Inference for Partially Synthetic, Public Use Microdata Sets

J. P. Reiter*

Key Words: Confidentiality, Disclosure, Multiple Imputation, Synthetic Data

Abstract

To avoid disclosures, one approach is to release partially synthetic, public use microdata sets. These are comprised of the units originally surveyed, but some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, are replaced with multiple imputations. Although partially synthetic approaches are currently used to protect public use data, valid methods of inference have not been developed for them. This article presents such methods. They are based on the concepts of multiple imputation for missing data but use different rules for combining point and variance estimates. The combining rules also differ from those for fully synthetic data sets developed by Raghunathan *et al.* (2003). The validity of these new rules is illustrated in simulation studies.

1 Introduction

When releasing data to the public, statistical agencies seek to provide detailed data without disclosing respondents' sensitive information. To reduce the risk of disclosures, agencies typically alter the original data for public release, for example by recoding variables, swapping data, or adding random noise to data values (Willenborg and de Waal, 2001). However, these methods can distort relationships among variables

*Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu

in the data set. They also complicate analyses for users: to analyze properly perturbed data, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach was proposed by Rubin (1993): release fully synthetic data sets comprised entirely of multiply-imputed rather than actual values. This can protect confidentiality, since identification of units and their sensitive data can be difficult when the released data are not actual, collected values. And, with appropriate imputation and estimation methods based on the concepts of multiple imputation (Rubin, 1987), the approach can allow data users to obtain valid inferences using standard, complete-data statistical methods and software. Such inferences can be made using the methods developed by Raghunathan *et al.* (2003), whose rules for combining point and variance estimates differ from those of Rubin (1987). Other discussions and variants of synthetic data approaches appear in Little (1993), Fienberg *et al.* (1996, 1998), Dandekar *et al.* (2002a,b), Franconi and Stander (2002, 2003), Polettini *et al.* (2002), Polettini (2003), and Reiter (2002, 2003).

Although no data producers have adopted the fully synthetic approach yet, some have adopted a variant of the approach: release partially synthetic data sets comprised of a mix of actual and multiply-imputed values. For example, to protect data in the U.S. Survey of Consumer Finances, the U.S. Federal Reserve Board replaces monetary values at high disclosure risk with multiple imputations, then releases a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). Another partially synthetic approach has been implemented by Abowd and Woodcock (2001) to protect data in longitudinal, linked data sets. They replace all values of some sensitive variables with multiple imputations, but leave other variables at their actual values. A third approach has been implemented by Liu and Little (2002), who develop an algorithm for simulating multiple values of key identifiers for selected units. All these partially synthetic approaches are appealing because they promise to maintain many of the benefits of fully synthetic data–

protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software–with decreased sensitivity to the specification of imputation models.

Surprisingly, even though partially synthetic data sets are being publicly released, the literature does not contain technical results on how to obtain inferences from them. At first glance, it may appear appropriate to use the inferential methods for multiple imputation of missing data in Rubin (1987). Unfortunately, as shown in this article, these methods can result in biased variance estimates. Furthermore, and also as shown, the methods developed by Raghunathan *et al.* (2003) for analyzing fully synthetic data are not valid when applied on partially synthetic data. New methods of inference are required.

This paper describes methods for obtaining inferences from multiply-imputed, partially synthetic data sets. The derivation of these methods also provides prescriptions for generating partially synthetic data. The paper is organized as follows. Section 2 presents the new methods of inference. Section 3 shows a derivation of these methods from a Bayesian perspective, and it discusses conditions under which the resulting inferences should be valid from a frequentist perspective. Section 4 describes simulation studies that illustrate the validity of these methods, as well as the ineffectiveness of competing rules for combining multiple point and variance estimates. Section 6 concludes with suggestions of future areas of research.

