn}\} \\ &+ E\{var(Q^*|Q^{**})\} + var\{E(Q^*|Q^{**})\} \\ &= \bar{u}_M - E(W1|d_{syn}) - E(W2|d_{syn}) + E\{(1+1/m_1)B_\infty|d_{syn}\} \\ &+ E(W1|d_{syn})/m_1 m_2 + E(W2|d_{syn})/m_1 m_2 m_3 \end{aligned} \tag{24}$$

We approximate the expectations by $E(W2|d_{syn}) \simeq \bar{w}2_M$, $E(W1|d_{syn}) \simeq \bar{w}1_M - \bar{w}2_M/m_3$, and $E(B_\infty|d_{syn}) \simeq b_M - \bar{w}1_M/m_2$, so

$$T_s = \bar{u}_M - (1+1/m_2)\bar{w}1_M - (1-1/r)\bar{w}2_M + (1+1/m_1)b_M \tag{25}$$

It is also possible that T_s is negative. We modify it as $T_s^* = \bar{u}_M + (1+1/m_1)b_M$ when $T_s < 0$. For the special cases, if we only has missing data without synthesis, then $W2_M = 0$, then $T = \bar{u}_M - (1+1/m_2)\bar{w}1_M + (1+1/m_1)b_M$, which is the same as the variance estimator of two-stage approach for missing data [7]; If there is no missing data and we only sample with synthesis, then $B_\infty = 0$ and $W1_M = 0$, so $T = \bar{u}_M - \bar{w}2_M + \frac{1}{r}\bar{w}2_M$, which is the same as the variance estimator under sampling with synthesis [2]. If including the finite population correction factor, then

$$T_s = (1-n/N)(\bar{u}_M - \bar{w}1_M - \bar{w}2_M) + (1+1/m_1)b_M + \bar{w}1_M/m_1 m_2 + \bar{w}2_M/m_1 m_2 m_3 \tag{26}$$

4. Combining rules for inference

We summarize the combining rules for inferences of the same sampling approaches.

4.1. Simple random sampling

For large m_1, m_2 and m_3, we approximate the posterior distribution of $(Q|d_{syn})$ as a normal distribution,

$$(Q|d_{syn}) \sim N(\bar{q}_M, T_s) \tag{27}$$

When m_1, m_2 and m_3 are modest, we use a t-distribution,

$$(Q|d_{syn}) \sim t_{v_s}(\bar{q}_M, T_s) \tag{28}$$

Quantities needed for the combining rules:

$$\begin{aligned}
\bar{q}_M &= \frac{1}{m_1 m_2 m_3}\sum_{l,i,j} q^{(l,i,j)} = \frac{1}{m_1 m_2}\sum_{l,i}\bar{q}^{(l,i)}_{m_3} = \frac{1}{m_1}\sum_{l}\bar{q}^{(l)}_{m_2 m_3} && (29)\\
\bar{w}2_M &= \frac{1}{m_1 m_2(m_3-1)}\sum_{l=1}^{m_1}\sum_{i=1}^{m_2}\sum_{j=1}^{m_3}(q^{(l,i,j)} - \bar{q}^{(l,i)}_{m_3})^2 && (30)\\
\bar{w}1_M &= \frac{1}{m_1(m_2-1)}\sum_{l=1}^{m_1}\sum_{i=1}^{m_2}(\bar{q}^{(l,i)}_{m_3} - \bar{q}^{(l)}_{m_2 m_3})^2 && (31)\\
b_M &= \frac{1}{m_1-1}\sum_{l=1}^{m_1}(\bar{q}^{(l)}_{m_2 m_3} - \bar{q}_M)^2 && (32)\\
\bar{u}_M &= \sum_{l,i,j} u^{(l,i,j)}/(m_1 m_2 m_3). && (33)
\end{aligned}$$

$$T_s = \bar{u}_M - (1+1/m_2)\bar{w}1_M - (1-1/r)\bar{w}2_M + (1+1/m_1)b_M \tag{34}$$

The number of degrees of freedom v_s is derived by matching the first two moments of

$$\frac{v_s T_s}{\bar{u}_M - W1 - W2 + (1+1/m_1)B_\infty + W1/m_1 m_2 + W2/m_1 m_2 m_3} \sim \chi^2_{v_s} \tag{35}$$

where $T_s = \bar{u}_M + (1+1/m_1)b_M - (1+1/m_2)\bar{w}1_M - (1-1/m_3)\bar{w}2_M$.

