Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality
Reiter, J. P.; Drechsler, Jörg

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
SSG Sozialwissenschaften, USB Köln

Empfohlene Zitierung / Suggested Citation:

Nutzungsbedingungen:
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. By using this particular document, you accept the above-stated conditions of use.
Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality

J.P. Reiter, Jörg Drechsler
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)


Also with its new series "IAB Discussion Paper" the research institute of the German Federal Employment Agency wants to intensify dialogue with external science. By the rapid spreading of research results via Internet still before printing criticism shall be stimulated and quality shall be ensured.
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: jorg.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, \( Y_{a} \), 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 \( f(Q \mid D_{syn}) \). The three-step process for creating \( D_{syn} \) suggests that

\[
f(Q \mid D_{syn}) = \int f(Q \mid D_{obs}, P_{syn}, D_{syn})f(D_{obs} \mid P_{syn}, D_{syn})f(P_{syn} \mid D_{syn})dD_{obs}dP_{syn},
\]

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 
\[
f(Q | D_{\text{obs}}, P_{\text{syn}}, D_{\text{syn}}) = f(Q | D_{\text{obs}}).
\]
We assume that 
\[
(Q | D_{\text{obs}}) \sim N(Q_{\text{obs}}, U_{\text{obs}}),
\]
where \( Q_{\text{obs}} \) and \( U_{\text{obs}} \) are the estimates of the mean and variance computed from \( D_{\text{obs}} \) if it were released.

The \( D_{\text{syn}} \) is irrelevant given \( P_{\text{syn}} \), so that 
\[
f(D_{\text{obs}} | P_{\text{syn}}, D_{\text{syn}}) = f(D_{\text{obs}} | P_{\text{syn}}).
\]
Because inferences for \( Q \) depend only on \( Q_{\text{obs}} \) and \( U_{\text{obs}} \), it is sufficient to determine \( f(Q_{\text{obs}}, U_{\text{obs}} | P_{\text{syn}}) \). Let \( Q^{(i,j)} \) be the estimate of \( Q \) in population \( P^{(i,j)} \). Let \( Q^{(i,j)}_{r} = \sum_j q^{(i,j)} / r \), and \( Q_M = \sum_i Q^{(i,j)}_r / m \).

Let \( B_M = \sum_i (Q^{(i,j)} - Q_M)^2 / (m - 1) \), and \( W_r = \sum_j (Q^{(i,j)} - Q^{(i,j)}_r)^2 / (r - 1) \). We assume the following sampling distributions:
\[
\begin{align*}
(Q^{(i,j)} | D_{\text{obs}}, B_{\infty}) & \sim N(Q_{\text{obs}}, B_{\infty}) \\
(Q^{(i,j)} | Q^{(i,j)}_r, W^{(i)}) & \sim N(Q^{(i,j)}_r, W^{(i)})
\end{align*}
\]
where the \( Q^{(i,j)}_r \), the \( W^{(i)} \), and \( B_{\infty} \) 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 \( U_{\text{obs}} = B_{\infty} + W_{\infty} \), where \( W_{\infty} = \lim \sum_i W^{(i)}_r / m \) as \( m \to \infty \). From (2), (3), and (4), for finite \( m \) and \( r \) we have 
\[
(Q | P_{\text{syn}}, B_{\infty}, W^{(1)}_{\infty}, \ldots, W^{(m)}_{\infty}) \sim N(Q_M, (1 + 1/m)B_{\infty} + (1 + 1/(mr))W_{\infty}).
\]
We also have
\[
\begin{align*}
((m - 1)B_M / (B_{\infty} + W_{\infty}) | P_{\text{syn}}, W_{\infty}) & \sim \chi^2_{m-1} \\
((r - 1)W_r / W_{\infty} | P_{\text{syn}}) & \sim \chi^2_{r-1}.
\end{align*}
\]
The posterior distribution of \( Q \) conditioning on \( P_{\text{syn}} \) alone is found by integrating (5) over the distributions in (6) and (7).