2 Inferences from multiply-imputed, partially synthetic data sets

Let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise. Let $I = (I_1, \ldots, I_N)$. Let Y_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; let Y_{nobs} be the $(N - n) \times p$ matrix of unobserved survey data for the units with $I_j = 0$; and, let $Y = (Y_{obs}, Y_{nobs})$. For simplicity, we assume that all sampled units fully respond to the survey. Let X be the $N \times d$ matrix of design variables for all N units in the population, e.g. stratum or cluster indicators or size measures. We assume that such design information is known approximately for all population units. It may come, for example, from census records or the sampling frame(s).

The agency releasing synthetic data, henceforth abbreviated as the *imputer*, constructs synthetic data sets based on the observed data, $D = (X, Y_{obs}, I)$, in a two-part process. First, the imputer selects the values from the observed data that will be replaced with imputations. Second, the imputer imputes new values to replace those selected values. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \ldots, Z_n)$. Let $Y_{rep,i}$ be all the imputed (replaced) values in the ith synthetic data set, and let Y_{nrep} be all unchanged (unreplaced) values of Y_{obs}. The $Y_{rep,i}$ are assumed to be generated from the Bayesian posterior predictive distribution of $(Y_{rep,i}|D, Z)$. The values in Y_{nrep} are the same in all synthetic data sets. Each synthetic data set, d_i, is then comprised of $(X, Y_{rep,i}, Y_{nrep}, I, Z)$. Imputations are made independently for $i = 1, \ldots, m$ times to yield m different synthetic data sets. These synthetic data sets are released to the public.

The values in Z can and frequently will depend on the values in D. For example, the imputer may choose to simulate sensitive variables or identifiers only for units in the sample with rare combinations of identifiers; or, the imputer may replace only those incomes above 100,000 with imputed values. To avoid bias, imputers should account for such selections by imputing from the posterior predictive distribution of Y for those units with $Z_j = 1$. In practice, this can be done by using only the units with $Z_j = 1$ as the data when finding the posterior distributions for imputations. Using all units with $I_j = 1$ can result in biased estimates or wider confidence intervals with overly conservative coverage rates, as illustrated in the simulations of Section 4.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the *analyst*, seeks inferences about some estimand $Q = Q(X, Y)$, where the notation $Q(X, Y)$ means that Q is a function of (X, Y). For example, Q could be the population mean of Y or the population regression coefficients of Y on X. In each synthetic data set d_i, the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator v. It is assumed that the analyst determines the q and v as if the synthetic data were in fact collected data from a random sample of (X, Y) based on the actual survey design used to generate I.

For $i = 1, \ldots, m$, let q_i and v_i be respectively the values of q and v in synthetic data set d_i. Under certain conditions to be described in Section 3, the analyst can obtain valid inferences for scalar Q by combining the q_i and v_i. Specifically, the following quantities are needed for inferences:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m. \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_p = b_m/m + \bar{v}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$. When q is a function of only (X, Y_{nrep}, I) and not any imputed values, the synthetic data inferences are identical to the observed data inferences; that is, the $q_i = q_{obs}$ and $v_i = v_{obs}$ for all i, and the $b_m = 0$. When n is large, inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_p = (m-1)(1 + r_m^{-1})^2$, where $r_m = (m^{-1} b_m/\bar{v}_m)$. In many cases, r_m^{-1} and hence ν_p will be large enough that a normal distribution provides an adequate approximation to the t-distribution. Extensions for multivariate Q are not presented here.

T_p differs from the variance estimator for multiple imputation of missing data, $T_m = (1 + 1/m)b_m + \bar{v}_m$ (Rubin, 1987). In the partially synthetic data context, the $\bar{v}_m$ estimates $\mathrm{Var}(q_{obs})$ and the b_m/m estimates the additional variance due to using a finite number of imputations. In the missing data context, the $\bar{v}_m$ and b_m/m have the same interpretations, but an additional b_m is needed to average over the nonresponse mechanism (Rubin, 1987, Ch. 4). This additional averaging is unnecessary in partially synthetic data settings, since the selection mechanism Z, which is set by the imputer, is not treated as stochastic.

