Statistica Sinica **22** (2012), 795-811
doi:http://dx.doi.org/10.5705/ss.2010.140

BAYESIAN FINITE POPULATION IMPUTATION FOR DATA FUSION

Jerome P. Reiter

Duke University

Abstract: In data fusion, data owners seek to combine datasets with disjoint observations and distinct variables to estimate relationships among the variables. One approach is to concatenate the files, specify models relating the variables not jointly observed, and use the models to generate multiple imputations of the missing data. We show that the standard multiple imputation estimator of the sampling variance can have positive bias in such contexts. We present an approach for correcting this problem based on Bayesian finite population inference. We also present an approach for data fusion when some values are confidential and cannot be shared.

Key words and phrases: Confidentiality, disclosure, matching, multiple, sharing, synthetic.

1. Introduction

In many settings, researchers, policymakers, and other data analysts require variables that are not found in the same dataset. Mounting new surveys to obtain records with all variables measured can be an expensive endeavor. Therefore, many analysts in this situation seek to combine data from different sources, for example administrative records and survey data. We consider a special case of such data integration contexts: the analyst seeks to combine two data files that have disjoint records and some distinct variables. This is known as data fusion or statistical matching.

Data fusion arises in a variety of settings. It is used in marketing science to combine data from separate surveys, for example product purchase and media viewing data (Kamakura and Wedel (1997); van der Putten, Kok, and Gupta (2002); Gilula, McCulloch, and Rossi (2006)). It is used by economists to facilitate policy microsimulation modeling (Moriarity and Scheuren (2003)). It is employed by national statistical agencies including, for example, the Italian National Statistical Institute (D'Orazio, Di Zio, and Scanu (2002)) and Statistics Canada (Rässler (2002, pp.60-63)). For applications in other areas, see Kadane (2001, reprinted from a 1978 manuscript), Rodgers (1984), Moriarity and Scheuren (2001), and D'Orazio, Di Zio, and Scanu (2006).

Data fusion can be treated as a missing data problem. For example, Rubin (1986) suggests that the data owners concatenate the files, specify models relating the variables not jointly observed, and use the models to generate multiple imputations of the missing data in the concatenated file. The data owners can repeat the multiple imputation analysis for several specifications of the joint distribution to assess sensitivity of conclusions to those specifications. Rubin (1986) recommends that agencies perform at least two multiple imputations per specification to enable ascertainment of sampling variability.

In this article, we focus on the validity of the standard multiple imputation variance estimator for assessing sampling variability in data fusion given specification of the imputation models. Using simulations, we show that the standard multiple imputation variance estimator can have positive bias in these contexts; in fact, the bias can be orders of magnitude in size. We then present an approach for correcting this problem that is based on Bayesian finite population inference. The idea is as follows: (i) obtain or generate a large population that includes the records in the concatenated file, (ii) consider any unknown values for records in the constructed population as missing, and (iii) repeatedly complete the missing data in the population by imputing from models that are coherent with the observed marginal and assumed joint distributions. The analyst computes the quantity of interest in each of the completed populations, and combines these quantities with simple rules to obtain variance and interval estimates. We show via simulation that this leads to proper estimation of sampling variability and hence, for correctly specified data fusion models, well-calibrated inferences.

As an extension of this idea, we present a multiple imputation approach for data fusion when two data owners consider some values to be confidential, so that they are not willing to share the sensitive values in their databases with each other. The approach builds on the idea of partially synthetic data (Rubin (1993)); Little (1993); Reiter (2003); Abowd and Woodcock (2004); Reiter and Drechsler (2010)). First, to protect confidentiality, each owner replaces sensitive values in its data with r imputations drawn from models fit with its own data. Second, the owners share and concatenate the protected datasets to create r partially synthetic replicates. Third, the owners use Bayesian finite population inference on each concatenated dataset to obtain point and variance estimates. These estimates are combined using simple formulas derived in the appendix of this article. These formulas differ from standard multiple imputation (Rubin (1987)) and from standard partial synthesis (Reiter (2003)), because they are designed specifically to enable Bayesian finite population inferences in the data fusion context.

The remainder of the article is organized as follows. Section 2 reviews standard multiple imputation approaches for data fusion, and illustrates the potential

for biased estimation of sampling variances. Section 3 presents the Bayesian finite population imputation approach and shows that it leads to valid inference. Section 4 extends this approach to multiple imputation for confidential data fusion. Section 5 concludes with remarks about implementation of these proposals. Throughout the rest of the article, we refer to all data owners and analysts—who may be economists, marketers, statistical organizations, etc.—as agencies.

2. Data Fusion and Multiple Imputation

To fix the data fusion setting, suppose that there are two datasets, $D_1 = (X_1, Y_1)$ owned by Agency 1 and $D_2 = (X_2, Z_2)$ owned by Agency 2. Here, X, Y, or Z can be multivariate. None of the n_1 records in D_1 are in D_2, and the variables in Y_1 and Z_2 do not overlap. Hence, Y_2 is not observed for the n_2 records in D_2, and Z_1 is not observed for the n_1 records in D_1. The same variables comprise X_1 and X_2. As an illustration of this setting, X_1 and X_2 could include demographic variables available for all individuals, Y_1 could include wealth measures collected only by Agency 1, and Z_2 could include health measures collected only by Agency 2. Let $D = (X, Y_1, Z_2)$ be the concatenated file, where $X = (X_1, X_2)$ is the concatenation of X_1 and X_2.