We determine v_s by matching the mean and variance of the chi-squared distribution to those of (35). Set $\alpha = \frac{B_\infty + W1/m_2 + W2/m_2 m_3}{b_M}$, $\beta = \frac{W1+W2/m_3}{\bar{w}1_M}$ and $\gamma = W2/\bar{w}2_M$, then

$$(m_1-1)\alpha^{-1}|d_{syn}, W1, W2 \sim \chi^2_{m_1-1} \tag{36}$$

$$m_1(m_2-1)\beta^{-1}|d_{syn}, W2 \sim \chi^2_{m_1(m_2-1)} \tag{37}$$

$$m_1 m_2(m_3-1)\gamma^{-1}|d_{syn} \sim \chi^2_{m_1 m_2(m_3-1)} \tag{38}$$

Let $f = (1+\frac{1}{m_1})b_M/\bar{u}_M$, $g = (1+\frac{1}{m_2})\bar{w}1_M/\bar{u}_M$, and $e = (1-\frac{1}{m_3})\bar{w}2_M/\bar{u}_M$. Write (35) as

$$\frac{T_s}{\bar{u}_M + (1+\frac{1}{m_1})(B_\infty + W1/m_2 + W2/m_2 m_3) - (1+\frac{1}{m_2})(W1 + W2/m_3) - (1-\frac{1}{m_3})W2} = \frac{1+f-g-e}{1+\alpha f - \beta g - \gamma e} \tag{39}$$

1. For the expectation of 39, use an iterated expectation and first-order Taylor series expansions in α^{-1}, β^{-1} and γ^{-1} around their expectations, which equal to one, and obtain

$$E\{E(\frac{1+f-g-e}{1+\alpha f - \beta g - \gamma e}|d_{syn}, W1, W2)|d_{syn}\} \simeq E(E(\frac{1+f-g-e}{1+f-\beta g - \gamma e}|d_{syn}, W2)|d_{syn}) \simeq 1 \tag{40}$$

2. For the variance of 39, use the iterated variance computation,

$$E\{var(\frac{1+f-g-e}{1+\alpha f-\beta g-\gamma e}|d_{syn},W1,W2)|d_{syn}\}+var\{E(\frac{1+f-g-e}{1+\alpha f-\beta g-\gamma e}|d_{syn},W1,W2)|d_{syn}\} \tag{41}$$

For the interior variance and expectation, use a first-order Taylor series expansion in α around 1. Since $var(\alpha^{-1}|d_{syn},W1,W2)=2/(m_1-1)$, the above expression equals approximately

$$E\{\frac{2(1+f-g-e)^2f^2}{(m_1-1)(1+f-\beta g-\gamma e)^4}|d_{syn}\}+var\{\frac{1+f-g-e}{1+f-\beta g-\gamma e}|d_{syn}\} \tag{42}$$

The first part of 42 is

$$E\{E(\frac{2(1+f-g-e)^2f^2}{(m_1-1)(1+f-\beta g-\gamma e)^4}|d_{syn},W2)|d_{syn}\}\simeq\frac{2f^2}{(m_1-1)(1+f-g-e)^2} \tag{43}$$

Use the iterated variance computation for the second part of (42),

$$E\{var(\frac{1+f-g-e}{1+f-\beta g-\gamma e}|d_{syn},W2)|d_{syn}\}+var\{E(\frac{1+f-g-e}{1+f-\beta g-\gamma e}|d_{syn},W2)|d_{syn}\} \tag{44}$$

