IAB Discussion Paper

Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung

No. 20/2007

Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality

J.P. Reiter, Jörg Drechsler

Bundesagentur für Arbeit

Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality

J.P. Reiter (Institute of Statistics and Decision Sciences, Box 90251,
Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu)
Jörg Drechsler (IAB)

Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality

J. P. Reiter* and Jörg Drechsler†

Abstract

To protect the confidentiality of survey respondents' identities and sensitive attributes, statistical agencies can release data in which confidential values are replaced with multiple imputations. These are called synthetic data. We propose a two-stage approach to generating synthetic data that enables agencies to release different numbers of imputations for different variables. Generation in two stages can reduce computational burdens, decrease disclosure risk, and increase inferential accuracy relative to generation in one stage. We present methods for obtaining inferences from such data. We describe the application of two stage synthesis to creating a public use file for a German business database.

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

1 INTRODUCTION

Many national statistical agencies, survey organizations, and researchers—henceforth all called agencies—disseminate microdata, i.e. data on individual units, in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases and improvements in record linkage technologies have increased the risk of disclosure to the point where most agencies alter microdata before release (Reiter, 2004a). For example, agencies globally recode variables, such as releasing ages in five year intervals or top-coding incomes above 100,000 as "100,000 or more" (Willenborg and de Waal, 2001); they swap data values for randomly selected units (Dalenius and Reiss, 1982); or, they add random noise to continuous data values (Fuller, 1993). When applied with high intensity, these strategies reduce the utility of the released data, making some analyses impossible and severely distorting the results of others. They also complicate analyses for users. To analyze perturbed data properly, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These can be difficult to use for

*Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu

†Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany. E-mail: joerg.drechsler@iab.de

non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach to disseminating public use data was suggested by Rubin (1993): release multiply-imputed, synthetic data sets. Specifically, he proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these data sets to the public. A related approach was suggested by Fienberg (1994). These are called *fully synthetic* data sets. Releasing fully synthetic data can protect confidentiality, since identification of the sampled units and their sensitive data is very difficult when the values in the released data are not actual, collected values. Furthermore, with appropriate synthetic data generation and the inferential methods developed by Raghunathan *et al.* (2003) and Reiter (2005c), users can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of fully synthetic data are described by Rubin (1993), Little (1993), Fienberg *et al.* (1998), Raghunathan *et al.* (2003), Abowd and Lane (2004), and Reiter (2002, 2005b).

Some agencies have adopted a variant of Rubin's original approach, suggested by Little (1993): release data sets 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* data sets. For example, the U.S. Federal Reserve Board protects data in the 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, 2004) protect data in longitudinal, linked data sets by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values. Liu and Little (2002) and Little *et al.* (2004) present a general algorithm, named SMIKe, for simulating multiple values of key identifiers for selected units. Partially synthetic, public use data products are in the development stage in the U.S. for the Survey of Income and Program Participation, the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey group quarters data.

Partially synthetic approaches are appealing because they promise to 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. Valid inferences from partially synthetic data sets can be obtained using the methods developed by Reiter (2003, 2005c), whose rules for combining point and variance estimates differ from those of Rubin (1987) and also from those of Raghunathan *et al.* (2003). Methods for handling missing data simultaneously with partially synthetic data are developed in Reiter (2004b). Other illustrations of partially synthetic data include Reiter (2005d) and Mitra and Reiter (2006).

In this article, we present a two-stage approach to generating fully and partially synthetic data, in which agencies impute some variables only a few times and other variables many times. Two stage synthesis can have advantages over one-stage synthesis. In some settings, it reduces disclosure risks while increasing data usefulness. For example, agencies may want to release only a few imputed values of quasi-identifiers or sensitive variables, since intruders can use information from multiple data sets to refine guesses of the true values (Liu and Little, 2002; Reiter, 2005d; Mitra and Reiter, 2006), but they may want to release large numbers of imputations for other variables to drive down the variance introduced by imputation. In other settings, it reduces the labor needed to generate synthetic data. This is the case for the two-stage synthesis of the public release data for the German Institute for Employment Research (IAB) Establishment Panel, which is described in Section 2. A related approach, called nested multiple imputation (Shen, 2000; Harel and Schafer, 2003; Rubin, 2003b), has been used to reduce labor in the context of imputation for missing data.

The paper is organized as follows. Section 2 motivates the usefulness of two-stage synthetic data for reducing disclosure risks or decreasing agencies' labor. Section 3 derives methods for obtaining inferences from two-stage fully or partially synthetic data. These methods account for the correlations among estimates within the same first-stage nest. Section 4 illustrates the performance of these methods via simulation studies. Section 5 concludes with general remarks about two-stage synthetic data.

2 Motivation for two-stage synthesis

In this section, we first review evidence from the literature on the implications for disclosure risk and inferential accuracy of releasing many synthetic data sets. Two stage synthesis allows agencies to compromise on the risk-accuracy trade-off. We then describe the synthesis of data from the IAB Establishment Panel, for which one-stage synthesis demands too high labor cost.

2.1 Implications of releasing many synthetic data sets

From the perspective of the data analyst, there are benefits when agencies release a large number of multiply-imputed, synthetic data sets. The variability in point estimates computed with synthetic data decreases with the number of replicates. The reduction can be substantial when many values are synthesized. For example, Reiter (2002) finds roughly a 30% increase in the variance of survey-weighted estimates of population means when dropping from one hundred to five fully synthetic data sets. Reiter (2003) finds nearly a 100% increase in variance of regression coefficients when going from fifty to two partially synthetic data sets in which all of the dependent variable is replaced with imputations. Increasing the number of replicates also reduces the variability in estimators of variance. This variability can be large when many values are synthesized; in fact, for fully synthetic data, Reiter (2005b) finds that some variance estimators computed with ten fully synthetic data sets are so poor as to be essentially worthless. Those variance estimators have acceptable properties with one hundred replicates. We note that the incremental benefits become minimal as the number of replicates gets large.

From the perspective of the agency, there are risks to releasing a large number of multiply-imputed, synthetic data sets. Increasing the number of replicates provides more information for intruders to estimate the original data values. To illustrate this, we extend the partial synthesis done by Mitra and Reiter (2006), which used the 1987 U.S. Survey of Youth in Custody. The survey interviewed youths in juvenile facilities about their family background, previous criminal history, and drug and alcohol use. The sample contains 2,621 youths in 50 facilities. Mitra and Reiter (2006) consider facility membership to be potentially identifying information. Therefore, they generate new facility identifiers for all 2,621 youths. This is done by (i) fitting multinomial regressions of facility identifiers on the survey variables, (ii) drawing new values of parameters for the regressions and computing the resulting predicted probabilities for each youth, and (iii) simulating new identifiers from the multinomial distributions based on the predicted probabilities. To assess disclosure risk, they assumed that the intruder uses the mode of each youth's multiply-imputed facility as the best guess of the youth's actual facility. When no unique mode exists, they randomly select one value. We follow the same procedures for different numbers of synthetic data sets. With three replicates, approximately 17% of intruders' guesses are correct. With ten replicates, this increases to 20%. With fifty replicates, this increases to 24%. While perhaps not alarming, the increasing identification rates certainly would push agencies to minimize the number of imputations of facilities.

For fully synthetic data, there has been little work on the impacts on disclosure risk of releasing many replicates. In part, this is because identification disclosure risks are low for fully synthetic data. Each data set contains different samples of records, and all survey variables are synthesized. However, the risks are not zero. When the imputation models are highly detailed, the imputations could reproduce combinations of quasi-identifiers for real records. Intruders might interpret this to mean that real-data records with those characteristics were in the original sample, which could result in identification disclosures if some of those records are unique in the population. This risk could be magnified when releasing multiple synthetic data sets, because (i) there are several opportunities to impute such records, and (ii) there could be repetitions of realistic synthetic records, which might strengthen the intruder's confidence that a similar real record was in the original data.