T_p also differs from the variance estimator for analyzing fully synthetic data, $T_s = (1 + 1/m)b_m - \bar{v}_m$ (Raghunathan *et al.*, 2003). To generate fully synthetic data, new units are sampled off the frame(s) for each synthetic data set, and their data are imputed. As shown by Raghunathan *et al.* (2003), this re-sampling and imputation process results in $b_m - \bar{v}_m$ as an appropriate estimate of $\text{Var}(q_{obs})$. For partially synthetic data, the original units are released for each data set, so that $\bar{v}_m$ is an appropriate estimate of $\text{Var}(q_{obs})$.

3 Justification of new combining rules

This section shows a Bayesian derivation of the inferences described in Section 2 and conditions under which these inferences are valid from a frequentist perspective. These results are based on, and closely follow, the theory developed in Raghunathan *et al.* (2003).

3.1 Bayesian derivation

For this derivation, we assume that the same posterior distributions are used for inferences and imputations. The posterior distribution for $(Q|d^m)$, where $d^m = \{d_1, d_2, \ldots, d_m\}$, can be decomposed as

$$f(Q|d^m) = \int f(Q|d^m, D, B) f(D|d^m, B) f(B|d^m) dD dB \tag{5}$$

where $B = Var(q_i|D, Z)$. The integration with respect to $f(D|d^m, B)dD$ is only over the values of Y_{obs} that are replaced with imputations; the (X, Y_{nrep}, I) components of D remain fixed.

Given D, the synthetic data are irrelevant, so that $f(Q|d^m, D, B) = f(Q|D)$. We assume standard Bayesian asymptotics hold, so that $f(Q|D) \sim N(q_{obs}, v_{obs})$, where q_{obs} and v_{obs} are the posterior mean and variance of Q determined using D.

Integrating (5) over D, we obtain $f(Q|d^m, B)$. Since only q_{obs} and v_{obs} are needed for inferences about $(Q|D)$, for $f(D|d^m, B)$ it is sufficient to determine $f(q_{obs}, v_{obs}|d^m, B)$. We assume imputations are made so

that, for all i, $(q_i|D,B) \sim N(q_{obs},B)$ and $(v_i|D,B) \sim (v_{obs}, << B)$. Here, the notation $F \sim (G, << H)$ means that the random variable F has a distribution with expectation of G and variability much less than H. In actuality, v_i is typically centered at a value larger than v_{obs}, since synthetic data incorporate uncertainty due to drawing values of the parameters. For large sample sizes n, this bias should be minimal. The assumption that $E(q_i|D,B) = q_{obs}$ should be reasonable when the imputations are drawn from the correct posterior distribution of Y for those units with $Z_j = 1$.

Assuming flat priors for q_{obs} and v_{obs}, standard Bayesian theory implies that $(q_{obs}|d^m,B) \sim N(\bar{q}_m, B/m)$ and $(v_{obs}|d^m,B) \sim (\bar{v}_m, << B/m)$. Hence, the posterior mean and variance of $(Q|d^m,B)$ are

$$E(Q|d^m,B) = E(E(Q|D,d^m,B)|d^m,B) = E(q_{obs}|d^m,B) = \bar{q}_m \quad (6)$$

$$Var(Q|d^m,B) = E(Var(Q|D,d^m,B)|d^m,B) + Var(E(Q|D,d^m,B)|d^m,B) = \bar{v}_m + B/m \quad (7)$$

Since all the convolutions involve normal distributions, $f(Q|d^m,B) \sim N(\bar{q}_m, \bar{v}_m + B/m)$.

To integrate this distribution over $f(B|d^m)$, we use the fact that $\left((m-1)b_m B^{-1}|d^m\right) \sim \chi^2_{m-1}$ and, following the approximation in Rubin (1987, pp. 90-92), fit the first two moments of $\bar{v}_m + B/m$ to a mean-square random variable. The resulting approximation to the posterior distribution of Q is $(Q|d^m) \sim t_{\nu_p}(\bar{q}_m, T_p)$, where ν_p is as defined in Section 2.