Use a first-order Taylor series expansion in β^{-1} around 1 with $var(\beta^{-1}|d_{syn},W2)=\frac{2}{m_1(m_2-1)}$. and use a first-order Taylor series expansion in γ^{-1} around 1 with $var(\gamma^{-1}|d_{syn})=\frac{2}{m_1m_2(rm_3-1)}$. The second part of (42) becomes

$$\frac{2g^2}{m_1(m_2-1)(1+f-g-e)^2}+\frac{2e^2}{m_1m_2(m_3-1)(1+f-g-e)^2} \tag{45}$$

We get

$$v_s=\frac{f^2}{(m_1-1)(1+f-g-e)^2}+\frac{g^2}{m_1(m_2-1)(1+f-g-e)^2}+\frac{e^2}{m_1m_2(m_3-1)(1+f-g-e)^2} \tag{46}$$

that is

$$v_s=\{\frac{[(1+\frac{1}{m_1})b_M]^2}{(m_1-1)T_s^2}+\frac{[(1+1/m_2)\bar{w}1_M]^2}{m_1(m_2-1)T_s^2}+\frac{[(1-1/m_3)\bar{w}2_M]^2}{m_1m_2(m_3-1)T_s^2}\}^{-1} \tag{47}$$

4.2. Stratified random sampling

In practice, comparing to simple random sampling, stratified random sampling approach is more common and attractive for taking stratifications and differences accross multiple domains into account. We modify the combining rules to make valid inferences based on stratified sampling estimation. Suppose N_h is the population size and n_h is the sample size in Stratum h, for $h=1,\ldots,H$. If Q is the quantity of interest in the census, let $\bar{q}_{Mh}$ be the estimate of Q in Stratum h and let T_{sh} be the value of T_s computed based on the records in Stratum h. If n_h is not too small comparing to N_h, we have to include the finite population correction factor into the computation of T_{sh}. Then we can combine $\bar{q}_{Mh}$ and T_{sh} across stratum to estimate Q in the census. In the combining rules, the point estimate is $\bar{q}_M=\Sigma_h(N_h/N)\bar{q}_{Mh}$ and $T_s=\Sigma_h(N_h/N)^2T_{sh}$. Inferences can be similarly made on $(Q-\bar{q}_M)\sim N(0,T_s)$. When m_1, m_2 and m_3 are modest, we propose

$(Q - \bar{q}_M) \sim t_{\nu_{st}}(0, T_s)$, where ν_{st} is derived similarly as ν_s by matching the mean and variance of the chi-squared distribution $B_{\infty h}$, $W1_h$,and $W2_h$, which are referring as B_∞, $W1$ and $W2$ in Stratum h. Let $\bar{u}_{Mh}$, b_{Mh}, $\bar{w}1_{Mh}$ and $\bar{w}2_{Mh}$ be the corresponding values of $\bar{u}_M$, b_M, $\bar{w}1_M$ and $\bar{w}2_M$ in Stratum h.

We have

$$v_{st} = \{\frac{[(1+\frac{1}{m_1})\Sigma_{h=1}^{H}(N_h/N)^2 b_M]^2}{(m_1-1)T_s^2} + \frac{[(1+1/m_2)\Sigma_{h=1}^{H}(N_h/N)^2 \bar{w}1_M]^2}{m_1(m_2-1)T_s^2} + \frac{[(1-1/m_3)\Sigma_{h=1}^{H}(N_h/N)^2 \bar{w}2_M]^2}{m_1 m_2(m_3-1)T_s^2}\}^{-1} \quad (48)$$

For most cases v_{st} is large enough that an approximate normal distribution is adequate for inferences.

5. Simulation

We implement simulations to show the derived combining rules for inference.