Ideally, when considering the release of public use data, the agency balances confidentiality protection and inferential accuracy; see, for example, Duncan *et al.* (2001), Skinner and Elliot (2002), Reiter (2005a), Gomatam *et al.* (2005), and Karr *et al.* (2006). Confidentiality concerns often trump accuracy concerns. With one stage synthetic data, favoring confidentiality over accuracy could lead agencies to release few replicates. With two stage synthesis, agencies can compromise on the risk-accuracy trade-off. Agencies can release few imputations of quasi-identifiers or other confidential variables to reduce disclosure risks, and release many imputations of other variables to enable analysts to improve precision for analyses involving those variables.

2.2 Synthesis of the IAB Establishment Panel

The IAB Establishment Panel, conducted since 1993, contains detailed information about German firms' personnel structure, development, and policy. Considered one of most important business panels in Germany, there is high demand for access to these data from external researchers. Because of the sensitive nature of the data, researchers desiring direct access to the data have to work on site at the IAB. Alternatively, researchers can submit code for statistical analyses to the IAB research data center, whose staff run the code on the data and send the results to the researchers. To help researchers develop code, the IAB provides remote access to a publicly available "dummy data set" with the same structure as the Establishment Panel. The dummy data set comprises random numbers generated without attempts to preserve the distributional properties of the variables in the Establishment Panel data. For all analyses done with the genuine data, researchers can publicize their analyses only after IAB staff check for potential violations of confidentiality.

Releasing public use files of the Establishment Panel would allow more researchers to access the data with fewer burdens, stimulating research on German business data. It also would free up staff time from running code and conducting confidentiality checks. Because there are so many sensitive variables in the data set, standard disclosure limitation methods like swapping or microaggregation would have to be applied with high intensity, which would severely compromise the utility of the released data. Therefore, the IAB decided to develop synthetic data, specifically (at this stage) fully synthetic data.

Each synthetic data set comprises establishments sampled from the sampling frame for the Establishment Panel. We sample records according to the design of the Establishment Panel—stratifying by region, establishment size, and industry—to take advantage of the efficiency gained by the original stratification. Let X be the variables corresponding to the stratum indicators.

We impute values of the Establishment Panel survey variables, Y_b, for all establishments in the synthetic data samples. These models are developed as follows. First, for all records in the original panel, we obtain establishment-level data, Y_a, from the German Social Security Data (GSSD). The GSSD contains information on individuals covered by social security, including data on their employer such as demographic characteristics and average wages of its employees. The employers are identified by the establishment identification numbers used in the Establishment Panel, which enables direct matching between the two data sources. Second, we build a statistical model relating Y_b to (X, Y_a) using the data from the original panel. Third, for each synthetic sample, we match the newly drawn establishments to the GSSD and append their values of Y_a to the synthetic data. Fourth, we simulate values of Y_b from $f(Y_b|X, Y_a)$, using the X and the appended values of Y_a for the new establishments. After the imputation, all variables in Y_a are deleted for confidentiality reasons. The result is a synthetic data set that mimics the structure of the Establishment Panel, comprising the stratification indicators X and the imputed survey variables Y_b.

Previous research has shown that releasing large numbers of fully synthetic data sets improves synthetic data inferences (Reiter, 2005b). The usual advice from multiple imputation for missing data—release five multiply-imputed data sets—tends not to work well for fully synthetic data because the fractions of "missing" information are large. Following Reiter (2005b), the IAB desired to generate and release one hundred fully synthetic data sets. However, doing so requires matching to the GSSD one hundred times and imputing Y_b for each matched sample. These are very labor intensive tasks. The matching has to be checked and corrected if necessary each time, and the matched data need to be transferred to different software platforms for the imputation of Y_b. Furthermore, each matched data file is re-configured manually to implement the imputation routines.

This led the IAB synthesis team to adopt a two stage approach to synthesis. We draw only ten synthetic samples, thus requiring only ten iterations of matching and data processing to obtain Y_a. For each sample, we impute Y_b another ten times, resulting in one hundred data sets. This two-stage method reduces the labor by a factor of ten while allowing us to release one hundred data sets containing information about Y_b as opposed to only ten. For more details about the imputation models in the synthesis, which are based on the sequential multivariate regression imputation strategy of Raghunathan *et al.* (2001), see Drechsler *et al.* (2007).

The ten sets of Y_b for each sample are correlated. Standard one-stage methods of inference do not account for this nested structure. Section 3 derives new methods of inference for two stage synthesis, both for fully and partially synthetic data. The methods are presented assuming all variables are released, but they apply when some variables are suppressed as in the synthesis of the Establishment Panel. The methods also assume for generality that (Y_a, Y_b) is known only for the sampled records.

3 Inferences with two-stage synthetic data

For a finite population of size N, let $I_l = 1$ if unit l is included in the survey, and $I_l = 0$ otherwise, where $l = 1, \ldots, N$. Let $I = (I_1, \ldots, I_N)$, and let the sample size $s = \sum I_l$. Let X be the $N \times d$ matrix of sampling design variables, e.g. stratum or cluster indicators or size measures. We assume that X is known approximately for the entire population, for example from census records or the sampling frame(s). Let Y be the $N \times p$ matrix of survey data for the population. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $s \times p$ sub-matrix of Y for all units with $I_l = 1$, where 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 R be an $N \times p$ matrix of indicators such that $R_{lk} = 1$ if the response for unit l to item k is recorded, and $R_{lk} = 0$ otherwise. The observed data is thus $D_{obs} = (X, Y_{obs}, I, R)$.

3.1 Fully synthetic data

Let Y_a be the values simulated in stage 1, and let Y_b be the values simulated in stage 2. The agency seeks to release fewer replications of Y_a than of Y_b, yet do so in a way that enables the analyst of the data to obtain valid inferences with standard complete data methods. To do so, the agency generates synthetic data sets in a three-step process. First, the agency fills in the unobserved values of Y_a by drawing values from $f(Y_a \mid D_{obs})$, creating a partially completed population. This is repeated independently m times to obtain $Y_a^{(i)}$, for $i = 1, \ldots, m$. Second, in each partially completed population defined by nest i, the agency generates the unobserved values of Y_b by drawing from $f(Y_b \mid D_{obs}, Y_a^{(i)})$, thus completing the rest of the population values. This is repeated independently r times for each nest to obtain $Y_b^{(i,j)}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, r$. The result is $M = mr$ completed populations, $P^{(i,j)} = (D_{obs}, Y_a^{(i)}, Y_b^{(i,j)})$, where $i = 1, \ldots, m$ and $j = 1, \ldots, r$. Third, the agency takes a simple random sample of size n_{syn} from each completed population $P^{(i,j)}$ to obtain $D^{(i,j)}$. These M samples, $D_{syn} = \{D^{(i,j)} : i = 1, \ldots, m; j = 1, \ldots, r\}$, are released to the public. Each released $D^{(i,j)}$ includes a label indicating its value of i, i.e. an indicator for its nest.

The agency can sample from the completed populations using designs other than simple random samples, for example the stratified sampling in the IAB Establishment Panel synthesis. When synthetic data are generated using complex samples, the analyst should account for the sampling design to obtain valid inferences, such as using survey-weighted estimates. One advantage of creating synthetic data by simple random sampling is that analysts need not deal with complex sampling designs; they can analyze the synthetic data as if they come from simple random samples.

The agency could simulate Y for all N units, thereby avoiding the release of any actual values of Y. In practice, it is not necessary to generate completed-data populations for constructing the $D^{(i,j)}$; the agency need only generate values of Y for units in the synthetic samples. The formulation of completing the population, then sampling from it, aids in deriving the methods for inference.

Let Q be the estimand of interest, such as a population mean or a regression coefficient. The analyst of synthetic data seeks $f(Q|D_{syn})$. The three-step process for creating D_{syn} suggests that

$$f(Q|D_{syn}) = \int f(Q|D_{obs}, P_{syn}, D_{syn}) f(D_{obs}|P_{syn}, D_{syn}) f(P_{syn}|D_{syn}) dD_{obs} dP_{syn}, \quad (1)$$

where $P_{syn} = \{P^{(i,j)} : i = 1, \ldots, m; j = 1, \ldots, r\}$. For all derivations in Section 3, we assume that the analyst's distributions are identical to those used by the agency 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 we derive using standard large sample Bayesian arguments. Diffuse priors are assumed for all parameters.