3.2 Randomization validity

For inferences based on (1) - (4) to have valid frequentist properties, we require two conditions. First, the analyst must use randomization valid estimators, q and v. That is, when q and v are applied on D to get q_{obs} and v_{obs}, the $(q_{obs}|X,Y) \sim N(Q,U)$ and $(v_{obs}|X,Y) \sim (U, << U)$, where the relevant distribution is that of I. Second, the synthetic data generation methods must be proper in a sense similar to Rubin (1987). Specifically, the data generation methods should satisfy the following conditions:

C1: Averaging over imputations of $Y_{rep,i}$, it is required that (i) $(q_i|X,Y,I,Z) \sim N(q_{obs}, B)$; (ii) $(b_m|X,Y,I,Z) \sim (B, << B)$; and, (iii) $(\bar{v}_m|X,Y,I,Z) \sim (v_{obs}, << B/m)$, where $B = Var(q_i|X,Y,I,Z)$.

C2: Averaging over the sampling and replacement mechanisms $(I,Z|X,Y)$, it is required that $(B|X,Y) \sim (B_0, << U)$ where $B_0 = E(b_m|X,Y)$.

Essentially, these conditions require the synthetic data be generated so that the q_i are unbiased for q_{obs}, the b_m is unbiased for B_0, and the $\bar{v}_m$ is unbiased for v_{obs}. Further discussion of proper imputation can be found in Rubin (1987, Ch. 4).

Using these assumptions, it follows that

$$\begin{aligned} E(\bar{q}_m|X,Y) &= E(E(\bar{q}_m|X,Y,I,Z)|X,Y) = E(q_{obs}|X,Y) = Q \qquad (8) \\ Var(\bar{q}_m|X,Y) &= E(Var(\bar{q}_m|X,Y,I,Z)|X,Y) + Var(E(\bar{q}_m|X,Y,I,Z)|X,Y) \\ &= E(B|X,Y)/m + Var(q_{obs}|X,Y) = B_0/m + U, \qquad (9) \end{aligned}$$

Since $(q_{obs}|X,Y)$ and the $(q_i|X,Y,I,Z)$ are assumed to have normal distributions, it follows that $(\bar{q}_m|X,Y) \sim N(Q, B_0/m + U)$.

When C1 and C2 hold, T_p is an unbiased estimator of $B_0/m + U$. The t-approximation is justified using the method outlined in Rubin (1987, pp. 128-129). Specifically, the t-approximation follows since $((m-1)b_m B_0^{-1}|X,Y) \sim \chi^2_{m-1}$, and the degrees of freedom of a chi-squared random variable equals two times the square of its expectation over its variance.

4 Simulation studies

This section illustrates the validity of these new combining rules, as well as the ineffectiveness of T_m and T_s as variance estimators, using simulation studies of partially synthetic strategies. Section 4.1 describes two

studies in which the imputer generates synthetic data only for selected units. Section 4.2 describes a study in which the imputer generates synthetic data for all values of one survey variable, leaving the others at their observed values. For illustrations, the simulations use artificial data and correct posterior distributions for imputations. Of course, in real settings the correct imputation model typically is not known and must be estimated using the observed data and subject-matter expertise. For all simulations, the population sizes are considered infinite so that finite population corrections factors are ignored.

4.1 Imputation for selected units

Imputers may decide to replace the observed values for some units in the collected data, then release a mixture of the imputed and observed values. This strategy is employed in two simplistic although illustrative simulations, the first involving a single variable and the second four variables.

4.1.1 Simulations using a single variable

Each observed dataset, D, is comprised of $n = 100$ values drawn randomly from $Y \sim N(0, 10^2)$. Two different schemes are used to specify the units with $Z_j = 1$, so that two sets of partially synthetic data sets are generated for each D. The first scheme, labelled "Random", replaces Y for 20 units randomly sampled from D. The second scheme, labelled "Big Y", replaces Y only for units with $Y_j > 10$.