In general, releasing \( P_{\text{syn}} \) is impractical for agencies, as it could require releasing \( M \) data files of very large size \( N \). We therefore take random samples of size \( n_{\text{syn}} \) from each population, i.e. the \( D^{(i,j)} \). We require the distributions of \( Q_M, B_{\infty} \), and \( W^{(i)}_{\infty} \) conditional on \( D_{\text{syn}} \). 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 \( n_{\text{syn}} = N \), the \( u^{(i,j)} = 0 \). Let \( \bar{q}_r = \sum_j q^{(i,j)} / r \), and \( \bar{q}_M = \sum_i \bar{q}_r / m \). Let \( b_M = \sum_i (\bar{q}_r - \bar{q}_M)^2 / (m - 1) \), and \( w_r = \sum_j (q^{(i,j)} - \bar{q}_r)^2 / (r - 1) \). Finally, let \( \bar{u}_M = \sum_{i,j} u^{(i,j)} / (mr) \).

For \( n_{\text{syn}} \) large, we assume the sampling distribution of each \( (q^{(i,j)} | P_{\text{syn}}) \) 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
\[
(q^{(i,j)} | P_{\text{syn}}) \sim N(q^{(i,j)}, \bar{u}_M).
\]
Using the standard Bayesian arguments based on these sampling distributions, we have

\[ (Q^{(i)}|\tilde{q}_r^{(i)}, \bar{u}_M) \sim N(\tilde{q}_r^{(i)}, \tilde{u}_M/r) \] (9)

and

\[ (\tilde{Q}_M|D_{syn}) \sim N(\tilde{q}_M, \bar{u}_M/(mr)). \] (10)

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 \( \tilde{Q}_M, B_\infty, \) and the \( W^{(i)}_\infty \) 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

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

Similarly,

\[ Var(Q|D_{syn}) = Var[E(Q|P_{syn}, B_\infty, W_\infty)|D_{syn}] + Var[E(Q|P_{syn}, B_\infty, W_\infty)|D_{syn}] \]

\[ = (1 + m^{-1}) Var(\tilde{Q}_M|D_{syn}) + (1 + 1/(mr)) Var(W_\infty|D_{syn}) + \bar{u}_M/(mr). \] (15)

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

\[ Var(Q|D_{syn}) \approx (1 + m^{-1})(b_M - \bar{w}_M/r) + (1 + 1/(mr))(\bar{w}_M - \bar{u}_M) + \bar{u}_M/(mr) \]

\[ = (1 + m^{-1})b_M + (1 - 1/r)\bar{w}_M - \bar{u}_M = T_f. \] (16)

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

\[ \nu_f = \left( \frac{(1 + 1/m)b_M^2}{(m - 1)T_f^2} + \frac{(1 - 1/r)\bar{u}_M^2}{(m(r - 1))} \right)^{-1}. \]
The degrees of freedom is derived by matching the first two moments of \((\nu_f T_f)/(\bar{\sigma}_M/(mr) + (1+1/m)B_\infty + (1+1/(mr))\bar{W}_\infty)\) to an inverse chi-squared distribution with \(\nu_f\) 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,

\[ T_f^* = T_f + \lambda \bar{\sigma}_M, \tag{17} \]

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 \(\nu_f\) 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 \(\nu_f^*\), where

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

The \(\nu_f\) can be very small even when \(T_f > 0\), which could result in excessively wide intervals. We evaluate a modification to \(\nu_f\) 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,

\[ f(Q \mid D_{syn}) = \int f(Q \mid D_{obs}, D_{syn}) f(D_{obs} \mid D_{syn}) dD_{obs}. \tag{19} \]

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

\[
\begin{align*}
(q^{(i)}_{\infty}|D_{obs}, b_{\infty}) & \sim N(Q_{obs}, b_{\infty}) \\
(q^{(i,j)}|D_{obs}, q^{(i)}_{\infty}, w^{(i)}_{\infty}) & \sim N(q^{(i)}_{\infty}, w^{(i)}_{\infty}).
\end{align*}
\]  

(20)  
(21)

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

\[
(Q|D_{syn}, b_{\infty}, w^{(1)}_{\infty}, \ldots, w^{(m)}_{\infty}) \sim N(q_{M}, U_{obs} + b_{\infty}/m + \bar{w}_{\infty}/(mr)).
\]  

(22)

Since the \( Y_o \) 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 \( b_{\infty} \) and the \( w^{(i)}_{\infty} \) are obtained from an analysis of variance setup. From (21), we have

\[
\frac{(r - 1)w_{M}^{(i)}}{w_{\infty}^{(i)}}|D_{syn} \sim \chi^2_{r-1}.
\]  

(23)