To begin, the synthetic data are irrelevant for inference about Q given the observed data, so that $f(Q|D_{obs}, P_{syn}, D_{syn}) = f(Q|D_{obs})$. We assume that

$$(Q|D_{obs}) \sim N(Q_{obs}, U_{obs}), \tag{2}$$

where Q_{obs} and U_{obs} are the estimates of the mean and variance computed from D_{obs} if it were released.

The D_{syn} is irrelevant given P_{syn}, so that $f(D_{obs}|P_{syn}, D_{syn}) = f(D_{obs}|P_{syn})$. Because inferences for Q depend only on Q_{obs} and U_{obs}, it is sufficient to determine $f(Q_{obs}, U_{obs}|P_{syn})$. Let $Q^{(i,j)}$ be the estimate of Q in population $P^{(i,j)}$. Let $\bar{Q}_r^{(i)} = \sum_j Q^{(i,j)}/r$, and $\bar{Q}_M = \sum_i \bar{Q}_r^{(i)}/m$. Let $B_M = \sum_i (\bar{Q}_r^{(i)} - \bar{Q}_M)^2/(m-1)$, and $W_r^{(i)} = \sum_j (Q^{(i,j)} - \bar{Q}_r^{(i)})^2/(r-1)$. We assume the following sampling distributions:

$$\begin{aligned} \left(\bar{Q}_\infty^{(i)}|D_{obs}, B_\infty\right) &\sim N(Q_{obs}, B_\infty) \qquad (3)\\ \left(Q^{(i,j)}|\bar{Q}_\infty^{(i)}, W_\infty^{(i)}\right) &\sim N(\bar{Q}_\infty^{(i)}, W_\infty^{(i)}) \qquad (4) \end{aligned}$$

where the $\bar{Q}_\infty^{(i)}$, the $W_\infty^{(i)}$, and B_∞ are the limits of the corresponding finite-sum quantities as $m \to \infty$ and $r \to \infty$. The process of repeatedly completing populations and estimating Q in this nested manner is equivalent to simulating the posterior distribution of Q. Hence, the $U_{obs} = B_\infty + \bar{W}_\infty$, where $\bar{W}_\infty = \lim \sum_i W_\infty^{(i)}/m$ as $m \to \infty$. From (2), (3), and (4), for finite m and r we have

$$\left(Q|P_{syn}, B_\infty, W_\infty^{(1)}, \ldots, W_\infty^{(m)}\right) \sim N(\bar{Q}_M, (1+1/m)B_\infty + (1+1/(mr))\bar{W}_\infty). \tag{5}$$

We also have

$$\begin{aligned} \left((m-1)B_M/(B_\infty + \bar{W}_\infty/r)|P_{syn}, \bar{W}_\infty\right) &\sim \chi^2_{m-1} \qquad (6)\\ \left((r-1)W_r^{(i)}/W_\infty^{(i)}|P_{syn}\right) &\sim \chi^2_{r-1}. \qquad (7) \end{aligned}$$

The posterior distribution of Q conditioning on P_{syn} alone is found by integrating (5) over the distributions in (6) and (7).

In general, releasing P_{syn} is impractical for agencies, as it could require releasing M data files of very large size N. We therefore take random samples of size n_{syn} from each population, i.e. the $D^{(i,j)}$. We require the distributions of $\bar{Q}_M, B_\infty$, and the $W_\infty^{(i)}$ conditional on D_{syn}. For all (i,j), let $q^{(i,j)}$ be the estimate of $Q^{(i,j)}$, and let $u^{(i,j)}$ be the estimate of the variance associated with $q^{(i,j)}$. The $q^{(i,j)}$ and $u^{(i,j)}$ are computed based on the design used to sample from $P^{(i,j)}$. Note that when $n_{syn} = N$, the $u^{(i,j)} = 0$. Let $\bar{q}_r^{(i)} = \sum_j q^{(i,j)}/r$, and $\bar{q}_M = \sum_i \bar{q}_r^{(i)}/m$. Let $b_M = \sum_i (\bar{q}_r^{(i)} - \bar{q}_M)^2/(m-1)$, and $w_r^{(i)} = \sum_j (q^{(i,j)} - \bar{q}_r^{(i)})^2/(r-1)$. Finally, let $\bar{u}_M = \sum_{i,j} u^{(i,j)}/(mr)$.

For n_{syn} large, we assume the sampling distribution of each $(q^{(i,j)}|P_{syn}^{(i)})$ is $N(Q^{(i,j)}, U^{(i)})$, where $U^{(i)}$ is an implied sampling variance. We further assume that the sampling variability in the $u^{(i,j)}$ is negligible, so that $u^{(i,j)} \approx U^{(i)}$. We also make the simplifying assumption that the variability in the $U^{(i)}$ across nests is small, so that $U^{(i)} \approx \sum U^{(i)}/m$. Thus, we have

$$(q^{(i,j)}|P_{syn}^{(i)}) \sim N(Q^{(i,j)}, \bar{u}_M). \tag{8}$$

Using the standard Bayesian arguments based on these sampling distributions, we have

$$\left(\bar{Q}^{(i)}|\bar{q}_r^{(i)}, \bar{u}_M\right) \sim N(\bar{q}_r^{(i)}, \bar{u}_M/r) \tag{9}$$

and

$$\left(\bar{Q}_M|D_{syn}\right) \sim N(\bar{q}_M, \bar{u}_M/(mr)). \tag{10}$$

To obtain the conditional distributions of B_∞ and the $W_\infty^{(i)}$, we use an analysis of variance setup. From (4) and (8), we have

$$\left(\frac{(r-1)w_r^{(i)}}{W_\infty^{(i)} + \bar{u}_M}|D_{syn}\right) \sim \chi^2_{r-1}. \tag{11}$$

From (3), (4), and (8), and making the simplifying assumption that the $W_\infty^{(i)} = \bar{W}_\infty$ for all i, we have

$$\begin{aligned} \left(\frac{(m-1)b_M}{B_\infty + \bar{W}_\infty/r + \bar{u}_M/r}|D_{syn}, \bar{W}_\infty\right) &\sim \chi^2_{m-1} & (12)\\ \left(\frac{m(r-1)\bar{w}_M}{\bar{W}_\infty + \bar{u}_M}|D_{syn}\right) &\sim \chi^2_{m(r-1)} & (13)\end{aligned}$$

where $\bar{w}_M = \sum_i w_r^{(i)}/m$.

To obtain the conditional distribution of Q given D_{syn}, we should integrate the distributions in (5), (6), and (7) with respect to the distributions of $\bar{Q}_M$, B_∞, and the $W_\infty^{(i)}$ in (10), (11), and (12). Although this integration can be carried out numerically, we desire a straightforward approximation that can be easily computed by analysts using D_{syn}. For large m and r, we can approximate $f(Q|D_{syn})$ by a normal distribution with mean $E(Q|D_{syn})$ and variance $Var(Q|D_{syn})$. Using (5) and (10), we have

$$E(Q|D_{syn}) = E[E(Q|\bar{Q}_M)|D_{syn}] = E(\bar{Q}_M|D_{syn}) = \bar{q}_M. \tag{14}$$

Similarly,

$$\begin{aligned} Var(Q|D_{syn}) &= E[Var(Q|P_{syn}, B_\infty, \bar{W}_\infty)|D_{syn}] + Var[E(Q|P_{syn}, B_\infty, \bar{W}_\infty)|D_{syn}] \\ &= (1+m^{-1})E(B_\infty|D_{syn}) + (1+1/(mr))E(\bar{W}_\infty|D_{syn}) + \bar{u}_M/(mr). \quad (15)\end{aligned}$$