2.1. Data fusion by multiple imputation

Data fusion can be treated as a missing data problem, where the complete dataset has (X, Y, Z) for all records in D_1 and D_2. At first glance, dealing with the missing data in the concatenated file may seem hopeless: there is no information about the joint distribution $f(Y, Z|X)$ in D. However, the agency can posit specifications of this joint distribution and perform analyses under those assumptions. For example, suppose that, possibly after suitable transformations, $f(Y, Z|X)$ is presumed to be a conditional bivariate normal distribution. The range of admissible values for $cov(Y, Z|X)$ is limited because the variance matrix must be positive definite. The agency can select several plausible values of $cov(Y, Z|X)$ from the admissible region, either manually or by drawing from a prior distribution, and perform the desired analyses under each selected covariance specification (Moriarity and Scheuren (2003); Rässler (2003)). The agency also may have auxiliary information about the unknown elements of the covariance matrix, for example from previous studies or from population data, that further constrains the admissible region; see D'Orazio, Di Zio, and Scanu (2006) for several examples of such constraints involving normal and multinomial data.

Viewing data fusion as a missing data problem suggests using missing data solutions for analyzing the concatenated file. In particular, the agency can use multiple imputation for data fusion, as we now describe. For any specification of $f(Y, Z|X)$, the agency creates $D^{(l)} = (X, Y^{(l)}, Z^{(l)})$, where $Y^{(l)} = (Y_1, Y_2^{(l)})$,

$Z^{(l)} = (Z_1^{(l)}, Z_2)$, and $Y_2^{(l)}$ and $Z_1^{(l)}$ are imputed values based on draws from the predictive distributions of Y and Z implied by $f(Y, Z|X)$. To enable estimation of sampling variability given $f(Y, Z|X)$, the agency creates several completed datasets, $(D^{(1)}, \ldots, D^{(m)})$, each containing independent draws of the missing Y_2 and Z_1. For $l = 1, \ldots, m$, let $q^{(l)}$ and $u^{(l)}$ be respectively the estimate of some population quantity Q and the estimate of the variance of $q^{(l)}$ in $D^{(l)}$. Analysts use $\bar{q}_m = \sum_{l=1}^m q^{(l)}/m$ to estimate Q, and use $T_m = (1+1/m)b_m + \bar{u}_m$ to estimate $\text{var}(\bar{q}_m)$, where $b_m = \sum_{l=1}^m (q^{(l)} - \bar{q}_m)^2/(m-1)$ and $\bar{u}_m = \sum_{l=1}^m u^{(l)}/m$. For large samples, inferences for Q are obtained from the t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom is $\nu_m = (m-1)(1+\bar{u}_m/((1+1/m)b_m))^2$.

Multiple imputation for data fusion is appealing. The agency can use the completed datasets for a variety of inferences. For example, it is straightforward to estimate the coefficients in the regression of X on (Y, Z), or any other regression for that matter, with completed datasets. The agency easily can ascertain the sampling uncertainty associated with these estimates: it need only combine point and variance estimates from the multiple datasets. Finally, the agency can share the datasets with others, which is an important benefit for government organizations and researchers charged with disseminating data.

2.3. Bias in T_m in data fusion

Unfortunately, T_m can be badly biased in data fusion settings, as we now illustrate via simulation. Let (X, Y, Z) have a multivariate normal distribution with means equal to zero, variances equal to one, $cov(X, Y) = 0.3$, $cov(X, Z) = 0.8$, and $cov(Y, Z) = 0.4$. The partial correlation of Y and Z given X is approximately $\rho_{YZ|X} = 0.2795$. To construct D, we simulate $n = 900$ values from this distribution. Let D_1 contain the values of (X, Y) for the first 500 records, and let D_2 contain the the values of (X, Z) for the next 400 records. Hence, Y and Z are never jointly observed.

We created multiply-imputed datasets using the Bayesian imputation approach of Rässler (2003). This approach enables imputation from theoretically correct models. Hence, any biases in T_m stem from inappropriateness of the multiple imputation combining rules rather than incorrect imputation models.

Let $\beta_{Y|X}$ and $\sigma^2_{Y|X}$ be, respectively, the true coefficient and residual variance in the regression of Y on X. Define $\beta_{Z|X}$ and $\sigma^2_{Z|X}$ analogously. For $i = 1, 2$, let $W_i = (1, X_i)$, i.e., a column of ones is appended to X_i for use in a regression. We use non-informative prior distributions for all parameters. The imputation strategy proceeded as follows.

2.1 Simulate values of $\sigma^2_{Y|X}$ and $\beta_{Y|X}$ from standard Bayesian posterior distributions estimated with a linear regression of Y_1 on X_1. Let $\sigma^{(l)2}_{Y|X}$ and $\beta^{(l)}_{Y|X}$ be the drawn values.

Table 1. Illustration of the potential bias in T_m and conservative 95% confidence interval coverage rates when using standard multiple imputation for data fusion.

Estimand	var($\bar{q}_{10}$)	Avg. T_{10}	95% CI Cov.
$\beta_{YZ\|X}$	0.0007	0.0056	100%
$\beta_{YX\|Z}$	0.0024	0.0057	99.7%
$\beta_{XY\|Z}$	0.0004	0.0010	99.7%
$\beta_{XZ\|Y}$	0.0008	0.0009	96.3%
μ_Y	0.0019	0.0020	94.9%
μ_Z	0.0017	0.0017	94.2%

2.2 Simulate values of $\sigma^2_{Z|X}$ and $\beta_{Z|X}$ from standard Bayesian posterior distributions estimated with a linear regression of Z_2 on X_2. Let $\sigma^{(l)2}_{Z|X}$ and $\beta^{(l)}_{Z|X}$ be the drawn values.

2.3 Compute the conditional covariance $\sigma^{(l)}_{YZ|X} = \rho_{YZ|X}\sigma^{(l)}_{Y|X}\sigma^{(l)}_{Z|X}$ using the posited $\rho_{YZ|X}$.