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

\[
\begin{align*}
\frac{(m - 1)b_{M}}{b_{\infty} + \bar{w}_{\infty}/r}|D_{syn}, \bar{w}_{\infty}) & \sim \chi^2_{m-1} \\
\frac{(m(r - 1)\bar{u}_M)}{\bar{w}_{\infty}}|D_{syn} & \sim \chi^2_{m(r-1)}.
\end{align*}
\]  

(24)  
(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 \( b_{\infty} \) 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 \( \nu_p \) 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
The observed data, $D_{\text{obs}}$, comprise the values of $(Y_1, \ldots, Y_5)$ for a simple random sample of $s = 1,000$ records from the 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 $i = 1, \ldots, m$ and $j = 1, \ldots, r$, we generate $Y_a^{(i)}$ by taking a random sample of $n_{\text{syn}} = 1,000$ 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, $f(Y_3, Y_4, Y_5|D_{\text{obs}}, Y_a^{(i)})$, with noninformative prior distributions on all parameters. That is, we draw $Y_3^{(i,j)}$ from the regression $f(Y_3|D_{\text{obs}}, Y_a^{(i)})$, we draw $Y_4^{(i,j)}$ from the regression $f(Y_4|D_{\text{obs}}, Y_a^{(i)}, Y_3^{(i,j)})$, and we draw $Y_5^{(i,j)}$ from the regression $f(Y_5|D_{\text{obs}}, Y_a^{(i)}, Y_3^{(i,j)}, Y_4^{(i,j)})$. 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 $Y_1$ ($\beta_1$) and of $Y_5$ ($\beta_5$) in a regression of $Y_3$ on all other variables, and the regression coefficients of $Y_2$ ($\alpha_2$) and of $Y_5$ ($\alpha_5$) 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 $\text{Var}(\bar{q}_M)$. The $T_f$ associated with $\alpha_2$ and $\alpha_5$ 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 $B_\infty$ 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 $\nu_f^*$ 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 $\nu_f^*$, 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,

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

<table>
<thead>
<tr>
<th>$m, r$</th>
<th>$Q$</th>
<th>$\text{Var}(\bar{q}_M)$</th>
<th>Avg. $T_f$</th>
<th>$% T_f &lt; 0$</th>
<th>Avg. $T_f^*$</th>
<th>95% CI Cov*</th>
<th>95% CI Cov**</th>
</tr>
</thead>
<tbody>
<tr>
<td>3, 3</td>
<td>$\beta_3$</td>
<td>.0409</td>
<td>.0389</td>
<td>15.7</td>
<td>.0448</td>
<td>94.2</td>
<td>93.8</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>.0537</td>
<td>.0533</td>
<td>12.3</td>
<td>.0587</td>
<td>98.0</td>
<td>95.9</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>.00108</td>
<td>.00106</td>
<td>12.2</td>
<td>.00117</td>
<td>98.0</td>
<td>96.2</td>
</tr>
<tr>
<td></td>
<td>$\alpha_2$</td>
<td>.000766</td>
<td>.000850</td>
<td>24.8</td>
<td>.00109</td>
<td>97.6</td>
<td>96.3</td>
</tr>
<tr>
<td></td>
<td>$\alpha_5$</td>
<td>.000121</td>
<td>.000126</td>
<td>19.3</td>
<td>.000151</td>
<td>97.8</td>
<td>95.7</td>
</tr>
<tr>
<td>5, 5</td>
<td>$\beta_3$</td>
<td>.0327</td>
<td>.0335</td>
<td>3.6</td>
<td>.0349</td>
<td>99.2</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>.0458</td>
<td>.0471</td>
<td>1.8</td>
<td>.0479</td>
<td>98.8</td>
<td>96.0</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>.000929</td>
<td>.000942</td>
<td>1.8</td>
<td>.000958</td>
<td>98.8</td>
<td>95.8</td>
</tr>
<tr>
<td></td>
<td>$\alpha_2$</td>
<td>.000615</td>
<td>.000686</td>
<td>12.1</td>
<td>.000802</td>
<td>99.6</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>$\alpha_5$</td>
<td>.0000980</td>
<td>.000109</td>
<td>6.0</td>
<td>.000116</td>
<td>99.6</td>
<td>95.6</td>
</tr>
<tr>
<td>5, 20</td>
<td>$\beta_3$</td>
<td>.0319</td>
<td>.0319</td>
<td>0.0</td>
<td>.0319</td>
<td>95.6</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>.0448</td>
<td>.0449</td>
<td>0.0</td>
<td>.0449</td>
<td>95.4</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>.000878</td>
<td>.000901</td>
<td>0.0</td>
<td>.000901</td>
<td>95.8</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td>$\alpha_2$</td>
<td>.000581</td>
<td>.000662</td>
<td>4.1</td>
<td>.000701</td>
<td>99.1</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>$\alpha_5$</td>
<td>.0000925</td>
<td>.000103</td>
<td>0.4</td>
<td>.000103</td>
<td>97.3</td>
<td>96.0</td>
</tr>
<tr>
<td>20, 5</td>
<td>$\beta_3$</td>
<td>.0303</td>
<td>.0308</td>
<td>0.0</td>
<td>.0308</td>
<td>95.9</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>.0454</td>
<td>.0450</td>
<td>0.0</td>
<td>.0450</td>
<td>95.1</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>.000885</td>
<td>.000890</td>
<td>0.0</td>
<td>.000890</td>
<td>95.1</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>$\alpha_2$</td>
<td>.000501</td>
<td>.000576</td>
<td>0.7</td>
<td>.000582</td>
<td>98.4</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>$\alpha_5$</td>
<td>.0000870</td>
<td>.0000953</td>
<td>0.1</td>
<td>.0000955</td>
<td>96.9</td>
<td>95.0</td>
</tr>
<tr>
<td>20, 20</td>
<td>$\beta_3$</td>
<td>.0312</td>
<td>.0305</td>
<td>0.0</td>
<td>.0305</td>
<td>94.6</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>.0426</td>
<td>.0444</td>
<td>0.0</td>
<td>.0444</td>
<td>95.5</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>.000850</td>
<td>.000885</td>
<td>0.0</td>
<td>.000885</td>
<td>95.6</td>
<td>95.6</td>
</tr>
<tr>
<td></td>
<td>$\alpha_2$</td>
<td>.000492</td>
<td>.000573</td>
<td>0.0</td>
<td>.000573</td>
<td>96.6</td>
<td>95.9</td>
</tr>
<tr>
<td></td>
<td>$\alpha_5$</td>
<td>.0000869</td>
<td>.0000946</td>
<td>0.0</td>
<td>.0000946</td>
<td>96.0</td>
<td>96.0</td>
</tr>
</tbody>
</table>