Based on (12) and (13), we approximate the expectations in (15) as $E(\bar{W}_\infty|D_{syn}) \approx \bar{w}_M - \bar{u}_M$ and $E(B_\infty|D_{syn}) \approx b_M - \bar{w}_M/r$. Substituting these approximate expectations in (15), we obtain

$$\begin{aligned} Var(Q|D_{syn}) &\approx (1+m^{-1})(b_M - \bar{w}_M/r) + (1+1/(mr))(\bar{w}_M - \bar{u}_M) + \bar{u}_M/(mr) \\ &= (1+m^{-1})b_M + (1-1/r)\bar{w}_M - \bar{u}_M = T_f. \quad (16)\end{aligned}$$

For modest m and r, we obtain inferences by using a t-distribution, $(\bar{q}_M - Q) \sim t_{\nu_f}(0, T_f)$. The degrees of freedom, ν_f, equal

$$\nu_f = \left(\frac{((1+1/m)b_M)^2}{(m-1)T_f^2} + \frac{((1-1/r)\bar{w}_M)^2}{(m(r-1))T_f^2}\right)^{-1}.$$

The degrees of freedom is derived by matching the first two moments of $(\nu_f T_f)/(\bar{u}_M/(mr)+(1+1/m)B_\infty + (1+1/(mr))\bar{W}_\infty)$ to an inverse chi-squared distribution with ν_f degrees of freedom. The derivation is presented in the appendix.

It is possible that $T_f < 0$, particularly for small values of m and r. To adjust for this possibility, one approach is to use the conservative and always positive variance estimator,

$$T_f^* = T_f + \lambda \bar{u}_M, \tag{17}$$

where $\lambda = 1$ when $T_f \leq 0$ and $\lambda = 0$ when $T_f > 0$. Generally, negative values of T_f can be avoided by making n_{syn} or m and r large.

When $T_f < 0$, using the degrees of freedom ν_f is overly conservative, since T_f^* tend to be already conservative when $\lambda = 1$. To avoid excessively wide confidence intervals, one approach is to base inferences on normal distributions in this case. Equivalently, and for notational simplicity, use t-distributions with degrees of freedom ν_f^*, where

$$\nu_f^* = \nu_f + \lambda\infty. \tag{18}$$

The ν_f can be very small even when $T_f > 0$, which could result in excessively wide intervals. We evaluate a modification to ν_f when it is small in Section 4.

3.2 Partially synthetic data

We assume that $Y_{inc} = Y_{obs}$, i.e. there is no missing data. Methods for handling missing data and one stage of partial synthesis simultaneously are presented by Reiter (2004b).

The agency generates the partially synthetic data in two stages. Let $Y_a^{(i)}$ be the values imputed in the first stage in nest i, for $i = 1, \ldots, m$. Let $Y_b^{(i,j)}$ be the values imputed in the second stage in data set j in nest i, for $j = 1, \ldots, r$. Let Y_{nrep} be the values of Y_{obs} that are not replaced with synthetic data and hence are released as is. Let $Z_{a,l} = 1$ if unit l, for $l = 1, \ldots, s$, is selected to have any of its first-stage data replaced with synthetic values, and let $Z_{a,l} = 0$ for those units with all first-stage data left unchanged. Let $Z_{b,l}$ be defined similarly for the second-stage values. Let $Z = (Z_{a,1}, \ldots, Z_{a,s}, Z_{b,1}, \ldots, Z_{b,s})$.

To create the $Y_a^{(i)}$ for those records with $Z_{a,l} = 1$, first the agency draws from $f(Y_a \mid D_{obs}, Z)$, conditioning only on values not in Y_b. Second, in each nest, the agency generates the $Y_b^{(i,j)}$ for those records with $Z_{b,l} = 1$ by drawing from $f(Y_b^{(i,j)} \mid D_{obs}, Z, Y_a^{(i)})$. Each synthetic data set, $D^{(i,j)}$, comprises $(X, Y_a^{(i)}, Y_b^{(i,j)}, Y_{nrep}, I, Z)$. The entire collection of $M = mr$ data sets, $D_{syn} = \{D^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, r\}$, with labels indicating the nests, is released to the public.

To obtain inferences from nested partially synthetic data, we assume the analyst acts as if each $D^{(i,j)}$ is a sample according to the original design. We require the integral,

$$f(Q|D_{syn}) = \int f(Q|D_{obs}, D_{syn}) f(D_{obs}|D_{syn}) dD_{obs}. \tag{19}$$

Unlike in fully synthetic data, there is no intermediate step of completing populations. Let $q^{(i,j)}$, $\bar{q}_r^{(i)}$, $\bar{q}_M$, b_M, and the $w_r^{(i)}$ be defined as in the previous section. Define $\bar{q}_\infty^{(i)} = \lim \bar{q}_r^{(i)}$, $b_\infty = \lim b_M$, and $w_\infty^{(i)} = \lim w_r^{(i)}$ as $m \to \infty$ and $r \to \infty$.

With large samples, we assume again that $f(Q|D_{obs}) = N(Q_{obs}, U_{obs})$. We assume that the sampling distributions of the synthetic data point estimators are

$$\left(\bar{q}_{\infty}^{(i)}|D_{obs}, b_{\infty}\right) \sim N(Q_{obs}, b_{\infty}) \tag{20}$$

$$\left(q^{(i,j)}|D_{obs}, \bar{q}_{\infty}^{(i)}, w_{\infty}^{(i)}\right) \sim N(\bar{q}_{\infty}^{(i)}, w_{\infty}^{(i)}). \tag{21}$$

When coupled with (2) and diffuse priors on all parameters, (20) and (21) imply that

$$\left(Q|D_{syn}, b_{\infty}, w_{\infty}^{(1)}, \ldots, w_{\infty}^{(m)}\right) \sim N(\bar{q}_M, U_{obs} + b_{\infty}/m + \bar{w}_{\infty}/(mr)). \tag{22}$$

Since the Y_a and Y_b are simulated from their conditional distributions, each $u^{(i,j)}$ approximates U_{obs}. We assume that the $u^{(i,j)}$ have low variability, so that $u^{(i,j)} \approx \bar{u}_M \approx U_{obs}$.

The posterior distributions of b_{∞} and the $w_{\infty}^{(i)}$ are obtained from an analysis of variance setup. From (21), we have

$$\left(\frac{(r-1)w_r^{(i)}}{w_{\infty}^{(i)}}|D_{syn}\right) \sim \chi^2_{r-1}. \tag{23}$$

From (20), (21), and (23), and making the simplifying assumption that the $w_{\infty}^{(i)} = \bar{w}_{\infty}$ for all i, we have

$$\left(\frac{(m-1)b_M}{b_{\infty} + \bar{w}_{\infty}/r}|D_{syn}, \bar{w}_{\infty}\right) \sim \chi^2_{m-1} \tag{24}$$

$$\left(\frac{m(r-1)\bar{w}_M}{\bar{w}_{\infty}}|D_{syn}\right) \sim \chi^2_{m(r-1)}. \tag{25}$$

To obtain the conditional distribution of Q, we should integrate (22) over the distributions in (23) and (24). For large m and r, we can approximate this with a normal distribution, substituting the approximate expected values of b_{∞} and $\bar{w}_{\infty}$ into the variance in (22). For large m and r, this variance simplifies to $T_p = \bar{u}_M + b_M/m$, so that the approximate normal distribution is $(\bar{q}_M - Q) \sim N(0, T_p)$.

For small m and r, we can use a t-distribution for inferences, $(\bar{q}_M - Q) \sim t_{\nu_p}(0, T_p)$. The degrees of freedom $\nu_p = (m-1)(1 + m\bar{u}_M/b_M)^2$. The degrees of freedom is derived by matching the first two moments of $(\nu_p(\bar{u}_M + b_M/m))/(\bar{u}_M + b_{\infty}/m + \bar{w}_{\infty}/(mr))$ to an inverse chi-squared distribution with ν_p degrees of freedom. The derivation is presented in the appendix.