5.1. Simple random sampling

We generate a population of $N = 10^5$ units comprising five variables $Y_1, \ldots, Y_5$ drawn from $N(0, \Sigma)$, where the element $\Sigma_{ij} = 5{*}0.8^{|i-j|}$. Suppose Y_5 has missing values. The missing percentage is p, where we consider $p = 15\%$. We construct regression models on the observed data and predict the missing values. For the same sample approach, we draw m_1 different values of the parameters and based on each parameter we generate m_2 multiple missing datasets. Therefore, we have $m_1 * m_2$ completed populations, corresponding to three-stage approach. We replace values of Y_4 for all units with $Y_1 > q$, where q is the 75th percentile of Y_1 in the complete population. The replacement values are generated from the predictive distribution $f(Y_4|P_{com}, Y_1 > q)$, with parameter values(here are the values for the regression coefficients) calculated from the population. We draw samples with the same records of size $n = 10000$ from the total $M = m_1 * m_2 * m_3$ synthetic populations. We run simulations under different scenarios of the values of (m_1, m_2, m_3).

Based on the generated samples, we implement the derived combining rules for twelve quantities in each simulation iteration, including the population means of Y_4 and Y_5, the coefficients from a regression of Y_3 on all other variables and the coefficients from a regression of Y_5 on all other variables.

We repeat the whole three-level imputation process 500 times and compare the values of the derived variance formula with their true values. Specifically, we calculate the ratio of the derived variance T_s to the true variance $var(\bar{q}_M)$ to check the validity. The closer the ratio is to 1, the more evidence supports T_s. Furthermore, we consider the nominal coverage rate and average length of the 95% confidence intervals of the twelve quantities of interest. No negative values of T_s happen here.

From the results in Figure 1, we can see that for the differently chosen values of (m_1, m_2, m_3), all the ratios are quite close to 1. Each T_s is approximately unbiased for $var(\bar{q}_M)$. We apply the student t distribution of the quantity and calculate the corresponding degrees of freedom. The nominal coverage rates of 95% confidence intervals are approximately 95%. Figure 1 illustrates that the derived combining rules provide a valid and approximately true variance estimation.

5.2. Stratified random sampling

We simulate a large census dataset comparing five variables $Y_1, \ldots, Y_5$ in four different strata.Y_1 is a binary variable with value 0 or 1 and the probability $P(Y_1 = 1) = 0.7$ keeps across the four strata. The generating distributions for $(Y_2, \ldots, Y_5)$ and the stratum sizes are shown in Table.