2.4 Impute Y_2 using $N(W_2\beta^{(l)}_{Y|X}+(Z_2–W_2\beta^{(l)}_{Z|X})\sigma^{(l)}_{YZ|X}/\sigma^{(l)2}_{Z|X}, \sigma^{(l)2}_{Y|X}–\sigma^{(l)2}_{YZ|X}/\sigma^{(l)2}_{Z|X})$. Let $Y_2^{(l)}$ be the imputed values.

2.5 Impute Z_1 using $N(W_1\beta^{(l)}_{Z|X}+(Y_1–W_1\beta^{(l)}_{Y|X})\sigma^{(l)}_{YZ|X}/\sigma^{(l)2}_{Y|X}, \sigma^{(l)2}_{Z|X}–\sigma^{(l)2}_{YZ|X}/\sigma^{(l)2}_{Y|X})$. Let $Z_1^{(l)}$ be the imputed values.

We add $Y_2^{(l)}$ and $Z_1^{(l)}$ to D to create $D^{(l)}$. This five step process was repeated $m = 10$ times resulting in $(D^{(1)}, \ldots, D^{(\text{A.6})})$, which were then used for analysis. In the simulation, we evaluated inferences for six estimands, including the means of Y (μ_Y) and Z (μ_Z), the coefficients of X ($\beta_{YX|Z}$) and Z ($\beta_{YZ|X}$) in the regression of Y on X and Z, and the coefficients of Y ($\beta_{XY|Z}$) and Z ($\beta_{XZ|Y}$) in the regression of X on Y and Z. Imputations in the simulations used the correct value of $\rho_{XY|Z}$ to illuminate the bias in T_m.

Table 1 summarizes the properties of the standard multiple imputation inferences obtained from 1,000 simulations. The averages of $\bar{q}_{10}$ are within simulation error of the true values and so are not reported. For the means, the multiple imputation inferences have good properties: T_m is approximately unbiased and coverage rates are near the nominal 95%. For $\beta_{XZ|Y}$, T_M appears to have a slightly positive bias, but the resulting inferences are reasonable. However, the multiple imputation inferences for the remaining three regression coefficients are unreliable. For these estimands, T_m has large positive bias, resulting in coverages rates near 100%. This is particularly troubling since estimation of relationships involving Y and Z is often the purpose of data fusion. We note that simulations in Rässler (2004) also exhibit conservative coverage.

Why is T_m unreliable for data fusion? The reason is that analysts who compute $\bar{u}_m$ and b_m in the standard way do not properly account for the informative prior distribution for $\rho_{YZ|X}$. To illustrate this, let $m = \infty$ for simplicity. Given a particular value of $\rho_{YZ|X}$, the analyst's posterior distribution of interest is $f(Q|D, \rho_{YZ|X})$. Assuming that the posterior distribution of Q is approximately normal—as is usual in multiple imputation contexts—the analyst must estimate $Var(Q|D, \rho_{YZ|X})$. As in Rubin (1987), this variance can be decomposed into $E(Var(Q|D^*, \rho_{YZ|X})) + Var(E(Q|D^*, \rho_{YZ|X}))$, where D^* represents the completed data. The first component is $\bar{u}_\rho = \lim \sum_{l=1}^{m} u_\rho^{(l)}/m$ as $m \to \infty$, where $u_\rho^{(l)} = Var(Q|D^{(l)}, \rho_{YZ|X})$. The second component is $b_\rho = \lim \sum_{l=1}^{m} (q_\rho^{(l)} - \bar{q}_\rho)^2/(m-1)$, where $q_\rho^{(l)} = E(Q|D^{(l)}, \rho_{YZ|X})$ and $\bar{q}_\rho = \lim_{l=1}^{m} q_\rho^{(l)}/m$ as $m \to \infty$. In standard multiple imputation, however, T_m is not based on $\bar{u}_\rho$ and b_ρ; rather, it uses $\bar{u}_\infty = \lim \sum_{l=1}^{m} u^{(l)}/m$ as $m \to \infty$ and $b_\infty = \lim \sum_{l=1}^{m} (q^{(l)} - \bar{q}_m)^2/(m-1)$ as $m \to \infty$. These latter two quantities are computed without considering $\rho_{YZ|X}$, i.e., $u^{(l)} = Var(Q|D^{(l)})$ and $q^{(l)} = E(Q|D^{(l)})$. In general, $\bar{u}_\rho \neq \bar{u}_\infty$, and $b_\rho \neq b_\infty$, which results in bias in T_m. For a simple but instructive example, let $\rho_{YZ|X}$ be the estimand Q of interest. We have $\bar{q}_\infty = \rho_{YZ|X}$, so that point estimation from standard multiple imputation is correct and has no sampling error. However, except in trivial cases, $\bar{u}_\infty > 0$ and $b_\infty > 0$, so that $T_\infty \neq Var(\bar{q}_\infty|D, \rho_{YZ|X}) = 0$. In contrast, $\bar{u}_\rho = b_\rho = 0$, since $E(Var(\rho_{YZ|X}|D^{(l)}, \rho_{YZ|X}) = 0$ and $Var(E(\rho_{YZ|X}|D^{(l)}, \rho_{YZ|X}) = 0$. We note that when $\bar{u}_\infty \approx \bar{u}_\rho$ and $b_\infty \approx b_\rho$, T_m should be an approximately valid estimate of variance.

Table 1 suggests that the relative magnitude of the bias in T_m increases with the influence of $\rho_{YZ|X}$ on the estimand. Among the six estimands, $\beta_{YZ|X}$ is most strongly affected by $\rho_{YZ|X}$, and its associated T_m is positively biased by a factor of 8.0 (.0056/.0007). Next in the order are $\beta_{YX|Z}$ and $\beta_{XY|Z}$ since, with $\rho_{XY} = .3$ and $\rho_{XZ} = .8$, the $\rho_{XY|Z}$ changes more dramatically with $\rho_{YZ|X}$ than $\rho_{XZ|Y}$ does. For these quantities, the associated T_m are positively biased by factors of about 2.5. Next is $\beta_{XZ|Y}$, which has a minor reliance on $\rho_{YZ|X}$ and only small bias in T_m. Finally, μ_Y and μ_X are independent of $\rho_{YZ|X}$, and their associated T_m are approximately unbiased.