4 Illustrative simulations

In this section, we present results from simulation studies of the inferential methods for two-stage, fully synthetic data. The studies are designed to resemble the synthesis for the IAB Establishment Panel. They include evaluations of adjustments for negative variance estimates and small degrees of freedom. We do not present the results from simulation studies of two-stage partially synthetic data. Results from those studies indicated that the inferential methods outlined in Section 3.2 have good frequentist properties without any need for adjustments.

We generate a population of $N = 100,000$ records comprising five variables, $Y_1, \ldots, Y_5$. The (Y_1, Y_2) are drawn from a joint t-distribution with 20 degrees of freedom and a correlation of 0.5. The (Y_3, Y_4, Y_5) are drawn from the joint normal distribution $N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 1.5Y_1 + 1.5Y_2 \\ 2.5Y_1 + 2.5Y_2 \\ -3.0Y_1 - 3.0Y_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 30 & 15 & 15 \\ 15 & 30 & 15 \\ 15 & 15 & 30 \end{pmatrix}.$$

The observed data, D_{obs}, comprise the values of $(Y_1, \dots, Y_5)$ for a simple random sample of $s = 1,000$ records from this population. We assume that (Y_1, Y_2) are known for all N records and that (Y_3, Y_4, Y_5) are known only for the s sampled records. Using an analogy with the IAB Establishment Panel synthesis, the (Y_1, Y_2) are like variables found in the German Social Security Data; the (Y_3, Y_4, Y_5) are like variables only found in the Establishment Panel; and, concatenating all five variables for the s records is like matching the information from the GSSD for the Establishment Panel respondents. For simplicity, we do not incorporate stratification in the sampling.

We treat $Y_a = (Y_1, Y_2)$ as the first stage variables and $Y_b = (Y_3, Y_4, Y_5)$ as the second stage variables. For each synthetic data set $D^{(i,j)}$, where $i = 1, \dots, m$ and $j = 1, \dots, r$, we generate $Y_a^{(i)}$ by taking a random sample of $n_{syn} = 1,000$ records from the population and using their values of (Y_1, Y_2). We generate the $Y_b^{(i,j)}$ for these records by sampling from the posterior predictive distribution, $f(Y_3, Y_4, Y_5 | D_{obs}, Y_a^{(i)})$, with noninformative prior distributions on all parameters. That is, we draw $Y_3^{(i,j)}$ from the regression $f(Y_3 | D_{obs}, Y_a^{(i)})$, we draw $Y_4^{(i,j)}$ from the regression $f(Y_4 | D_{obs}, Y_a^{(i)}, Y_3^{(i,j)})$, and we draw $Y_5^{(i,j)}$ from the regression $f(Y_5 | D_{obs}, Y_a^{(i)}, Y_3^{(i,j)}, Y_4^{(i,j)})$. The released data comprise the mr copies of the $(Y_a^{(i)}, Y_b^{(i,j)})$. By including the imputations for the first stage variables in the released data, we deviate from the IAB Establishment Panel synthesis. However, this enables evaluations of inferences for relationships between variables imputed at different stages.

To evaluate the performance of the inferential methods, we estimate five quantities: the population mean of Y_3 ($\bar{Y}_3$), the regression coefficients of Y_1 (β_1) and of Y_5 (β_5) in a regression of Y_3 on all other variables, and the regression coefficients of Y_2 (α_2) and of Y_5 (α_5) in a regression of Y_1 on all other variables. We repeat the process of drawing D and generating synthetic data sets 5,000 times. For simplicity, we do not utilize the small finite population correction factors when computing the $u^{(i,j)}$.

Table 1 summarizes the results for several combinations of m and r. The averages of the $\bar{q}_M$ across the iterations are within simulation error of their corresponding population values; we do not report them in the table. For most estimands, the T_f are nearly unbiased for the $Var(\bar{q}_M)$. The T_f associated with α_2 and α_5 tend to have positive bias. For $m = r = 3$, the values of T_f are frequently negative. This results from high variability in b_M and $\bar{w}_M$, making them unstable estimates of B_∞ and $\bar{W}_\infty$. Negative variance estimates become less frequent as M increases, since the variability in b_M and $\bar{w}_M$ decreases. The always positive variance estimator T_f^* is, as expected, conservative.

The column labeled "95% CI Cov*" displays the percentages of the 5,000 synthetic 95% confidence intervals that cover their corresponding Q. The intervals are based on T_f^* and on t-distributions with ν_f^* defined in (18). For scenarios with low m and r, the procedure generally produces intervals with greater than nominal coverage rates. In part this is due to the conservative nature of T_f^*. It also results from small values of ν_f^*, sometimes less than one, that arise because of inadequacies in the approximations for modest m and r. To avoid using unrealistically small degrees of freedom, we construct the modified degrees of freedom,

$$\nu_f^{**} = \max\{(m-1), \nu_f^*\}. \tag{26}$$

As displayed in the column labeled "95% CI Cov**," these coverage rates are closer to 95%. We note that confidence intervals based on a normal distribution for all iterations led to consistently lower than nominal coverage rates.

m, r	Q	Var($\bar{q}_M$)	Avg. T_f	$\%T_f < 0$	Avg. T_f^*	95% CI Cov*	95% CI Cov**
$3, 3$	$\bar{Y}_3$	.0409	.0389	15.7	.0448	94.2	93.8
	β_1	.0537	.0533	12.3	.0587	98.0	95.9
	β_5	.00108	.00106	12.2	.00117	98.0	96.2
	α_2	.000766	.000850	24.8	.00109	97.6	96.3
	α_5	.0000121	.0000126	19.3	.0000151	97.8	95.7
$5, 5$	$\bar{Y}_3$	.0327	.0335	3.6	.0349	99.2	95.5
	β_1	.0458	.0471	1.8	.0479	98.8	96.0
	β_5	.000929	.000942	1.8	.000958	98.8	95.8
	α_2	.000615	.000686	12.1	.000802	99.6	95.0
	α_5	.00000980	.0000109	6.0	.0000116	99.6	95.6
$5, 20$	$\bar{Y}_3$	.0319	.0319	0.0	.0319	95.6	95.4
	β_1	.0448	.0449	0.0	.0449	95.4	95.4
	β_5	.000878	.000901	0.0	.000901	95.8	95.7
	α_2	.000581	.000662	4.1	.000701	99.1	95.0
	α_5	.00000925	.0000103	0.4	.0000103	97.3	96.0
$20, 5$	$\bar{Y}_3$	.0303	.0308	0.0	.0308	95.9	94.8
	β_1	.0454	.0450	0.0	.0450	95.1	94.9
	β_5	.000885	.000890	0.0	.000890	95.1	94.8
	α_2	.000501	.000576	0.7	.000582	98.4	94.1
	α_5	.00000870	.0000953	0.1	.00000955	96.9	95.0
$20, 20$	$\bar{Y}_3$	.0312	.0305	0.0	.0305	94.6	94.6
	β_1	.0426	.0444	0.0	.0444	95.5	95.5
	β_5	.000850	.000885	0.0	.000885	95.6	95.6
	α_2	.000492	.000573	0.0	.000573	96.6	95.9
	α_5	.00000869	.00000946	0.0	.00000946	96.0	96.0

Table 1: Simulation results for two stage fully synthetic data

Although not displayed in the table, we also evaluated inferences for the population mean of Y_1. The T_f was again unbiased, but it was negative in many iterations. This problem can be traced to an inconsistency between the derivations and the simulation design. The derivations assume that Y_a is not known for the population, so that $f(Y_a|D_{obs})$ is an estimated rather than exact distribution. In the simulation, we sample directly from the population and from $f(Y_a)$. Hence, the estimated variance of $\sum_i \bar{Y}_1^{(i)}/m$ equals $\bar{u}_M/m$. The T_f is still correct in expectation because the $w_r^{(i)} = 0$ for $i = 1, 2, ..., m$ and the $E(b_M) = \bar{u}_M$. However, the variability in b_M in this simulation is large enough to result in many instances where $T_f < 0$. There is a simple fix to this problem: for settings where the values in Y_a are sampled from a known population of values, use $\bar{u}_M/m$ instead of T_f to calculate the variance of estimates involving only Y_a.