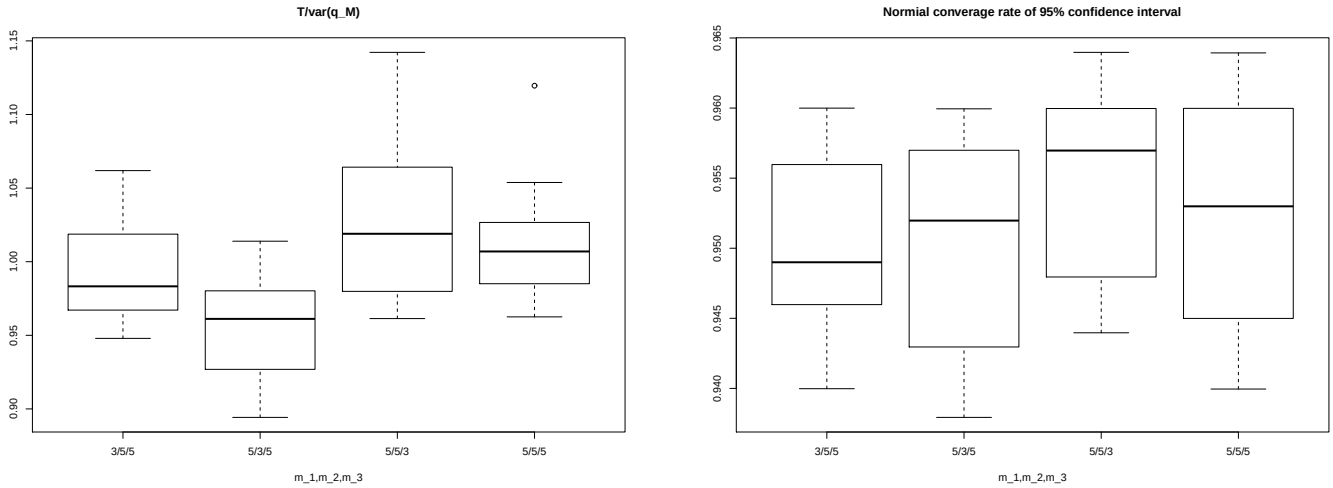


Figure 1: Boxplots of the calculated ratios of $T_s/var(q_M)$ and the nominal coverage rate of 95% confidence intervals under different value settings of (m_1, m_2, m_3)

The statistical agencies can apply proportional allocation and choose the same number to set the sample sizes for release from each stratum. First we use a common sample size $n = 10000$, where the finite population correction factors are non-ignorable, especially in Stratum 3 and Stratum 4.

We report the nominal coverage rate and average length of 95% confidence intervals of the mean of Y_5 and Y_4. The output is shown in Table 2. The ratio $Avg(T_s)/var(\bar{q}_M)$ (1.007) for $\bar{Y}_4$ is approximately as 1. The ratio $Avg(T_s)/var(\bar{q}_M)$ (1.007) for $\bar{Y}_5$ has 13% positive bias. The nominal coverage rates of 95% confidence intervals are both close to 95%. We can see that the derived combining rules adjusted by stratified sampling perform well here.

5.3. Exploration for suitable values of (m_1, m_2, m_3)

To choose a set of suitable values of (m_1, m_2, m_3) is a crucial problem [6] when it comes to practice, which is related to balancing the efficiency of combining rules and the computation burden. Too large values of (m_1, m_2, m_3) will cause heavy computation burdens for the statistical agencies, while too small values of (m_1, m_2, m_3) will reduce the efficiency of the whole imputation process. We implement several simulation studies to explore the importance roles or affects of different value choices of (m_1, m_2, m_3). Here we set the population size as $N = 10000$ and repeat the whole simulation for $T = 5000$ times. The data generating systhem is the same as that in the simple random sampling of Section 5.1. Because the population size is smaller, some values of the estimated variances become negative. The negative rater will be reported here. We first set the missing proportion of Y_5 as 50% and q as the 90% quantile of Y_1, where the missing percentage is quite large while the replacement percentage is small. Missingness dominates the most variability. The second senario is that 10% of the values of Y_1 are missing and q is 50% quantile of Y_1. Sampling with synthesis plays a more important role for the second case. We choose different sets of values of (m_1, m_2, m_3) as repectively $(10, 3, 3)$, $(3, 3, 10)$ and $(3, 10, 3)$. This setting will help find the values of which of m_1, m_2 and m_3 will affect the most to the estimation of variance.

Figure 2 displays the simulation results for the two cases. The first column is referring the fist case where the missing rate is 50% and the replacement rate is 10%; the second column is

Table 1: Distributions for simulating $(Y_2, \ldots, Y_5)$ across strata

	Stratum size-N_h	Simulating models CI length
Stratum 1	80000	$logY_2 \sim N(-1 + 0.5Y_1, 1)$
		$Y_3 \sim N(Y_1 + logY_2, 1)$
		$Y_4 \sim N(Y_1 + 1.5logY_2 + Y_3, 1)$
		$Y_5 \sim N(-2.5logY_2 + Y_3 + 0.1Y_4, 1)$
Stratum 2	50000	$logY_2 \sim N(-2 + Y_1, 3)$
		$Y_3 \sim N(2Y_1 + 2logY_2, 3)$
		$Y_4 \sim N(2Y_1 + 3logY_2 + 2Y_3, 3)$
		$Y_5 \sim N(-5logY_2 + Y_3 + 0.2Y_4, 3)$
Stratum 3	30000	$logY_2 \sim N(-3 + 1.5Y_1, 5)$
		$Y_3 \sim N(3Y_1 + 3logY_2, 5)$
		$Y_4 \sim N(3Y_1 + 4.5logY_2 + 3Y_3, 5)$
		$Y_5 \sim N(-7.5logY_2 + Y_3 + 0.3Y_4, 5)$
Stratum 4	10000	$logY_2 \sim N(-4 + 2Y_1, 7)$
		$Y_3 \sim N(4Y_1 + 4logY_2, 7)$
		$Y_4 \sim N(4Y_1 + 6logY_2 + 4Y_3, 7)$
		$Y_5 \sim N(-10logY_2 + Y_3 + 0.4Y_4, 7)$

