Selecting the Number of Imputed Datasets When Using Multiple Imputation for Missing Data and Disclosure Limitation

Jerome P. Reiter, Duke University*

Abstract

Multiple imputation can handle missing data and disclosure limitation simultaneously. First, fill in the missing data to generate m completed datasets, then replace confidential values in each completed dataset with r imputations. I investigate how to select m and r.

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

1 Introduction

When statistical agencies disseminate data to the public, they strive to release files that are safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, informative for a wide range of statistical analyses, and easy for users to analyze with standard statistical methods. Often, however, agencies cannot release data in their collected form, because doing so would disclose some survey respondents' identities or attributes. Agencies do the obvious things to protect confidentiality before releasing data, such as stripping unique identifiers like names, social security numbers, and addresses. However, these actions alone may not eliminate the risk of disclosures when key identifying variables—e.g., age, sex, race,

*Jerome P. Reiter, Institute of Statistics and Decision Sciences, Box 90251, Durham, NC 27708. This research was supported by the NSF-ITR grant, Info Tech Challenges for Secure Access to Confidential Social Science Data, # 0427889.

and marital status—remain on the file. These keys can be used to match units in the released data to other databases. Many agencies therefore alter values of key identifiers, and possibly values of sensitive variables, before releasing the data (Willenborg and de Waal, 2001).

Several authors, beginning with Rubin (1993), have proposed using multiple imputation as a way to release disclosure-proofed datasets with high utility. In Rubin's original approach, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic dataset, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these datasets to the public. These are called *fully synthetic* datasets. Releasing fully synthetic data can preserve confidentiality, since identification of units and their sensitive data can be difficult when the released data are not actual, collected values. Furthermore, using appropriate data generation and estimation methods of Raghunathan *et al.* (2003)—based on the concepts of multiple imputation (Rubin, 1987) for missing data—analysts can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software, at least for inferences congenial to the model used to generate the data. Provided the agency releases some description of this model, analysts can determine whether or not their questions can be answered using the fully synthetic data. Other attractive features of fully synthetic data are described by Rubin (1993), Raghunathan *et al.* (2003), Raghunathan (2003), and Reiter (2002, 2004a, 2005a,b).

Although no statistical agencies have released fully synthetic datasets as of this writing, some agencies use or are considering a variant of the multiple imputation approach proposed by Little (1993): release datasets comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic* datasets. For example, the U.S. Federal Reserve Board protects data in the U.S. Survey of Consumer Finances by replacing monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). The U.S. Bureau of the Census and

Abowd and Woodcock (2001) protect data in longitudinal, linked datasets by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values. The Bureau of the Census currently is researching the possibility of releasing partially synthetic public use files for the Survey of Income and Program Participation and the American Communities Survey. Partially synthetic approaches are appealing because they can maintain the primary 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 (Reiter, 2003).

When some data are missing, it is logical to impute the missing and partially synthetic data simultaneously. Multiple imputation is appealing for handling nonresponse because it moves the burden of dealing with the missing data off of data analysts and on to data producers, who typically have greater resources than analysts. When the imputation models meet certain conditions (Rubin, 1987, Chapter 4), analysts of the completed datasets can obtain valid inferences using complete-data statistical methods and software. Specifically, the analyst computes point and variance estimates of interest with each dataset and combines these estimates using simple formulae developed by Rubin (1987). These formulae automatically propagate the uncertainty introduced by imputation through the analysts' inferences, enabling analysts to focus on modeling issues rather than estimation technicalities.

To adapt multiple imputation to handle nonresponse and disclosure simultaneously, agencies can use the procedure proposed by Reiter (2004b). First, the agency fills in the missing data, generating m multiply-imputed datasets. Second, the agency replaces the values at risk of disclosure in each imputed dataset with r multiple imputations, ultimately releasing $M = m * r$ nested multiply-imputed datasets. Valid inferences can be obtained by combining point and variance estimates from the multiply-imputed datasets as described in Reiter (2004b).

In this paper, I consider the use of simultaneous multiple imputation for missing data and for partially synthetic data. The existing literature contains little guidance on implementing this approach in practice.

One critical issue is the selection of m and r. For example, for a fixed value of M, is it preferable to select large m or large r? This paper provides such guidance by using simulation studies to explore the effects on data utility of different allocations of m and r. The studies suggest that agencies can improve accuracy by making m large relative to r. The improvements can be substantial when the fractions of missing information are large.