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 $\sum_i Y_1^{(i)} / m$ 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 \( \tilde{u}_M/m \) 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

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

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

Let $\gamma = (B_{\infty} + W_{\infty}/r + \bar{u}_M/r)/b_M$, and let $\delta = (W_{\infty} + \bar{u}_M)/\bar{w}_r$. 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 $\gamma$ and $\delta$ into (27), the random variable is

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

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

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

where the $D_{syn}$ is suppressed from both expectations for brevity. We approximate the expectations using first order Taylor series expansions in $\gamma^{-1}$ and $\delta^{-1}$ around their expectations, which equal one. The approximation boils down to substituting ones for $\gamma$ and $\delta$. 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

$$
Var \left( E \left( \frac{T_f}{\bar{u}_M/(mr) + (1 + 1/m)(\gamma b_M - \delta \bar{w}_M/r) + (1 + 1/(mr))(\delta \bar{w}_M - \bar{u}_M)} \mid \delta \right) \right) + E \left( Var \left( \frac{T_f}{\bar{u}_M/(mr) + (1 + 1/m)(\gamma b_M - \delta \bar{w}_M/r) + (1 + 1/(mr))(\delta \bar{w}_M - \bar{u}_M)} \mid \delta \right) \right). \quad (30)
$$
For the interior expectation and variance, we use first order Taylor series expansions in $\gamma^{-1}$ around its expectation. The first term in (30) approximately equals

$$\text{Var}(T_f)^{\gamma^{-1}} \| \frac{1}{\mu_{M}/(mr) + (1 + 1/m)(b_M - \delta \bar{w}_M/r) + (1 + 1/(mr))(\delta \bar{w}_M - \bar{u}_M)} .$$

(31)

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

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

(32)

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

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

(33)

The expectation in (32) approximately equals

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

(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

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

(35)

A.2 Partially synthetic data

We approximate the distribution of

$$\left(\begin{array}{c}
\nu_p T_p \\
\bar{u}_M + b_{\infty}/m + \bar{w}_{\infty}/(mr)
\end{array} \ | \ D_{syn}\right)$$

as a chi-squared distribution with $\nu_p$ degrees of freedom. The $\nu_p$ is determined by matching the mean and variance of the inverted $\chi^2$ 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 $w^{(i)}_{\infty} = \bar{w}_M$ for all $i$, the $(\phi^{-1} \ | \ D_{syn}, \bar{w}_{\infty})$ and $(\psi^{-1} \ | \ D_{syn})$ have mean square distributions with degrees of freedom $m - 1$ and $m(r - 1)$, respectively. We write the random variable in (36) as

$$\frac{T_p}{\bar{u}_M + \phi b_M/m} .$$

(37)