We also examined the performance of the variance estimator for one stage fully synthetic data developed by Raghunathan *et al.* (2003). That is, we ignored the nesting. The one-stage variance estimator tends to underestimate variances. This underestimation becomes less severe as m and r increase.

5 Concluding Remarks

The key to any synthetic data approach is the imputation models. When high fractions of values are synthesized, the validity of inferences depends critically on the validity of the models used to generate the synthetic data. The synthetic data reflect only those relationships included in the data generation models. When the models fail to reflect accurately certain relationships, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses. In practice, this dependence means that some analyses cannot be performed accurately, and that agencies need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses. For example, agencies might include summaries of the posterior distributions of parameters in the data generation models as attachments to public releases of data. Or, they might include generic statements that describe the imputation models, such as "Main effects for age, sex, and race are included in the imputation models for education." This transparency also is a benefit of the synthetic data approach: analysts are given indications of which analyses can be reliably performed with the synthetic data. Analysts who desire finer detail than afforded by the imputations may have to apply for special access to the observed data.

As with multiple imputation for missing data, the inferential methods in Section 3 are derived from Bayesian perspectives and presume that the analyst and imputer use the same models for inferences about Q (Rubin, 1987, Chapter 3). This typically is not the case in public use data. Many analysts of public use data files estimate domain means and basic regressions, whereas agencies generate imputations from more complicated models. There has been little work on the properties of synthetic data inferences when the imputation and analysis models differ. Frequentist evaluations based on genuine data (Reiter, 2005b,d) suggest that one stage synthetic data inferences have good properties—in the sense that coverage rates of confidence intervals are near or exceed nominal rates—when the imputation models are more general than the analysts' inferences. Similar results are found in the missing data literature for congenial imputations (Meng, 1994; Schafer, 1997; Rubin, 2003a). The simulation results in this paper are in accord with these findings. These results notwithstanding, more research on congeniality issues for two stage synthetic data is needed.

Additional topics for future research specific to two stage synthesis include methods for selecting m and r based on risk-utility evaluations, for using the M data sets to do significance tests of multi-component hypotheses and other multivariate inference, and for handling missing data and confidentiality simultaneously, perhaps in a three stage imputation procedure.

For many data sets, concerns over confidentiality make it nearly impossible to release public use data. As resources available to malicious data users attempting re-identifications continue to expand, the alterations needed to protect data with traditional disclosure limitation techniques—such as swapping, adding noise, or microaggregation—may become so extreme that, for many analyses, the released data are no longer useful. Synthetic data, on the other hand, have the potential to enable data dissemination while preserving data utility. By synthesizing in two stages, data producers can improve the risk-utility profile, or reduce the labor costs, of their data releases.

Appendix: Derivation of Approximate Degrees of Freedom

Here we derive the degrees of freedom for the approximate t-distributions for two stage fully and partially synthetic data.

A.1 Fully synthetic data

The key step is to approximate the distribution of

$$\left(\frac{\nu_f T_f}{\bar{u}_M/(mr) + (1+1/m)B_\infty + (1+1/(mr))\bar{W}_\infty} \mid D_{syn}\right) \tag{27}$$

as a chi-squared distribution with ν_f degrees of freedom. The ν_f is determined by matching the mean and variance of the inverted χ^2 distribution to the mean and variance of (27).

Let $\gamma = (B_\infty + \bar{W}_\infty/r + \bar{u}_M/r)/b_M$, and let $\delta = (\bar{W}_\infty + \bar{u}_M)/\bar{w}_r$. Making the approximation that the $W_\infty^{(i)} = \bar{W}_\infty$ for all i, the $(\gamma^{-1} \mid b_M)$ and $(\delta^{-1} \mid \bar{w}_M)$ have mean square distributions with degrees of freedom $m-1$ and $m(r-1)$, respectively. Substituting γ and δ into (27), the random variable is

$$\frac{T_f}{\bar{u}_M/(mr) + (1+1/m)(\gamma b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M)}. \tag{28}$$

We need to approximate the expectation and variance of (28) and match them to a mean square random variable with ν_f degrees of freedom. We write the expectation as

$$E\left(E\left(\frac{T_f}{\bar{u}_M/(mr) + (1+1/m)(\gamma b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M)} \mid \delta\right)\right), \tag{29}$$

where the D_{syn} is suppressed from both expectations for brevity. We approximate the expectations using first order Taylor series expansions in γ^{-1} and δ^{-1} around their expectations, which equal one. The approximation boils down to substituting ones for γ and δ. After substitution, the denominator in (28) approximately equals T_f, and the expectation approximately equals one.

For the variance, we use the conditional variance representation

$$Var\left(E\left(\frac{T_f}{\bar{u}_M/(mr) + (1+1/m)(\gamma b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M)} \mid \delta\right)\right)$$
$$+E\left(Var\left(\frac{T_f}{\bar{u}_M/(mr) + (1+1/m)(\gamma b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M)} \mid \delta\right)\right). \tag{30}$$

For the interior expectation and variance, we use first order Taylor series expansions in γ^{-1} around its expectation. The first term in (30) approximately equals

$$Var\left(\frac{T_f}{\bar{u}_M/(mr) + (1+1/m)(b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M)}\right). \quad (31)$$

Since $Var(\gamma^{-1} \mid D_{syn}, \delta) = 2/(m-1)$, the second term in (30) approximately equals

$$E\left(\frac{(2/(m-1))T_f^2((1+1/m)b_M)^2}{(\bar{u}_M/(mr) + (1+1/m)(b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M))^4}\right) \quad (32)$$

We next approximate the the variance in (31) and the expectation in (32) using first order Taylor series expansions in δ^{-1} around its expectation. Since $Var(\delta^{-1} \mid D_{syn}) = 2/(m(r-1))$, the variance in (31) approximately equals

$$\frac{2/(m(r-1))T_f^2((1-1/r)\bar{w}_M)^2}{T_f^4}. \quad (33)$$

The expectation in (32) approximately equals

$$\frac{(2/(m-1))T_f^2((1+1/m)b_M)^2}{T_f^4}. \quad (34)$$

The variance in (30) is approximately the sum of (33) and (34). Since a mean square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that the

$$\nu_f = \left(\frac{((1+1/m)b_M)^2}{(m-1)T_f^2} + \frac{((1-1/r)\bar{w}_M)^2}{(m(r-1))T_f^2}\right)^{-1}. \quad (35)$$

A.2 Partially synthetic data

We approximate the distribution of

$$\left(\frac{\nu_p T_p}{\bar{u}_M + b_\infty/m + \bar{w}_\infty/(mr)} \mid D_{syn}\right) \quad (36)$$

as a chi-squared distribution with ν_p degrees of freedom. The ν_p is determined by matching the mean and variance of the inverted χ^2 distribution to the mean and variance of (36).

Let $\phi = (b_\infty + \bar{w}_\infty/r)/b_M$, and let $\psi = \bar{w}_\infty/\bar{w}_M$. Making the approximation that the $w_\infty^{(i)} = \bar{w}_\infty$ for all i, the $(\phi^{-1} \mid D_{syn}, \bar{w}_\infty)$ and $(\psi^{-1} \mid D_{syn})$ have mean square distributions with degrees of freedom $m-1$ and $m(r-1)$, respectively. We write the random variable in (36) as

$$\frac{T_p}{\bar{u}_M + \phi b_M/m}. \quad (37)$$

To match moments, we need to approximate the expectation and variance of (37) and match them to a mean square random variable with ν_p degrees of freedom.