Table 2: Outputs under Stratified random sampling with $(m_1, m_2, m_3) = (5, 5, 5)$

Q	$Avg(T_s)/var(\bar{q}_M)$	95% CI nominal coverage(%)	95% CI average length
$\bar{Y}_5$	1.135	95.2	0.135
$\bar{Y}_4$	1.007	96.0	0.453

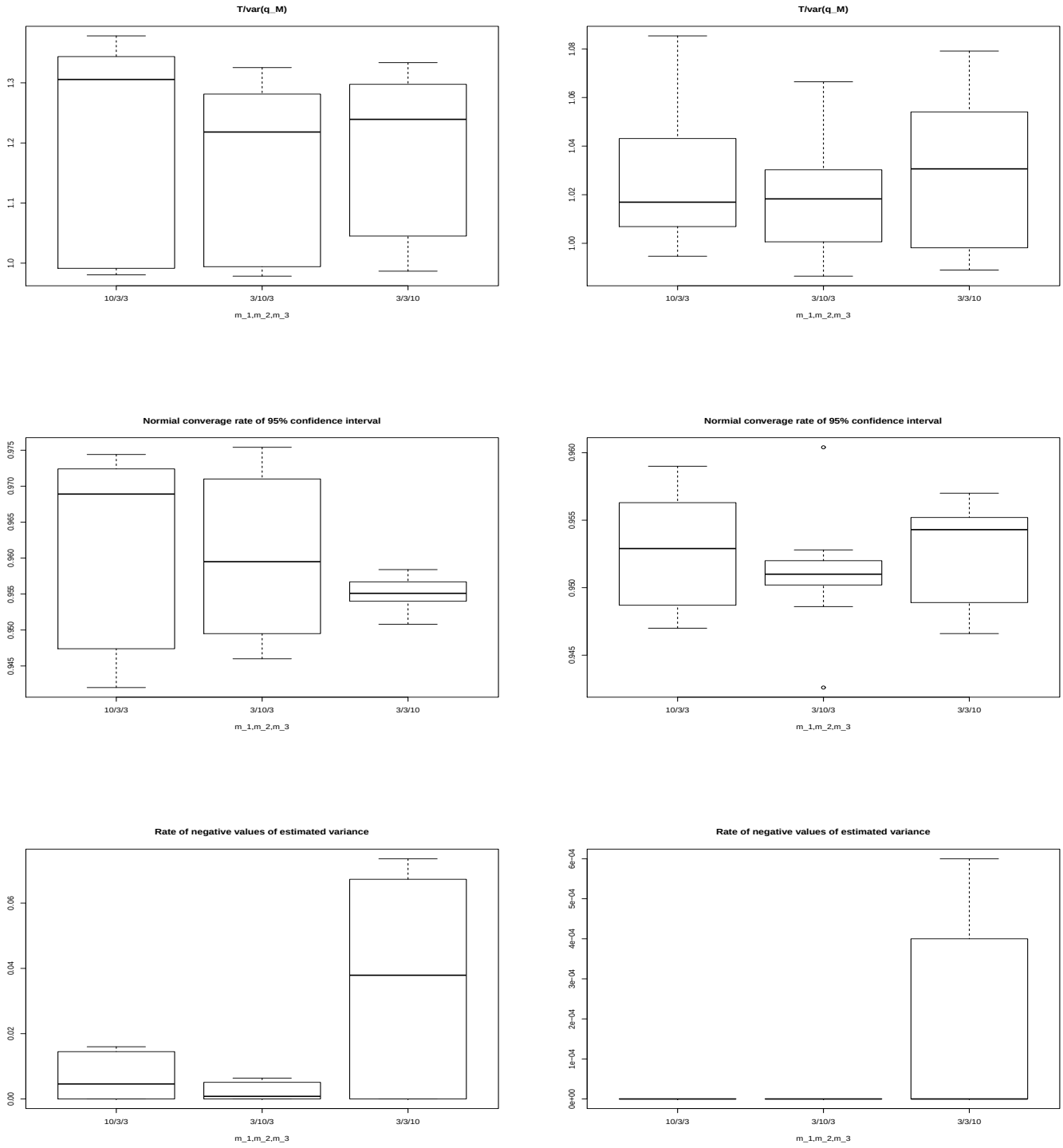


Figure 2: Simulation results for different missing percentages and different replacement rates: the left column represents the case when 50% values are missing and the threshold percentile of Y_1 is 90%; the right column represents the case when 10% values are missing and the threshold percentile of Y_1 is 50%

the result where the missing rate is 10% and the replacement rate is 50%. In the first case with large missingness, the estimated variances have a bit positive bias. When $m_3(=10)$ is large, the boxplot of the nominal coverage rates is more compact and around 95%. Negative values of estimated variances appear in all the three setting of (m_1, m_2, m_3) with large missingness. While the synthesis rate is large, only when $m_3(=10)$ is large, the estimated variances has negative values. The setting $(m_1 = 3, m_2 = 10, m_3 = 3)$ gives the most compact boxplot and the median closest to 95% with repect to the nominal coverage rate. Meanwhile, when $(m_1 = 3, m_2 = 10, m_3 = 3)$ most values of the estimated variance are close to the truth. In general, when the replacement rate is large, we recommend a larger value for m_2 comparing to m_1 and m_3. When the missingness proportion is large, larger values of m_2 and m_3 will generate better estimates.

6. Conclusions

We proposed a new approach to protect the privacy of census public use microdata when some values are missing. A new set of combining rules are derived to enable valid inference from the released data samples. However, some challenging issues still exist. How to adjust the negative estimated