To examine the nature of the bias further, we repeated the simulation under different scenarios. We first set $\rho_{YZ|X} = 0$ to represent conditional independence between Y and Z, which is implicitly assumed in many data fusion applications (D'Orazio, Di Zio, and Scanu (2006)). For $\beta_{YZ|X}$, the positive bias in T_m jumps to a factor of 20; for $\beta_{YX|Z}$ and $\beta_{XY|Z}$, the factors are 2.8 and 2.2, respectively. We then set $\rho_{YZ|X} = 0.9$ to represent a high partial correlation between Y and Z. For the three regression coefficients, T_m continued to have positive bias, but

the magnitudes were greatly decreased. Bias factors for the three regression coefficients ranged from roughly 1.1 to 1.4. Thus, as the strength of $\rho_{YZ|X}$ increases, it appears that there is reduced potential for bias in T_m.

Finally, to illustrate that these bias trends can be present even in seemingly innocuous scenarios, we performed a new simulation assuming joint independence of X, Y, and Z. We again used steps 2.1 – 2.5 for multiple imputation with $m = 10$. For $\beta_{YZ|X}$, T_m was positively biased by a factor of about 20; for the other estimands, T_m was approximately unbiased. These results are in accordance with the trends in Table 1. The bias factor was large for $\beta_{YZ|X}$, which is depends heavily on $\rho_{YZ|X}$. The bias factors were essentially zero for the other estimands, which do not depend on $\rho_{YZ|X}$ because of independence. We also repeated this simulation using $\rho_{YZ|X} = 0.9$. All bias factors were close to one. These results confirm what was seen previously: the relative bias in T_m decreases with the strength of $\rho_{YZ|X}$.

3. Bayesian Finite Population Imputation

An alternative would be to replace the multiple imputation approximation with an exact Bayesian inference, as in Gilula, McCulloch, and Rossi (2006). Exact Bayesian inferences properly account for information in prior distributions. However, some agencies prefer the simplicity of combining point estimates to deriving posterior distributions in the presence of missing data, particularly when the integrated data could be used for a variety of analyses. To facilitate these preferences, we approximate the exact Bayesian inference by treating data fusion inference as a problem in Bayesian finite population inference (Gelman et al. (2004, Chap. 7)). As we will see, this enables derivation of simple combining rules that can enable valid inferences.

Suppose that the n records in the concatenated data, D, are a subset of a much larger population, P, of N records. Further, suppose that a finite population analogue of the parameter of interest can be described. For example, for the coefficients in the regression of Y on (X, Z), the corresponding finite population quantity is $(W'W)^{-1}W'Y$, where W is an $N \times p + 1$ matrix containing a vector of ones and the values of the p variables in (X, Z). Finite population analogues exist for many common estimands, including summary statistics and logistic regression coefficients. Let Q generically represent the finite population representation of the quantity of interest.

Following the logic of Bayesian finite population inference, the analyst imputes the missing values for the $N - n$ records not in D, as well as any missing values in D, e.g., Y_2 and Z_1. The result is an entire completed population, $P^{(l)}$. The analyst then computes the value of Q in $P^{(l)}$; call this $Q^{(l)}$. For example, if Q is the population mean of Y, the analyst takes the sum of the n observed values of

Y_1 and the $N-n$ imputed values of Y. Each $Q^{(l)}$ is a draw from the posterior distribution of Q under the posited models for (X, Y, Z), so that the analyst can generate many $Q^{(l)}$ to summarize the posterior distribution of Q. When the posterior distribution of Q given $Var(Q|D)$ is normal, the analyst can generate a modest number of draws, say $m = 10$ draws, and use $Q - \bar{Q} \sim t_{m-1}(0, (1 + 1/m)W_m)$, where $\bar{Q} = \sum Q^{(l)}/m$ and $W_m = \sum(Q^{(l)} - \bar{Q})^2/(m-1)$.

For many data fusion contexts, D_1 and D_2 are not probability samples from a well-defined target population. In such cases, the agency can generate a hypothetical population on which to implement Bayesian finite population inference. The agency should set N much larger than n to minimize the impact of finite population correction factors on variance estimation. This process proceeds as follows. First, the agency generates $Y_2^{(l)}$ and $Z_1^{(l)}$ using the data fusion models and assumed distribution for $f(Y, Z|X)$. Second, the agency generates X for the $N - n$ records not in D using a model for X, which can be estimated from the marginal distribution of (X_1, X_2). Alternatively, the agency could use a Bayesian bootstrap (Rubin (1981)) from the completed data. Third, the agency generates values of Y for the $N - n$ records using a model for $f(Y|X)$, which can be estimated with D_1. Finally, the agency generates values of Z for the $N - n$ records using the implied data fusion model for $f(Z|X, Y)$. The result is one completed population, $P^{(l)}$.

For data fusion contexts in which D_1 and D_2 are random samples from a target population, the target population is a natural candidate for P. In general, the records not in D_1 and D_2 are missing all of (Y, Z) and most variables in X, except possibly for design variables like stratum or cluster indicators. The agency needs to impute plausible values for all the missing variables to generate m completed populations. The agency should take the design information into account when imputing; for example, include indicators for strata in imputation models (Reiter, Raghunathan, and Kinney (2006)), or use Bayesian bootstraps within strata when generating X.