For each D, and for each scheme, there are $m = 5$ synthetic data sets $d_i = (Y_{rep,i}, Y_{nrep}, I, Z)$, for $i = 1, \ldots, 5$. The $Y_{rep,i}$ are generated by using a Bayesian bootstrap (Rubin, 1987, pp.123-124), which draws values of Y from a donor pool comprised of selected values of Y_{obs}. Let Y_{elig} be the $n_0 \times 1$ vector of values of Y_{obs} that make up the donor pool. Let $n_{rep} = \sum_{j=1}^{100} Z_j$. The Bayesian bootstrap proceeds as follows:

1. Draw $(n_0 - 1)$ uniform random numbers. Sort these numbers in ascending order. Label these ordered numbers as $a_0 = 0, a_1, a_2, \ldots, a_{n_0-1}, a_{n_0} = 1$.

Table 1: Simulation results when imputing single variable

Scheme and Imputation Method	Avg. $\bar{q}_5$	Var $\bar{q}_5$	Avg. T_p	Avg. T_m	Coverage of 95% CIs Using T_p	Using T_m
$Z_j = 1$ for 20 randomly selected units						
SELECT	0.024	1.097	1.067	1.420	94.5%	96.7%
ALL	0.020	1.233	1.044	1.281	92.6%	94.9%
$Z_j = 1$ for units with $Y_j > 10$						
SELECT	0.016	1.031	1.011	1.068	94.5%	95.0%
ALL	-2.383	0.796	0.736	0.921	20.7%	28.8%
Observed data results*	0.016	1.021	1.000		94.5%	

* The column labels do not apply for this row. The average of the $q_{obs} = 0.016$, the variance of the $q_{obs} = 1.021$, the average of the $v_{obs} = 1.000$, and 94.5% of the five thousand 95% observed-data confidence intervals cover zero.

2. Draw n_{rep} uniform random numbers, $u_1, u_2, \ldots, u_j, \ldots, u_{n_{rep}}$. For each of these u, impute $Y_{elig,j}$ when $a_{j-1} < u \leq a_j$.

This Bayesian bootstrap is not likely to be used to impute data in real settings, since data sets contain more than one variable. It is used here because it provides straightforward, proper imputations for this illustration.

As mentioned in Section 2, the correct posterior predictive distribution is $f(Y|D, Z)$, not $f(Y|D)$. This implies that the donor pool, Y_{elig}, should equal the set $\{Y_j : Z_j = 1\}$. This set is labelled "SELECT." For comparisons, synthetic values also are imputed using the donor set $\{Y_j : I_j = 1\}$. This set is labelled "ALL." Imputations based on ALL donors do not meet condition C1 in Section 3.2, since $E(q_i|X, Y, I, Z) = \left(\sum_{j=1}^{100-n_{rep}} y_{nrep,j} + n_{rep}\bar{y}_{obs}\right)/n \neq \bar{y}_{obs}$, whereas imputations based on SELECT donors are proper.

Table 1 summarizes the results from 5,000 runs of this simulation. For both the Random and Big Y schemes, the averages of the $\bar{q}_5$ based on the SELECT donors approximately equal the average of q_{obs}. In the Random scheme, the $\bar{q}_5$ based on ALL donors is also unbiased, because $E(\bar{y}_{nrep}|X, Y, I) = q_{obs}$ when averaged over Z (which is in fact stochastic in this scheme). However, when using ALL donors in the Big Y

scheme, $\bar{q}_5$ has a large, negative bias. This results because imputed values are not restricted to be greater than 10 when using ALL donors.

In both the Random and Big Y schemes, 94.5% of the 5,000 synthetic 95% confidence intervals based on T_p and the SELECT donors cover zero. This rate is identical to the 94.5% coverage rate for the confidence intervals based on the observed data ($q_{obs} \pm 1.96\sqrt{v_{obs}}$). The nominal rates are less than 95% due to simulation error. The 2-3% difference between the averages of the T_p and the $\text{Var}(\bar{q}_5)$ roughly equals the difference between the average v_{obs} and $\text{Var}(q_{obs})$. The usual multiple imputation variance estimator, T_m, tends to overestimate the $\text{Var}(\bar{q}_5)$, leading to overly conservative confidence interval coverage rates, showing that T_m is not the correct variance estimator when analyzing properly imputed, partially synthetic data.