values of variances needs to be explored more. The direct posterior simulations can be a good solution. Furthermore, how to choose a proper set of the values of (m_1, m_2, m_3) is crucial. We implement some simulationss to study which one will affect a lot repectively for large missingness proportion and large replacement rate.

Acknowledgements

This work are supported by National Science Foundation grant NSF-SES-0751671.

[1] Barnard, J. and X. Meng (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research 8*, 17–36.

[2] Drechsler, J. and J. Reiter (2010). Sampling with synthesis: a new approach for releasing public use census microdata. *Journal of the American Statistical Associationw*.

[3] Heitjan, D. F. and R. J. A. Little (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics 40*, 13–29.

[4] Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*, 407–426.

[5] Raghunathan, T. E. and G. S. Paulin (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. In *Proceedings of the Section on Business and Economic Statistics of the American Statistical Association*, pp. 1–10.

[6] Reiter, J. (2008a). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitatio. *Statistics and Probability Letter 78*.

[7] Reiter, J. P. (2008b). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika 95*, 933–946.

[8] Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9*, 462–468.

[9] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*, 473–489.

[10] Schafer, J. L., T. M. Ezzati-Rice, W. Johnson, M. Khare, R. J. A. Little, and D. B. Rubin (1998). The NHANES III multiple imputation project. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 28–37.

[11] Schenker, N., T. E. Raghunathan, P. L. Chiu, D. M. Makuc, G. Zhang, and A. J. Cohen (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association 101*, 924–933.