We now illustrate the validity of the finite population imputation approach for data fusion inferences. We use the same simulation design as in Section 2. We again use the data fusion methods of Rässler (2003) to facilitate evaluation of the approach. However, rather than implement standard multiple imputation, we constructed hypothetical populations with 100,900 records, i.e., we repeatedly generated 100,000 additional records, as follows.

3.1 Complete D using the Steps 2.1 – 2.5 from Section 2, assuming $\rho_{YZ|X} = 0.2795$, to obtain $D^{(l)}$.

3.2 Simulate X for the 100,000 records excluded from D by drawing from the posterior predictive distribution, $f(X|X_1, X_2)$, based on noninformative prior

Table 2. Illustration of valid inferences when using finite population imputation for data fusion.

Estimand	var($\bar{q}_{10}$)	Avg. W_{10}	95% CI Cov.
$\beta_{YZ\|X}$	0.0005	0.0005	94.4%
$\beta_{YX\|Z}$	0.0025	0.0025	94.2%
$\beta_{XY\|Z}$	0.0004	0.0004	95.7%
$\beta_{XZ\|Y}$	0.0009	0.0009	95.3%
μ_Y	0.0023	0.0022	94.9%
μ_Z	0.0017	0.0018	96.2%

distributions on all parameters (which include the mean and variance of X in this simulation). Let $X_{exc}^{(l)}$ be the drawn values of X for these records. Let $W_{exc}^{(l)} = (1, X_{exc}^{(l)})$.

3.3 Simulate Y for the 100,000 records by drawing from the posterior predictive distribution, $f(Y|D_1, W_{exc}^{(l)})$, based on noninformative prior distributions; this is estimated using D_1. The parameters are the same as those drawn in Step 2.1. Let $Y_{exc}^{(l)}$ be the drawn values of Y for these records.

3.4 Simulate Z for the 100,000 records as in Step 2.5 of Section 2. Specifically, draw from $N(W_{exc}^{(l)}\beta_{Z|X}^{(l)} + (Y_{exc}^{(l)} - W_{exc}^{(l)}\beta_{Y|X}^{(l)})\sigma_{YZ|X}^{(l)}/\sigma_{Y|X}^{(l)2}, \sigma_{Z|X}^{(l)2} - \sigma_{YZ|X}^{(l)2}/\sigma_{Y|X}^{(l)2})$. Here the parameters are the same as those used in Step 2.5.

We repeated Steps 3.1 – 3.4 $m = 10$ times, so that $Q - \bar{Q} \sim t_9(0, 1.1W_{10})$. Similar steps can be implemented with other imputation models and data fusion techniques based on matching, as we discuss in Section 5.

Table 2 summarizes the properties of the inferences for 1,000 independent runs of the simulation design that produced the results in Table 1. The averages of $\bar{q}_{10}$ are within simulation error of the true values and so are not reported. For all estimands, W_{10} is approximately unbiased, and coverage rates are near the nominal 95% level. The same qualitative results were obtained for other simulation designs discussed in Section 2. Thus, the Bayesian finite population imputation approach avoids the large biases that can arise when using the standard multiple imputation variance estimator. Yet, it retains a desirable feature of multiple imputation: straightforward estimation of uncertainty for a variety of estimands by combining point estimates across datasets.

4. Confidential Data Fusion

When two or more agencies coordinate a data fusion, the agencies may not be willing to share some of their data values with each other. For example, two national statistical institutes may have collected their data under pledges of confidentiality that they are legally bound to keep. In this section, we describe

how multiple imputation can be used to preserve confidentiality while enabling valid data fusion inference. For further discussion of integration of confidential data in contexts other than data fusion, see Kohnen and Reiter (2009) and Reiter (2009).

Let $D_1 = (D_{1S}, D_{1C})$ and $D_2 = (D_{2S}, D_{2C})$, where the subscript S corresponds to values that are not confidential and can be shared between agencies without disclosure limitation, and subscript C corresponds to confidential values that require disclosure limitation methods prior to sharing. To begin the procedure, each agency generates new data for its D_{iC}, where $i = 1$ or $i = 2$, by simulating replacement values from the posterior predictive distribution, $f(D_{iC}|D_i)$. The posterior distributions should respect any mechanisms used to select the values to synthesize. For example, if all incomes above \$250,000 are to be synthesized, the synthesis models for income should condition on this fact; see Reiter (2003) for further discussion of this issue. We assume that each agency uses non-informative prior distributions. Each agency creates r partially synthetic datasets, $D_1^{(l)} = (D_{1S}, D_{1C}^{(l)})$ and $D_2^{(l)} = (D_{2S}, D_{2C}^{(l)})$, where $l = 1, \ldots, r$.

Each agency should evaluate the disclosure risks associated with sharing their partially synthetic copies. Approaches for evaluating identification and attribute disclosure risks with partially synthetic data are described by Reiter (2005), Reiter and Mitra (2009), and Drechsler and Reiter (2008).

After sharing the partially synthetic data, the agencies concatenate the datasets—this is done arbitrarily, since the replications are done independently—to create r versions of the complete data. Let $D_{syn}^{(l)} = (D_1^{(l)}, D_2^{(l)})$, where $l = 1, \ldots, r$. Each agency is now free to pursue its own analysis of the concatenated datasets. We note that agencies need not specify $f(Y, Z|X)$ to share each $D_{syn}^{(l)}$, so that they only have to create r datasets once for use with any $f(Y, Z|X)$ that they wish to consider.