2 Review of inferential methods for multiple imputation

For a finite population of size N, let $I_j = 1$ if unit j is selected in the survey, and $I_j = 0$ otherwise, where $j = 1, 2, \ldots, N$. Let $I = (I_1, \ldots, I_N)$. Let R_j be a $p \times 1$ vector of response indicators, where $R_{jk} = 1$ if the response for unit j to survey item k is recorded, and $R_{jk} = 0$ otherwise. Let $R = (R_1, \ldots, R_N)$. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $n \times p$ matrix of survey data for the n units with $I_j = 1$; Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. 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 for all population units, for example from census records or the sampling frame(s). When it is not known for some units, X can be treated as part of Y for those units. Finally, we write the observed data as $D = (X, Y_{obs}, I, R)$.

Let $Z_j = 1$ if unit j is selected to have any of its 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_{nrep} be the values in Y_{obs} that are not replaced; these remain constant across all synthetic datasets.

To generate the M synthetic datasets, first the agency fills in values for Y_{mis} with draws from the Bayesian posterior predictive distribution of $(Y_{mis} \mid D)$, or approximations of that distribution such as those of Raghunathan *et al.* (2001). These draws are repeated independently $l = 1, \ldots, m$ times to obtain m completed datasets, $D^{(l)} = (D, Y_{mis}^{(l)})$. Once these are generated, in each $D^{(l)}$ the agency imputes replacement values $Y_{rep}^{(l,i)}$ for those units with $Z_j = 1$, drawing from the posterior distribution for $(Y_{rep} \mid D^{(l)}, Z)$, or a

close approximation of it. These draws are repeated independently r times, so that we obtain $i = 1, \ldots, r$ synthetic datasets, $d^{(l,i)} = (X, Y_{nrep}, Y_{mis}^{(l)}, Y_{rep}^{(l,i)}, I, R, Z)$ for each l. Each $d^{(l,i)}$ includes a label indicating the l of the $D^{(l)}$ from which it was drawn. A total of $M = mr$ datasets are generated and are released to the public.

The user of these synthetic public use datasets, henceforth labeled the *analyst*, seeks inferences about some estimand $Q = Q(X, Y)$, for example the population mean of Y or the population regression coefficients of Y on X. In each $d^{(l,i)}$, the analyst estimates Q with some estimator q and the variance of q with some estimator u. It is assumed the analyst specifies the point and variance estimators, q and u, by acting as if each $d^{(l,i)}$ was in fact collected data from a random sample of (X, Y) based on the original sampling design I. For $l = 1, \ldots, m$ and $i = 1, \ldots, r$, let $q^{(l,i)}$ and $u^{(l,i)}$ be respectively the values of q and u in dataset $d^{(l,i)}$. The following quantities are needed for inferences about scalar Q:

$$\bar{q}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} q^{(l,i)}/(mr) = \sum_{l=1}^{m} \bar{q}^{(l)}/m \quad (1)$$

$$\bar{w}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} (q^{(l,i)} - \bar{q}^{(l)})^2/m(r-1) = \sum_{l=1}^{m} w^{(l)}/m \quad (2)$$

$$b_M = \sum_{l=1}^{m} (\bar{q}^{(l)} - \bar{q}_M)^2/(m-1) \quad (3)$$

$$\bar{u}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} u^{(l,i)}/(mr). \quad (4)$$

The $\bar{q}^{(l)}$ is the average of the point estimates in each group of datasets indexed by l, and the $\bar{q}_M$ is the average of these averages across l. The $w^{(l)}$ is the variance of the point estimates for each group of datasets indexed by l, and the $\bar{w}_M$ is average of these variances. The b_M is the variance of the $\bar{q}^{(l)}$ across synthetic datasets. The $\bar{u}_M$ is the average of the estimated variances of q across all synthetic datasets. This notation differs slightly from that in Reiter (2004b) to distinguish the within-group and between-group quantities more clearly.

As described in Reiter (2004b), the analyst can use $\bar{q}_M$ to estimate Q and

$$T_M = (1 + 1/m)b_M - \bar{w}_M/r + \bar{u}_M \tag{5}$$

to estimate the variance of $\bar{q}_M$. Inferences can be based on the t-distribution, $(Q - \bar{q}_M) \sim t_{\nu_M}(0, T_M)$, with degrees of freedom

$$\nu_M = \left(\frac{((1 + 1/m)b_M)^2}{(m-1)T_M^2} + \frac{(\bar{w}_M/r)^2}{m(r-1)T_M^2} \right)^{-1}. \tag{6}$$

It is necessary for $r > 1$ to enable the analyst to estimate $\bar{w}_M$. If $r = 1$, the T_M collapses to Rubin's (1987) variance formula for multiple imputation for missing data, $T = (1+1/m)b_m+\bar{u}_m$. This is a biased estimator of the variance when multiple imputation is used for replacing observed data (Reiter, 2003).