To match moments, we need to approximate the expectation and variance of (37) and match them to a mean square random variable with $\nu_p$ 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 $\psi^{-1}$ and $\phi^{-1}$ around their expectations, which equal one. The approximation boils down to substituting one for $\psi$, as the $\psi$ never enters the computations except in the conditioning arguments for $\phi$. 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 $\phi^{-1}$ and $\psi^{-1}$ 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
\[
E \left( \frac{2T^2_p(b_M/m)^2}{(m - 1)(\bar{u}_M + b_M/m)^4} \mid D_{syn} \right) = \frac{2(b_M/m)^2}{(m - 1)T^2_p}. \tag{40}
\]
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


# Recently published

<table>
<thead>
<tr>
<th>No.</th>
<th>Author(s)</th>
<th>Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/2004</td>
<td>Gartner, H. Stephan, G.</td>
<td>How collective contracts and works councils reduce the gender wage gap</td>
<td>12/04</td>
</tr>
<tr>
<td>1/2005</td>
<td>Blien, U. Suedekum, J.</td>
<td>Local economic structure and industry development in Germany, 1993-2001</td>
<td>1/05</td>
</tr>
<tr>
<td>3/2005</td>
<td>Lechner, M. Miquel, R. Wunsch, C.</td>
<td>Long-run effects of public sector sponsored training in West Germany</td>
<td>1/05</td>
</tr>
<tr>
<td>5/2005</td>
<td>Gartner, H. Rässler, S.</td>
<td>Analyzing the changing gender wage gap based on multiply imputed right censored wages</td>
<td>2/05</td>
</tr>
<tr>
<td>6/2005</td>
<td>Alda, H. Bender, S. Gartner, H.</td>
<td>The linked employer-employee dataset of the IAB (LIAB) published as: <em>The linked employer-employee dataset created from the IAB establishment panel and the process-produced data of the IAB (LIAB)</em>. In: <em>Schmollers Jahrbuch, Zeitschrift für Wirtschafts- und Sozialwissenschaften</em> 125 (2005), S. 327-336 (shortened)</td>
<td>3/05</td>
</tr>
<tr>
<td>7/2005</td>
<td>Haas, A. Rothe, T.</td>
<td>Labour market dynamics from a regional perspective : the multi-account system</td>
<td>4/05</td>
</tr>
<tr>
<td>8/2005</td>
<td>Caliendo, M. Hujer, R. Thomsen, S. L.</td>
<td>Identifying effect heterogeneity to improve the efficiency of job creation schemes in Germany</td>
<td>4/05</td>
</tr>
<tr>
<td>Date</td>
<td>Authors</td>
<td>Title</td>
<td>Publication</td>
</tr>
<tr>
<td>------</td>
<td>------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>10/2005</td>
<td>Gerlach, K.</td>
<td>Individual tenure and collective contracts</td>
<td></td>
</tr>
<tr>
<td>12/2005</td>
<td>Alda, H.</td>
<td>Churning and institutions: Dutch and German establishments compared with micro-level data</td>
<td></td>
</tr>
<tr>
<td>13/2005</td>
<td>Caliendo, M.</td>
<td>Individual employment effects of job creation schemes in Germany</td>
<td>With respect to sectoral heterogeneity</td>
</tr>
<tr>
<td>14/2005</td>
<td>Lechner, M.</td>
<td>The curse and blessing of training the unemployed in a changing economy: the case of East Germany after unification</td>
<td></td>
</tr>
<tr>
<td>15/2005</td>
<td>Jensen, U.</td>
<td>Where have all the data gone?: stochastic production frontiers with multiply imputed German establishment data</td>
<td>Published in: <em>Zeitschrift für ArbeitsmarktForschung</em>, Jg. 39, H. 2, 2006, S. 277-295</td>
</tr>
<tr>
<td>16/2005</td>
<td>Schnabel, C.</td>
<td>Collective bargaining structure and its determinants: an empirical analysis with British and German establishment data</td>