To make inferences, the agency implements the Bayesian finite population imputation approach described in Section 3 for each $D_{syn}^{(l)}$. Recall that the agency simulates m draws of Q for a given dataset; hence, there are $M = mr$ total completed populations. Let $q^{(l,j)}$ be the jth draw of Q for $D_{syn}^{(l)}$; let $\bar{q}_m^{(l)} = \sum_{j=1}^m q^{(l,j)}/m$; and, let $w_m^{(l)} = \sum_{j=1}^m (q^{(l,j)} - \bar{q}_m^{(l)})^2/(m-1)$. The following quantities are then needed for inferences:

$$\bar{q}_M = \sum_{l=1}^r \frac{1}{r} \bar{q}_m^{(l)} \tag{4.1}$$

$$\bar{w}_M = \sum_{l=1}^r \frac{1}{r} \bar{w}_m^{(l)} \tag{4.2}$$

$$b_M = \sum_{l=1}^r \frac{(\bar{q}_m^{(l)} - \bar{q}_M)^2}{r-1}. \tag{4.3}$$

Table 3. Illustration of valid inferences when using finite population imputation for confidential data fusion ($r = 5, m = 10$).

Estimand	var($\bar{q}_{50}$)	Avg. T_{50}	95% CI Cov.
$\beta_{YZ\|X}$	0.0006	0.0006	93.2%
$\beta_{YX\|Z}$	0.0024	0.0025	95.1%
$\beta_{XY\|Z}$	0.0004	0.0004	95.3%
$\beta_{XZ\|Y}$	0.0011	0.0010	93.2%
μ_Y	0.0021	0.0020	94.1%
μ_Z	0.0019	0.0020	95.1%

The analyst uses $\bar{q}_M$ as the point estimate of Q and $T_M = \bar{w}_M + b_M/r$ as the variance estimate. For inference, the analyst uses a t-distribution, $Q - \bar{q}_M \sim t_{v_M}(0, T_M)$, with degrees of freedom given by

$$v_M = (r-1)(1 + \frac{r\bar{w}_M}{b_M})^2. \tag{4.4}$$

Derivations of these inferential methods are presented in the appendix.

We illustrate the validity of these inferential methods using the simulation design of Section 3 and $m = 10$. We replace all of Z_2 with $r = 5$ partially synthetic datasets generated from $f(Z|D_2)$ before concatenating the files. Table 3 summarizes the properties of T_{50} and the interval estimation procedure for 1,000 simulation runs. The averages of $\bar{q}_{50}$ are within simulation error of the true values and so are not reported. The simulated averages of T_{50} are approximately unbiased for the corresponding variances of $\bar{q}_{50}$, and the coverage rates are approximately equal to the nominal 95% level. Hence, the inferential methods for confidential data fusion can enable valid estimation of sampling variances and intervals.

5. Concluding Remarks

The simulations used the true partial correlations to demonstrate clearly that standard multiple imputation combining rules do not result in valid variance estimates in multiple imputation for data fusion. In genuine settings, however, the true partial correlation is unknown. To account for this uncertainty—which often is larger than the sampling and imputation variability—agencies can follow one of two general approaches. This first is akin to the approach described in Rubin (1986): select a modest number of representative and scientifically meaningful values of the partial correlations, run the Bayesian finite population imputation procedure for each specification to get inferences, and interpret the set of inferences as a sensitivity analysis. The second approach is more Bayesian in spirit: specify and repeatedly sample from prior distributions for the partial

correlations, run the Bayesian finite population imputation procedure for each specification with $m \geq 1$ to obtain at least one draw of the quantity of interest per specification, and mix all the draws. The mixed draws approximate the posterior distribution of the quantity of interest.

The Bayesian finite population imputation approach relies on generating $(X_{exc}, Y_{exc}, Z_{exc})$ for many more records than are in the concatenated data D. The resulting inferences will be sensitive to the choice of imputation models. In contrast, standard multiple imputation for data fusion requires imputation only for the missing Y_2 and Z_1, so that D contains comparatively larger fractions of observed values than are present in the completed populations from Bayesian finite population imputation. Thus, when comparing the Bayesian finite imputation approach to standard multiple imputation for data fusion, it is legitimate to ask the question: are potentially correct variance estimates worth the extra reliance on the imputations? The answers to this question depend on the type of analysis, as we now discuss.

For analyses involving variables that are marginally or jointly observed in just one of the datasets, e.g., the mean of Z or the regression of Z on X, analysts can avoid reliance on the imputation models by using only the relevant observed data, e.g., use only D_2 to estimate the regression of Z_2 on X_2. If the analyst instead estimates such quantities using the fused data, the point estimates for standard multiple imputation and Bayesian finite population imputation will be the same in expectation when the imputations are from correct models. All bets are off when the imputations are not from correct models: the mean squared error for either approach could dominate depending on how implausible the imputations are, although one generally expects the standard multiple imputation approach to be the lesser affected. This suggests that agencies disseminating fused data should include indicators of each record's data source, so as to enable secondary analysts to utilize single-source estimates either directly or to check the reasonableness of point estimates from the fused data.

Typically, the main point of data fusion is to estimate quantities involving relationships from the concatenated data. For such estimands, all the information in the fused data comes from the agency's joint imputation model. Unlike multiple imputation for missing data, in the fusion context there is essentially no observed information to anchor estimates of these associations should the imputation model be incorrect. Accordingly, for parameters involving variables from both agencies' datasets, point estimates from both standard multiple imputation and Bayesian finite population imputation are fully sensitive to the joint model specification.

To reduce reliance on parametric models, agencies can perform data fusion using predictive mean matching (Little (1988)), as suggested by Rubin (1986).

This approach substitutes observed values from Y_1 for missing Y_2 and observed values from Z_2 for missing Z_1, where substitutions are selected based on predictions from the agency's joint model for $(Y, Z|X)$; see Rubin (1986) for details. Data fusion by predictive mean matching does not immunize the standard multiple imputation variance estimator from the potential biases demonstrated in Section 2, which were evident even though the true models were used for imputation.