It is possible for $T_M < 0$, especially when m and r are small enough so that b_M and $\bar{w}_M$ are estimated with high variance. Adjustments for negative variance estimates have not been proposed in the existing literature. Here I propose that, when $T_M < 0$, analysts use the conservative approximation,

$$T_M^{adj} = (1 + 1/m)b_M + \bar{u}_M. \tag{7}$$

This mimics the variance estimator for multiple imputation for missing data. The corresponding degrees of freedom is

$$\nu_M^{adj} = (m-1)(1 + m\bar{u}_m/((m+1)b_m))^2 \tag{8}$$

which is the degrees of freedom for multiple imputation for missing data (Rubin, 1987). This proposal will be employed in the simulation studies.

3 Simulation studies

Agencies utilizing multiple imputation for missing data and disclosure limitation must choose values of m and r. This choice is guided by two concerns: data utility and disclosure risk. Generally, the larger the values of m and r, the smaller the variances of the $\bar{q}_M$, but the more information available for ill-intentioned users to attempt to learn about respondents' identities and sensitive attributes (Reiter, 2004b). To evaluate different allocations of m and r, I assume a fixed M. Agencies would select M after studying the risk and utility of different values, for example following the methods outlined in Reiter (2005c).

For a fixed value of M, there is little difference in the disclosure risks associated with different allocations of m and r. This is because the Y_{nrep}, which contain actual data and so provide some information about identities and attributes, are identical across all $d^{(l,i)}$ regardless of the allocation. Additionally, there are M available imputations for each replaced value of Y_{obs}—which the user can average in attempts to make disclosures, as illustrated by Reiter (2005c)—again regardless of the allocation.

The utility of the released data are directly affected by the choice of m and r, since these impact the $Var(\bar{q}_M)$. To evaluate these effects, I use simulation studies of three allocations: $(m = 8, r = 2), (m = 4, r = 4), (m = 2, r = 8)$. The $M = 16$ is selected for convenience but is roughly the size of M that agencies might use in practice. Each complete dataset, D, comprises $n = 1000$ values drawn randomly from $Y \sim N(0, 10^2)$. The population size is considered infinite so that finite population correction factors are ignored. To create missing data, I select random samples of units in D and make their Y missing. I then select random samples of units in Y_{obs} and replace their values with imputations, which simulates making partially synthetic data. This simulation design is simpler than what can arise in practice, but it is sufficient for illustrating the effects of different allocations of m and r on data utility.

The first stage of the process is to impute missing data to generate the $D^{(l)} = (Y_{obs}, Y_{mis}^{(l)}, I, R)$. The $Y_{mis}^{(l)}$ are drawn using a Bayesian bootstrap (Rubin, 1987, pp.123-124). This draws values of $Y_{mis}^{(l)}$ from donor pools comprising the n_{obs} values of Y_{obs}. The next stage is to replace selected values of Y_{obs} to generate

the partially synthetic datasets. As mentioned in Section 2, the correct posterior predictive distribution is $f(Y \mid D^{(l)}, Z)$, not $f(Y \mid D^{(l)})$. I generate the $Y_{rep}^{(l,i)}$ using standard Bayesian normal distribution theory, using only the values with $Z_j = 1$ to determine the posterior distributions. When $Z = I$, i.e. when all values are replaced, I simulate n values of $Y_{rep}^{(l,i)}$ so that each $d^{(l,i)}$ contains no values of Y_{obs} and no values of $Y_{mis}^{(l)}$.

The estimand of interest is the population mean, $Q = 0$. Each $q^{(l,i)}$ is the sample average of $(Y_{nrep}, Y_{mis}^{l}, Y_{rep}^{(l,i)})$, and each $u^{(l,i)}$ is the sample variance of $(Y_{nrep}, Y_{mis}^{(l)}, Y_{rep}^{(l,i)})$ divided by 1000. Table 1 summarizes the results from 5000 simulations involving several values of m and r, and differing rates of missing and synthetic data. For all scenarios, the average of the $\bar{q}_M$ is within simulation error of zero and so is not reported. Across all scenarios, the T_M is approximately unbiased for the $Var(\bar{q}_M)$, and the confidence interval coverages are close to nominal. As expected, variances increase with the fraction of missing data and with the fraction of replaced values. The fraction of missing data plays a larger role in the variance than does the fraction of replaced data. For example, going from 30% missing to 50% missing increases variances by around 33%, whereas going from 30% replaced to 100% replaced increases variances by 10% or less.