We write the expectation of (37) as

$$E\left(E\left(\frac{T_p}{\bar{u}_M + \phi b_M/m} \mid D_{syn}, \bar{w}_\infty\right) \mid D_{syn}\right). \tag{38}$$

We approximate these expectations using first order Taylor series expansions in ψ^{-1} and ϕ^{-1} around their expectations, which equal one. The approximation boils down to substituting one for ϕ, as the ψ never enters the computations except in the conditioning arguments for ϕ. After substitution, the denominator in (36) approximately equals T_p, and the expectation approximately equals one.

For the variance, we use the conditional variance representation

$$E\left(Var\left(\frac{T_p}{\bar{u}_M + \phi b_M/m} \mid d^M, \bar{w}_\infty\right) \mid D_{syn}\right) + Var\left(E\left(\frac{T_p}{\bar{u}_M + \phi b_M/m} \mid d^M, \bar{w}_\infty\right) \mid d^M\right). \tag{39}$$

For the interior expectation and variance, we use first order Taylor series expansions in ϕ^{-1} and ψ^{-1} around their expectations. The interior expectation equals approximately one, so that the variance in the second term equals zero. Since $Var(\phi^{-1} \mid D_{syn}, \bar{w}_\infty) = 2/(m-1)$, the interior variance in (39) approximately equals

$$E\left(\frac{2T_p^2(b_M/m)^2}{(m-1)(\bar{u}_M + b_M/m)^4} \mid D_{syn}\right) = \frac{2(b_M/m)^2}{(m-1)T_p^2}. \tag{40}$$

Since a mean square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that

$$\nu_p = (m-1)(T_p/(b_M/m))^2 = (m-1)(1 + m\bar{u}_M/b_M)^2. \tag{41}$$

References

Abowd, J. M. and Lane, J. I. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 282–289. New York: Springer-Verlag.

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.

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

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2007). A new approach for disclosure control in the IAB establishment panel–Multiple imputation for a better data access. Tech. rep., IAB Discussion Paper, No.11/2007.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Tech. rep., U.S. National Institute of Statistical Sciences.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University.

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.

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

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statistical Science* **20**, 163–177.

Harel, O. and Schafer, J. (2003). Multiple imputation in two stages. In *Proceedings of Federal Committee on Statistical Methodology 2003 Conference*.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60**, 224–232.

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., Liu, F., and Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Persepectives*, 141–152. New York: John Wiley & Sons.

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). Selective multiple imputation of keys for statistical disclosure control in microdata. In *ASA Proceedings of the Joint Statistical Meetings*, 2133–2138.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science* **9**, 538–558.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrer and L. Franconi, eds., *Privacy in Statistical Databases*, 177–188. New York: Springer-Verlag.

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). Estimating identification risks in microdata. *Journal of the American Statistical Association* **100**, 1103–1113.

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

Reiter, J. P. (2005d). 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.

Rubin, D. B. (2003a). Discussion on multiple imputation. *International Statistical Review* **71**, 619–625.

Rubin, D. B. (2003b). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.

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

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Dept. of Statistics.

Skinner, C. J. and Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B* **64**, 855–867.

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

Recently published

No.	Author(s)	Title	Date
1/2004	Bauer, T. K. Bender, S. Bonin, H.	Dismissal protection and worker flows in small establishments published in: Economica, (2007)	7/04
2/2004	Achatz, J. Gartner, H. Glück, T.	Bonus oder Bias? : Mechanismen geschlechtsspezifischer Entlohnung published in: Kölner Zeitschrift für Soziologie und Sozialpsychologie 57 (2005), S. 466-493 (revised)	7/04
3/2004	Andrews, M. Schank, T. Upward, R.	Practical estimation methods for linked employer-employee data	8/04
4/2004	Brixy, U. Kohaut, S. Schnabel, C.	Do newly founded firms pay lower wages? : first evidence from Germany published in: Small Business Economics, (2007)	9/04
5/2004	Kölling, A. Rässler, S.	Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models published in: Zeitschrift für ArbeitsmarktForschung 37 (2004), S. 306-318	10/04
6/2004	Stephan, G. Gerlach, K.	Collective contracts, wages and wage dispersion in a multi-level model published as: Wage settlements and wage setting : results from a multi-level model. In: Applied Economics, Vol. 37, No. 20 (2005), S. 2297-2306	10/04
7/2004	Gartner, H. Stephan, G.	How collective contracts and works councils reduce the gender wage gap	12/04
1/2005	Blien, U. Suedekum, J.	Local economic structure and industry development in Germany, 1993-2001	1/05
2/2005	Brixy, U. Kohaut, S. Schnabel, C.	How fast do newly founded firms mature? : empirical analyses on job quality in start-ups published in: Michael Fritsch, Jürgen Schmude (Ed.): Entrepreneurship in the region, New York et al., 2006, S. 95-112	1/05
3/2005	Lechner, M. Miquel, R. Wunsch, C.	Long-run effects of public sector sponsored training in West Germany	1/05
4/2005	Hinz, T. Gartner, H.	Lohnunterschiede zwischen Frauen und Männern in Branchen, Berufen und Betrieben published in: Zeitschrift für Soziologie 34 (2005), S. 22-39, as: Geschlechtsspezifische Lohnunterschiede in Branchen, Berufen und Betrieben	2/05
5/2005	Gartner, H. Rässler, S.	Analyzing the changing gender wage gap based on multiply imputed right censored wages	2/05
6/2005	Alda, H. Bender, S. Gartner, H.	The linked employer-employee dataset of the IAB (LIAB) published as: The linked employer-employee dataset created from the IAB establishment panel and the process-produced data of the IAB (LIAB). In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 125 (2005), S. 327-336 (shortened)	3/05
7/2005	Haas, A. Rothe, T.	Labour market dynamics from a regional perspective : the multi-account system	4/05
8/2005	Caliendo, M. Hujer, R. Thomsen, S. L.	Identifying effect heterogeneity to improve the efficiency of job creation schemes in Germany	4/05

9/2005	Gerlach, K. Stephan, G.	Wage distributions by wage-setting regime published as: Bargaining regimes and wage dispersion. In: Jahrbücher für Nationalökonomie und Statistik, Bd. 226, H. 6 (2006)	4/05
10/2005	Gerlach, K. Stephan, G.	Individual tenure and collective contracts	4/05
11/2005	Blien, U. Hirschenauer, F.	Formula allocation : the regional allocation of budgetary funds for measures of active labour market policy in Germany published in: Economics Bulletin, Vol. 18, no. 7 (2006)	4/05
12/2005	Alda, H. Allaart, P. Bellmann, L.	Churning and institutions : Dutch and German establishments compared with micro-level data	5/05
13/2005	Caliendo, M. Hujer, R. Thomsen, S. L.	Individual employment effects of job creation schemes in Germany with respect to sectoral heterogeneity	5/05
14/2005	Lechner, M. Miquel, R. Wunsch, C.	The curse and blessing of training the unemployed in a changing economy : the case of East Germany after unification	6/05
15/2005	Jensen, U. Rässler, S.	Where have all the data gone? : stochastic production frontiers with multiply imputed German establishment data published in: Zeitschrift für ArbeitsmarktForschung, Jg. 39, H. 2, 2006, S. 277-295	7/05
16/2005	Schnabel, C. Zagelmeyer, S. Kohaut, S.	Collective bargaining structure and its determinants : an empirical analysis with British and German establishment data published in: European Journal of Industrial Relations, Vol. 12, No. 2. S. 165-188	8/05
17/2005	Koch, S. Stephan, G. Walwei, U.	Workfare: Möglichkeiten und Grenzen published in: Zeitschrift für ArbeitsmarktForschung 38 (2005), S. 419-440	8/05
18/2005	Alda, H. Bellmann, L. Gartner, H.	Wage structure and labour mobility in the West German private sector 1993-2000	8/05
19/2005	Eichhorst, W. Konle-Seidl, R.	The interaction of labor market regulation and labor market policies in welfare state reform	9/05
20/2005	Gerlach, K. Stephan, G.	Tarifverträge und betriebliche Entlohnungsstrukturen published in: C. Clemens, M. Heinemann & S. Soretz (Hg.): Auf allen Märkten zu Hause, Marburg 2006, S. 123-143	11/05
21/2005	Fitzenberger, B. Speckesser, S.	Employment effects of the provision of specific professional skills and techniques in Germany published in: Empirical Economics, Vol. 32, No. 2/3 (2007), S. 529-573	11/05
22/2005	Ludsteck, J. Jacobebbinghaus, P.	Strike activity and centralisation in wage setting	12/05
1/2006	Gerlach, K. Levine, D. Stephan, G. Struck, O.	The acceptability of layoffs and pay cuts : comparing North America with Germany	1/06
2/2006	Ludsteck, J.	Employment effects of centralization in wage setting in a median voter model	2/06
3/2006	Gaggermeier, C.	Pension and children : Pareto improvement with heterogeneous preferences	2/06
4/2006	Binder, J. Schwengler, B.	Korrekturverfahren zur Berechnung der Einkommen über der Beitragsbemessungsgrenze	3/06
5/2006	Brixy, U.	Regional patterns and determinants of new firm formation	4/06