The Bayesian finite population imputation approach also can be adapted to work with predictive mean matching. In particular, the agency can construct completed populations by using (i) the matching methods of Rubin (1986) to complete D, (ii) a Bayesian bootstrap on the completed D to generate X_{exc}, (iii) predictive mean matching to generate Y_{exc} given X_{exc} based on a regression of Y_1 on X_1, and (iv) Rubin's (1986) predictive mean matching method to generate Z_{exc} given (X_{exc}, Y_{exc}). Semi-parametric methods like predictive mean matching are especially appealing for Bayesian finite population imputation, since the validity of the results depends on large amounts of simulated data.

Turning to confidential data fusion, agencies have to decide on r, the number of first-stage datasets. The choice of r involves trade offs between inferential accuracy, disclosure risks, and computational convenience: relatively large values of r result in smaller variances, greater disclosure risks, and greater computation. When only modest amounts of data, e.g., 25% or less, need to be synthesized for adequate protection, agencies can create small numbers of synthetic first-stage datasets, e.g., set $r = 3$, since little efficiency gains are expected as r increases. When large amounts of data need to be synthesized to protect confidentiality, agencies should make r as large as they are willing to bear, since efficiency gains can be substantial. Agencies can reduce computational burdens by using parallel computing to generate the first-stage partially synthetic datasets, as well as to create the hypothetical populations after sharing.

Confidential data fusion requires the agencies to synthesize their own data in ways that preserve salient features of the distributions yet protect confidentiality. Practically, this means that some confidential data fusion inferences are degraded compared to those based on the original data, since it is impossible to preserve all features of the original data unless the agencies share them outright—which would not protect confidentiality. Such degradations are arguably unavoidable when agencies seek to share confidential data in the fusion context. In a sense, however, the most appropriate comparison of data fusion based on synthetic data is not with data fusion based on the original data, as the latter is not possible with confidential data. Rather, it is with data fusion based on otherwise altered data—for which currently there are no principled methods of obtaining valid inferences—or perhaps with no data fusion at all. The extra step of synthesizing part of their data is the price agencies have to pay to protect confidentiality.

For any approach to data fusion, the characteristics of the assumed distribution and external information limit the range of admissible specifications for the unknown $f(Y, Z|X)$. In confidential data fusion, each synthetic dataset could admit different ranges. In such cases, one approach is to let the admissible range contain only those specifications that are coherent with all shared $D_{syn}^{(l)}$. A second approach is to perform inference for a given specification of $f(Y, Z|X)$ using only those $D_{syn}^{(l)}$ that are coherent with $f(Y, Z|X)$. Evaluating the trade-offs of these two approaches is an area for future research.

Acknowledgements

This research was supported by a grant from the National Science Foundation (NSF-MMS-0751671). The author thanks Christine Kohnen for helpful discussions about confidential data sharing.

Appendix: Derivation of Inferences for Confidential Data Fusion

The analyst of the concatenated partially synthetic datasets seeks to estimate $f(Q|D_{syn})$, where $D_{syn} = (D_{syn}^{(1)}, \ldots, D_{syn}^{(r)})$. For each $D_{syn}^{(l)}$, let $Q_\infty^{(l)}$ and $W_\infty^{(l)}$ be the point estimate of Q and its posterior variance that would be computed with $m = \infty$ draws from $D_{syn}^{(l)}$. Let $\bar{Q}_r = (1/r)\sum_{l=1}^r Q_\infty^{(l)}$, and let $\bar{Q}_\infty = \lim \bar{Q}_r$ and $B = \lim \sum_{l=1}^r (Q^{(l)} - \bar{Q}_\infty)^2/(r-1)$ both as $r \to \infty$. Let $Q^* = \{Q_\infty^{(l)} : l = 1, \ldots, r\}$, and let $W^* = \{W_\infty^{(l)} : l = 1, \ldots, r\}$. Then, $f(Q|D_{syn})$ can be written as

$$f(Q|D_{syn}) = \int f(Q|D, Q^*, B, W^*, D_{syn}) f(D|Q^*, B, W^*, D_{syn}) \times f(Q^*|B, W^*, D_{syn}) f(B, W^*|D_{syn}) dD dQ^* dB dW^*. \quad \text{(A.1)}$$

As in other applications of synthetic data, we find each component of this integral by assuming that the analyst's distributions are identical to those used for creating D_{syn}. We also assume that the sample sizes are large enough to permit normal approximations for these distributions. Thus, we require only the first two moments for each distribution, which can be derived using standard large sample Bayesian arguments. Diffuse priors are assumed for all parameters.

To begin, given D, the synthetic data are irrelevant for inferences, so that

$$f(Q|D, Q^*, B, W^*, D_{syn}) = N(Q_{obs}, V_{obs}). \quad \text{(A.2)}$$

Here, Q_{obs} and V_{obs} are the mean and variance of the posterior distribution of Q that would be obtained by performing the finite population imputation procedure with the original data D and an infinite number of draws.

Next, since all we require for inference are the first two moments, it is sufficient to consider $f(D|Q^*, B, W^*, D_{syn}) = f(Q_{obs}, V_{obs}|Q^*, B, W^*, D_{syn})$. We

presume the sampling distributions, $(Q_\infty^{(l)}|Q_{obs}, B) \sim N(Q_{obs}, B)$, for all l. This is reasonable when the replacement values are generated from predictive distributions based on the original data. With noninformative prior distributions, it follows that

$$f(Q_{obs}|V_{obs}, Q^*, B, W^*, D_{syn}) = N(\bar{Q}_r, \frac{B}{r}). \tag{A.3}$$

We also presume that $V_{obs} \approx W_\infty^{(l)}$ for any l. By extension, this implies $W_\infty^{(l)} \approx \bar{W}_\infty$ for any l. These assumptions are akin to those made about the complete-data variance in standard multiple imputation. Presuming $V_{obs} \approx W_\infty^{(l)}$ is reasonable here since these are complete-data variance estimators computed in the same way, i.e., the analyst uses the same Bayesian finite population imputation scheme for each $D^{(l)}$ that he or she would use for D, and the variability in posterior variances tends to be smaller than the variability in posterior means (Rubin (1987, p.89)). Hence, we set $V_{obs} = \bar{W}_\infty$.