The variances are smaller when m is relatively large than when r is relatively large. This trend becomes pronounced as the fraction of missing data increases. Additionally, with large fractions of missing information, using $m = 2$ results in intervals that have less than 95% coverage. These results suggest that for large amounts of missing data, it is preferable to allocate more resources to the missing data imputations than to the replacement imputations. This is consistent with advice from the literature on multiple imputation for missing data: increase m when the fraction of missing data is large (Rubin, 1987).

However, making m large is not the proverbial free lunch: when replacing all values with synthetic data, increasing m increases the risks of obtaining negative variance estimates. This forces us to use the adjusted variance estimator and adjusted degrees of freedom more frequently. The resulting inferences are conservative, with higher than 95% coverage for the confidence intervals. The risk of negative variance estimates decreases as the amount of replaced data decreases.

4 Conclusions

Agencies considering the use of multiple imputation for missing data and disclosure limitation select the total number of synthetic datasets to release based on disclosure risk and data utility considerations. Once M is selected, the agency needs to allocate resources to imputing missing data and to replacing values. This paper illustrates that inferences can be made more efficient by allocating more to m than to r, especially for substantial fractions of missing information. For small values of M, a reasonable approach is to use similar values of m and r. This realizes some of the payoffs from reducing the variance due to missing data and reduces the risks of negative variance estimates.

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.

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.

Raghunathan, T. E. (2003). Evaluation of Inferences from Multiple Synthetic Data Sets Created Using Semiparametric Approach. Tech. rep. Report for the National Academy of Sciences Panel on Access to Confidential Research Data.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* 181–189.

Reiter, J. P. (2004a). New approaches to data dissemintation: A glimpse into the future (?). *Chance* **17**, 3, 12–16.

Reiter, J. P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

Reiter, J. P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

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.

Table 1: Results of 5000 simulations for each scenario.

	$Var(\bar{q}_M)$	$E(T_M)$	$T_M < 0$	95% CI Coverage Unadj.	Adj.
10% missing, 30% replaced					
$m = 8$ and $r = 2$	.115	.116	0	95.4	
$m = 4$ and $r = 4$	.118	.117	0	94.6	
$m = 2$ and $r = 8$	.121	.121	0	95.0	
30% missing, 30% replaced					
$m = 8$ and $r = 2$	.156	.151	0	94.7	
$m = 4$ and $r = 4$	.157	.157	0	94.8	
$m = 2$ and $r = 8$	.167	.167	0	93.4	
50% missing, 30% replaced					
$m = 8$ and $r = 2$	.207	.214	0	95.0	
$m = 4$ and $r = 4$	.225	.226	0	94.1	
$m = 2$ and $r = 8$	.244	.255	0	92.2	
10% missing, 100% replaced					
$m = 8$ and $r = 2$	.128	.125	264	93.4	98.6
$m = 4$ and $r = 4$	.128	.127	9	95.6	95.7
$m = 2$ and $r = 8$	.125	.130	0	94.1	
30% missing, 100% replaced					
$m = 8$ and $r = 2$	.156	.159	178	95.1	98.5
$m = 4$ and $r = 4$	.170	.166	3	94.8	94.8
$m = 2$ and $r = 8$	.170	.175	0	92.8	
50% missing, 100% replaced					
$m = 8$ and $r = 2$	.221	.223	81	96.4	97.9
$m = 4$ and $r = 4$	.243	.240	4	94.5	94.5
$m = 2$ and $r = 8$	.264	.260	0	90.5	

The first two columns are the variance of the $\bar{q}_M$ and the average of T_M across the 5,000 simulation runs. The third column is the number of times the $T_M < 0$. The last two columns show the percentages of the five thousand 95% confidence intervals based on that contain zero. The penultimate column is for intervals based on the t-distribution with ν_M degrees of freedom, and the last column is for intervals based on T_M and ν_M when $T_M > 0$ and on T_M^{adj} and ν_M^{adj} when $T_M < 0$.