	Grotz, R.	and survival in western Germany	
6/2006	Blien, U. Sanner, H.	Structural change and regional employment dynamics	4/06
7/2006	Stephan, G. Rässler, S. Schewe, T.	Wirkungsanalyse in der Bundesagentur für Arbeit : Konzeption, Datenbasis und ausgewählte Befunde published as: Das TrEffeR-Projekt der Bundesagentur für Arbeit : die Wirkung von Maßnahmen aktiver Arbeitsmarktpolitik. In: Zeitschrift für ArbeitsmarktForschung, Jg. 39, H. 3/4 (2006)	4/06
8/2006	Gash, V. Mertens, A. Romeu Gordo, L.	Are fixed-term jobs bad for your health? : a comparison of West-Germany and Spain published in: European Societies, 2007	5/06
9/2006	Romeu Gordo, L.	Compression of morbidity and the labor supply of older people	5/06
10/2006	Jahn, E. J. Wagner, T.	Base period, qualifying period and the equilibrium rate of unemployment	6/06
11/2006	Jensen, U. Gartner, H. Rässler, S.	Measuring overeducation with earnings frontiers and multiply imputed censored income data	6/06
12/2006	Meyer, B. Lutz, C. Schnur, P. Zika, G.	National economic policy simulations with global interdependencies : a sensitivity analysis for Germany published in: Economic systems research, Vol. 19, No. 1 (2007), S. 37-55	7/06
13/2006	Beblo, M. Bender, S. Wolf, E.	The wage effects of entering motherhood : a within-firm matching approach	8/06
14/2006	Niebuhr, A.	Migration and innovation : does cultural diversity matter for regional R&D activity?	8/06
15/2006	Kiesl, H. Rässler, S.	How valid can data fusion be? published in: Journal of Official Statistics, (2006)	8/06
16/2006	Hujer, R. Zeiss, C.	The effects of job creation schemes on the unemployment duration in East Germany	8/06
17/2006	Fitzenberger, B. Osikominu, A. Völter, R.	Get training or wait? : long-run employment effects of training programs for the unemployed in West Germany	9/06
18/2006	Antoni, M. Jahn, E. J.	Do changes in regulation affect employment duration in temporary work agencies?	9/06
19/2006	Fuchs, J. Söhnlein, D.	Effekte alternativer Annahmen auf die prognostizierte Erwerbsbevölkerung	10/06
20/2006	Lechner, M. Wunsch, C.	Active labour market policy in East Germany : waiting for the economy to take off	11/06
21/2006	Kruppe, T.	Die Förderung beruflicher Weiterbildung : eine mikroökonometrische Evaluation der Ergänzung durch das ESF-BA-Programm	11/06
22/2006	Feil, M. Klinger, S. Zika, G.	Sozialabgaben und Beschäftigung : Simulationen mit drei makroökonomischen Modellen	11/06
23/2006	Blien, U. Phan, t. H. V.	A pilot study on the Vietnamese labour market and its social and economic context	11/06
24/2006	Lutz, R.	Was spricht eigentlich gegen eine private Arbeitslosenversicherung?	11/06
25/2006	Jirjahn, U. Pfeifer, C.	Mikroökonomische Beschäftigungseffekte des Hamburger Modells zur Beschäftigungsförderung	11/06

	Tsertsvadze, G.		
26/2006	Rudolph, H.	Indikator gesteuerte Verteilung von Eingliederungsmitteln im SGB II : Erfolgs- und Effizienzkriterien als Leistungsanreiz?	12/06
27/2006	Wolff, J.	How does experience and job mobility determine wage gain in a transition and a non-transition economy? : the case of east and west Germany	12/06
28/2006	Blien, U. Kirchhof, K. Ludewig, O.	Agglomeration effects on labour demand	12/06
29/2006	Blien, U. Hirschenauer, F. Phan, t. H. V.	Model-based classification of regional labour markets : for purposes of labour market policy	12/06
30/2006	Krug, G.	Kombilohn und Reziprozität in Beschäftigungsverhältnissen : eine Analyse im Rahmen des Matching-Ansatzes	12/06
1/2007	Moritz, M. Gröger, M.	The German-Czech border region after the fall of the Iron Curtain: Effects on the labour market : an empirical study using the IAB Employment Sample (IABS)	1/07
2/2007	Hampel, K. Kunz, M. Schanne, N. Wapler, R. Weyh, A.	Regional employment forecasts with spatial interdependencies	1/07
3/2007	Eckey, H.- F. Schwengler, B. Türck, M.	Vergleich von deutschen Arbeitsmarktregionen	1/07
4/2007	Kristen, C. Granato, N.	The educational attainment of the second generation in Germany : social origins and ethnic inequality	1/07
5/2007	Jacob, M. Kleinert, C.	Does unemployment help or hinder becoming independent? : the role of employment status for leaving the parental home	1/07
6/2007	Konle-Seidl, R. Eichhorst, W. Grienberger-Zingerle, M.	Activation policies in Germany : from status protection to basic income support	1/07
7/2007	Lechner, M. Wunsch, C.	Are training programs more effective when unemployment is high?	2/07
8/2007	Hohendanner, C.	Verdrängen Ein-Euro-Jobs sozialversicherungspflichtige Beschäftigung in den Betrieben?	2/07
9/2007	Seibert, H.	Frühe Flexibilisierung? : regionale Mobilität nach der Lehrausbildung in Deutschland zwischen 1977 und 2004	2/07
10/2007	Bernhard, S. Kurz, K.	Familie und Arbeitsmarkt : eine Längsschnittstudie zum Einfluss beruflicher Unsicherheiten auf die Familienerweiterung	2/07
11/2007	Drechsler, J. Dundler, A. Bender, S. Rässler, S. Zwick, T.	A new approach for disclosure control in the IAB Establishment Panel : multiple imputation for a better data access	2/07
12/2007	Fuchs, J. Söhnlein, D.	Einflussfaktoren auf das Erwerbspersonenpotenzial : Demografie und Erwerbsverhalten in Ost- und Westdeutschland	3/07
13/2007	Hartmann, J. Krug, G.	Verknüpfung von Befragungs- und Prozessdaten : Selektivität durch fehlende Zustimmung der Befragten?	3/07
14/2007	Baltagi, B. H. Blien, U. Wolf, K.	Phillips Curve or wage curve? : evidence from West Germany: 1980-2004	4/07
15/2007	Blien, U.	Expensive and low-price places to live : regional price levels	4/07

Stand: 29.6.2007

Imprint

IABDiscussionPaper
No. 20 / 2007

Editorial address
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit
Weddigenstr. 20-22
D-90478 Nürnberg

Editorial staff
Regina Stoll, Jutta Palm-Nowak

Technical completion
Jutta Sebald

Download of this DiscussionPaper:
http://doku.iab.de/discussionpapers/2007/dp2007.pdf

Website
http://www.iab.de

For further inquiries contact the author:
Jörg Drechsler, Tel. 0911/179-4021,
or e-mail: joerg.drechsler@iab.de