We next consider $f(Q_\infty^{(l)}|B, W^*, D_{syn})$. In each $D_{syn}^{(l)}$, each draw of Q from the Bayesian finite population imputation procedure is an estimate of $Q_\infty^{(l)}$. Hence, we have

$$f(Q_\infty^{(l)}|B, W^*, D_{syn}^{(l)}) \sim N(\bar{q}_m^{(l)}, \frac{1}{m} W_\infty^{(l)}), \tag{A.4}$$

where $\bar{q}_m^{(l)}$ is the average of the m draws of Q computed with $D^{(l)}$. As a result, we have

$$f(\bar{Q}_r|B, W^*, D_{syn}) = N(\bar{q}_M, \frac{1}{rm}\bar{W}_\infty). \tag{A.5}$$

Thus, given B and W^*, from (A.2), (A.3), and (A.5), we have

$$f(Q|B, W^*, D_{syn}) = N(\bar{q}_M, \bar{W}_\infty + \frac{B}{r} + \frac{1}{rm}\bar{W}_\infty). \tag{A.6}$$

We now turn to the distributions of the remaining variance components. For $f(\bar{W}_\infty|D_{syn})$, from the posterior normality of Q we have

$$\frac{(m-1)w_m^{(l)}}{W_\infty^{(l)}}\Big| D_{syn}^{(l)} \sim \chi^2_{m-1}, \tag{A.7}$$

so that, assuming $W_\infty^{(l)} = W_\infty^{(j)}$ for all (l, j), we have

$$\frac{r(m-1)\bar{w}_M}{\bar{W}_\infty}\Big| D_{syn} \sim \chi^2_{r(m-1)}. \tag{A.8}$$

For $f(B|W^*, D_{syn})$, we apply Bayesian analysis of variance to (A.3) and (A.4), so that

$$\frac{(r-1)b_M}{B + \bar{W}_\infty/m}\Big| D_{syn} \sim \chi^2_{r-1}. \tag{A.9}$$

We now need to average (A.6) over the distributions in (A.8) and (A.9). As an approximation to this integral, for large m and r we can substitute the approximate expectations for $\bar{W}_\infty$ and B, namely $\bar{w}_M$ and b_M, into (A.6), so that

$$f(Q|D_{syn}) \approx N(\bar{q}, \bar{w}_M + \frac{b_M}{r}). \tag{A.10}$$

For data fusion contexts with large M and modest r, we use a t-distribution with degrees of freedom given by (4.4). This degrees of freedom can be derived by matching the first and second moments of $T_M/(\bar{W}_\infty + B/r + (1/rm)\bar{W}_\infty)$ to a chi-squared distribution with v_M degrees of freedom.

References

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases*, (Edited by J. Domingo-Ferrer and V. Torra), 290-297. Springer-Verlag: New York.

D'Orazio, M., Di Zio, M. and Scanu, M. (2002). Statistical matching and official statistics. *Rivista di Statistica Ufficiale* **1**, 5-24.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley, New York.

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases* (LNCS 5262), (Edited by J. Domingo-Ferrer and Y. Saygin), 227-238. Springer-Verlag: New York.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall, London.

Gilula, Z., McCulloch, R. E. and Rossi, P. E. (2006). A direct approach to data fusion. *J. Marketing Res.* **43**, 73-83.

Kadane, J. (2001). Some statistical problems in merging datasets. *J. Official Statist.* **17**, 423-433.

Kamakura, W. A. and Wedel, M. (1997). Statistical data fusion for cross-tabulation. *J. Marketing Res.* **34**, 485-498.

Kohnen, C. N. and Reiter, J. P. (2009). Multiple imputation for combining confidential data owned by two agencies. *J. Roy. Statist. Soc. Ser. A* **172**, 511-528.

Little, R. J. A. (1988). Missing data adjustments in large surveys. *J. Bus. Econom. Statist.* **6**, 287-296.

Little, R. J. A. (1993). Statistical analysis of masked data. *J. Official Statist.* **9**, 407-426.

Moriarity, C. and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *J. Official Statist.* **17**, 407-422.

Moriarity, C. and Scheuren, F. (2003). A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **21**, 65–73.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Lecture Notes in Statistics 168, Springer: New York.

Rässler, S. (2003). A non-iterative Bayesian approach to statistical matching. *Statist. Neerlandica* **57**, 58-74.

Rässler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austral. J. Statist.* **33**, 153-171.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181-189.

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

Reiter, J. P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *Internat. Statist. Rev.* **77**, 179-195.

Reiter, J. P. and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statist. Sinica* **20**, 405-422.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *J. Privacy and Confidentiality* **1**, 99-110.

Reiter, J. P., Raghunathan, T. E. and Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology* **32**, 143-150.

Rodgers, W. L. (1984). An evaluation of statistical matching. *J. Bus. Econom. Statist.* **2**, 91–102.

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9**, 130-134.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **4**, 87-94.

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

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *J. Official Statist.* **9**, 462-468.

van der Putten, P., Kok, J. N. and Gupta, A. (2002). Why the information explosion can be bad for data mining, and how data fusion provides a way out. In *Proceedings of the Second SIAM International Conference on Data Mining*, (Edited by R. L. Grossman, J. Han, V. Kumar, H. Mannila, and R. Motwant), SIAM: Arlington, VA.

Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251.

E-mail: jerry@stat.duke.edu

(Received June 2010; accepted March 2011)