When imputations are based on ALL donors–an improper imputation method–in the Random scheme, T_p is negatively biased, and only 92.6% of the synthetic 95% confidence intervals cover zero. Using T_m increases the coverage rate to 95%, suggesting that it is safer to use T_m instead of T_p when ALL units are used for imputations. The confidence intervals based on ALL and T_m are on average wider than those based on SELECT and T_p. This illustrates the advantage of conditioning on Z to obtain proper imputations, even when the scheme used to set the $Z_j = 1$ does not depend on the values of Y.

Although not shown in Table 1, the variance estimator for fully synthetic data, T_s, is negative in every one of the 5,000 simulations for both schemes and both imputation methods. Clearly, although valid for fully synthetic data (Raghunathan *et al.*, 2003), T_s is not generally appropriate for partially synthetic data.

4.1.2 Simulations using four variables

Each observed dataset, D, is comprised of $n = 200$ values of four variables, (Y_1, Y_2, Y_3, Y_4), generated as follows: $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ has all variances equal to one and all covariances equal to 0.5; and, $(y_4|y_1, y_2, y_3) \sim N(10y_1 + 7y_2 + 4y_3, 25^2)$. To fix ideas, the variable Y_1 can be considered a key identifier and Y_4 the sensitive variable. The plan is to simulate values of the sensitive Y_4 for all units with "unusual"

Table 2: Simulation results when imputing Y_4 for units with $Y_1 > 1$

Type of Inference	Avg. $\bar{q}_5$	Var($\bar{q}_5$)	Avg. T_p	Avg. T_m	Coverage of 95% CIs Using T_p	Using T_m
Estimand is β						
SELECT	10.02	5.45	5.68	8.97	95.3%	98.2%
ALL	10.04	5.89	5.28	7.57	93.7%	96.9%
Observed data*	10.00	4.70			95.5%	
Estimand is α						
SELECT	9.25×10^{-3}	4.49×10^{-6}	4.76×10^{-6}	6.97×10^{-6}	95.4%	97.9%
ALL	9.59×10^{-3}	5.03×10^{-6}	4.75×10^{-6}	6.31×10^{-6}	94.1%	96.5%
Observed data*	9.66×10^{-3}	4.26×10^{-6}			95.4%	
Estimand is $\bar{Y}_4$						
SELECT	-1.45×10^{-2}	4.97	5.01	6.09	95.0%	96.6%
ALL	-1.24×10^{-3}	5.19	4.82	5.59	93.8%	95.4%
Observed data*	-2.34×10^{-3}	4.76			94.5%	

* The column labels do not apply for this row. These are the averages of the q_{obs}, the variance of the q_{obs}, and the percentage of 95% observed-data confidence intervals that cover their Q.

values of the key identifier, defined as $Y_1 > 1$. Hence, Y_{nrep} is comprised of sampled values of (Y_1, Y_2, Y_3) and values of Y_4 for those units with $Y_1 \leq 1$. Typically, around 30 units per observed data set have $Y_1 > 1$.

As before, we examine two schemes for determining the posterior predictive distribution for imputations. SELECT uses only the units with $Z_j = 1$ as the data for the posteriors, and ALL uses all observed units. Imputations under each scheme are made by (i) drawing values of the parameters of the regression of Y_4 on (Y_1, Y_2, Y_3) from their posterior distribution, which is estimated using either the SELECT or ALL units, and (ii) drawing values of Y_4 for units with $Z_j = 1$ using the drawn values of parameters. There are $m = 5$ synthetic data sets generated for each observed data set D.

The estimands of interest include β, the regression coefficient of Y_1 in the linear regression of Y_4 on (Y_1, Y_2, Y_3); α, the regression coefficient of Y_4 in the regression of Y_1 on (Y_2, Y_3, Y_4); and $\bar{Y}_4$, the population average of Y_4. For inferences about β and α, q is the usual ordinary least squares estimator and v its variance estimator. For inferences about $\bar{Y}_4$, q is the sample average and v its standard error.

Table 2 summarizes results from 5,000 runs of this simulation. When imputations are based on the SELECT units, the averages of the $\bar{q}_5$ and T_p are within simulation errors of the averages of the q_{obs} and Var(q_5). Additionally, the coverage rates for the synthetic 95% confidence intervals are similar to the coverage rates for the observed data 95% confidence intervals. The T_m are substantially larger than the Var(q_5), resulting in coverage rates around 97%. Although not shown in Table 2, T_s is negative in all 5,000 simulation runs. Taken together, these results are consistent with the findings in Section 4.1.1: when imputations are drawn from a posterior distribution that conditions on Z, point and interval estimates based on T_p are more accurate than those based on T_m and T_s.

Although imputations based on ALL units are not proper, it is informative to examine the performances of T_p and T_m for such imputations. Imputers might base imputations on all observed units for practical reasons, for example because the units with $Z_j = 1$ do not provide sufficient data to fit the imputation models. The results mirror those in Section 4.1.1: the T_p underestimate the Var(q_5), leading to coverage rates around 94%, whereas using T_m increases coverage rates to around 96%, primarily due to the positive bias in T_m. This again suggests that, when imputers do in fact base imputations on all observed units even though only some $Z_j = 1$, analysts are safer using T_m as the variance estimator rather than T_p. Just as seen in Section 4.1.1, the intervals based on ALL units are typically wider than those based on SELECT units, suggesting that, when possible, imputers are better off basing imputations only on the units with $Z_j = 1$.

4.2 Imputation of all values of Y for one variable

Each observed data set is comprised of $n = 200$ values of four variables generated as follows: $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \mathbf{I})$ where $\mathbf{I}$ is the identity matrix; and, $(y_4|y_1, y_2, y_3) \sim N(10y_1 + 10y_2 + 10y_3, 25^2)$. Hence, the $Y_{nrep} = (Y_1, Y_2, Y_3)$. Values of Y_4 are imputed from the Bayesian posterior predictive distribution of $(Y_4|Y_{obs})$, dervied by fitting the regression of Y_4 on (Y_1, Y_2, Y_3). All units have $Z_j = 1$ and are used as data for the posterior distributions. The estimands are the same as those described in Section 4.1.2.

Table 3: Simulation results when imputing an entire variable.

Estimand	Avg. q_{obs}	Avg. $\bar{q}_5$	Var q_{obs}	Var $\bar{q}_5$	Avg. T_p	Avg. T_m	Avg. T_s
β	9.95	9.94	3.19	4.46	4.54	11.10	4.63
α	.0137	.0135	6.12	7.69	7.94	17.30	5.17
$\bar{Y}_4$	.00	.00	4.55	5.83	6.00	12.30	2.87

Variances associated with α are multiplied by 10^6.

Table 3 summarizes the results from 5,000 simulation runs using $m = 5$ partially synthetic data sets. For all estimands, the averages of the $\bar{q}_5$ are practically identical to those of the q_{obs}. Additionally, the estimated variances based on T_p are close to the actual variances of the $\bar{q}_5$. The slight upward bias results because $\bar{v}_m$ tends to overestimate v_{obs}, as explained in Section 3.1. The T_m on average overestimate the Var($\bar{q}_5$) by factors of more than two, and the T_s severely underestimate the Var($\bar{q}_5$) for α and $\bar{Y}_4$. These problems are not due to small m; in simulations with large m they persist. Although errors of these magnitudes may not occur in other settings, the results in this simple setting again indicate that T_m and T_s are not appropriate in general for analyzing partially synthetic data, especially when synthesizing entire variables.

Imputers have incentive to release small numbers of synthetic data sets. Each additional data set requires extra storage, and more importantly, releasing too many data sets might jeopardize confidentiality if intruders somehow combine the imputed values to learn about the actual values. Table 4 displays results of independent replications of 5000 simulation runs using different values of m. Point estimates are unbiased for all three estimands and so are not displayed in the table. The 95% confidence interval coverage rates are close to 95% for all values of m greater than two. The inflations in the T_p are again due to positive biases in the $\bar{v}_m$.

Table 4 illustrates that, when imputing entire variables, substantial efficiency gains can be made by increasing m beyond five. The amount of efficiency gain depends on the magnitude of b_m. When b_m is small relative to $\bar{v}_m$, for example when imputing values only for a small number of selected units, efficiency gains from increasing m will not be large. For any partially synthetic strategy, imputers can compare gains in efficiency with potential tradeoffs in confidentiality by simulation studies of intruder behavior on different

Table 4: Sensitivity of partially synthetic inferences to value of m.

Setting	Var($\bar{q}_m$)	Avg. T_p	95% CI cov.
Inference for β			
$m = 2$	6.52	6.50	92.7
$m = 3$	5.38	5.38	94.4
$m = 4$	4.64	4.89	95.4
$m = 5$	4.46	4.54	95.1
$m = 10$	3.87	3.88	94.4
$m = 50$	3.30	3.37	95.1
Inference for α			
$m = 2$	10.62	10.89	93.4
$m = 3$	8.92	9.15	94.9
$m = 4$	8.41	8.45	94.9
$m = 5$	7.69	7.94	95.4
$m = 10$	6.99	7.02	94.8
$m = 50$	6.05	6.28	95.5
Inference for $\bar{Y}$			
$m = 2$	8.13	7.96	93.4
$m = 3$	6.51	6.86	95.5
$m = 4$	6.11	6.33	95.6
$m = 5$	5.83	6.00	95.3
$m = 10$	5.13	5.38	95.4
$m = 50$	4.66	4.87	95.5

Variances associated with α are multiplied by 10^6.

numbers of released synthetic data sets.

5 Concluding Remarks

The simulations in this article illustrate that the usual rules for combining multiply-imputed data sets result in positively biased variance estimates when applied on partially synthetic data. The new rules presented here appear to remedy this problem, thereby leading to more reliable inferences. Further research is needed to assess the performance of these new rules when using partially synthetic strategies for genuine data, for which the correct imputation models are unlikely to be known. The simulations and theory also suggest that, when possible, imputers should use only units with values selected for replacement as the data when estimating posterior predictive distributions for imputations. Further examination of this prescription when

simulating more than one variable in genuine data sets would be valuable. Lastly, this article does not examine the implications of various partially synthetic data strategies for protecting confidentiality, nor does it compare partially synthetic approaches to alternative techniques for disclosure control. Such comparisons would help imputers determine whether partially synthetic approaches are appropriate for their public use microdata releases.

Acknowledgments

This work was supported by the United States Bureau of the Census through a contract with Datametrics Research. The author thanks Trivellore Raghunathan, Donald Rubin, Laura Zayatz for providing statistical support and general motivation for this research, and two referees and an associate editor for their valuable comments and suggestions.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.

Dandekar, R. A., Cohen, M., and Kirkendall, N. (2002a). Sensitive micro data protection using Latin hypercube sampling technique. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 117–125. Berlin: Springer-Verlag.

Dandekar, R. A., Domingo-Ferrer, J., and Sebe, F. (2002b). LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 153–162. Berlin: Springer-Verlag.

Fienberg, S. E., Makov, 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., Steele, R. J., and Makov, U. E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. In *Proceedings of Bureau of Census 1996 Annual Research Conference*, 87–105.

Franconi, L. and Stander, J. (2002). A model based method for disclosure limitation of business microdata. *The Statistician* **51**, 1–11.

Franconi, L. and Stander, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing* forthcoming.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

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

Liu, F. and Little, R. J. A. (2002). Multiple imputation and statistical disclosure control in microdata. Presentation at *2002 Joint Statistical Meetings* in New York.

Polettini, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing* forthcoming.

Polettini, S., Franconi, L., and Stander, J. (2002). Model-based disclosure protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 83–96. Berlin: Springer-Verlag.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* forthcoming.

Reiter, J. P. (2002). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Reiter, J. P. (2003). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* forthcoming.

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

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

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.