<td>Published in: <em>European Journal of Industrial Relations</em>, Vol. 12, No. 2, S. 165-188</td>
</tr>
<tr>
<td>18/2005</td>
<td>Alda, H.</td>
<td>Wage structure and labour mobility in the West German private sector 1993-2000</td>
<td></td>
</tr>
<tr>
<td>19/2005</td>
<td>Eichhorst, W.</td>
<td>The interaction of labor market regulation and labor market policies in welfare state reform</td>
<td></td>
</tr>
<tr>
<td>22/2005</td>
<td>Ludsteck, J.</td>
<td>Strike activity and centralisation in wage setting</td>
<td></td>
</tr>
<tr>
<td>1/2006</td>
<td>Gerlach, K.</td>
<td>The acceptability of layoffs and pay cuts: comparing North America with Germany</td>
<td></td>
</tr>
<tr>
<td>2/2006</td>
<td>Ludsteck, J.</td>
<td>Employment effects of centralisation in wage setting in a median voter model</td>
<td></td>
</tr>
<tr>
<td>3/2006</td>
<td>Gaggermeier, C.</td>
<td>Pension and children: Pareto improvement with heterogeneous preferences</td>
<td></td>
</tr>
<tr>
<td>4/2006</td>
<td>Binder, J.</td>
<td>Korrekturverfahren zur Berechnung der Einkommen über der Beitragsbemessungsgrenze</td>
<td></td>
</tr>
<tr>
<td>5/2006</td>
<td>Brixy, U.</td>
<td>Regional patterns and determinants of new firm formation</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Authors</td>
<td>Title</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>6/2006</td>
<td>Grotz, R.</td>
<td>and survival in western Germany</td>
<td></td>
</tr>
<tr>
<td>7/2006</td>
<td>Blien, U. Sanner, H.</td>
<td>Structural change and regional employment dynamics</td>
<td></td>
</tr>
<tr>
<td>9/2006</td>
<td>Romeu Gordo, L.</td>
<td>Compression of morbidity and the labor supply of older people</td>
<td></td>
</tr>
<tr>
<td>10/2006</td>
<td>Jahn, E. J. Wagner, T.</td>
<td>Base period, qualifying period and the equilibrium rate of unemployment</td>
<td></td>
</tr>
<tr>
<td>13/2006</td>
<td>Beblo, M. Bender, S. Wolf, E.</td>
<td>The wage effects of entering motherhood: a within-firm matching approach</td>
<td></td>
</tr>
<tr>
<td>14/2006</td>
<td>Niebuhr, A.</td>
<td>Migration and innovation: does cultural diversity matter for regional R&amp;D activity?</td>
<td></td>
</tr>
<tr>
<td>15/2006</td>
<td>Kiesl, H. Rässler, S.</td>
<td>How valid can data fusion be?</td>
<td></td>
</tr>
<tr>
<td>16/2006</td>
<td>Hujer, R. Zeiss, C.</td>
<td>The effects of job creation schemes on the unemployment duration in East Germany</td>
<td></td>
</tr>
<tr>
<td>17/2006</td>
<td>Fitzengerber, B. Osikominu, A. Völter, R.</td>
<td>Get training or wait?: long-run employment effects of training programs for the unemployed in West Germany</td>
<td></td>
</tr>
<tr>
<td>18/2006</td>
<td>Antoni, M. Jahn, E. J.</td>
<td>Do changes in regulation affect employment duration in temporary work agencies?</td>
<td></td>
</tr>
<tr>
<td>19/2006</td>
<td>Fuchs, J. Söhnlein, D.</td>
<td>Effekte alternativer Annahmen auf die prognostizierte Erwerbsbevölkerung</td>
<td></td>
</tr>
<tr>
<td>20/2006</td>
<td>Lechner, M. Wunsch, C.</td>
<td>Active labour market policy in East Germany: waiting for the economy to take off</td>
<td></td>
</tr>
<tr>
<td>21/2006</td>
<td>Kruppe, T.</td>
<td>Die Förderung beruflicher Weiterbildung: eine mikroökometrische Evaluation der Ergänzung durch das ESF-BA-Programm</td>
<td></td>
</tr>
<tr>
<td>22/2006</td>
<td>Feil, M. Klinger, S. Zika, G.</td>
<td>Sozialabgaben und Beschäftigung: Simulationen mit drei makroökonomischen Modellen</td>
<td></td>
</tr>
<tr>
<td>24/2006</td>
<td>Lutz, R.</td>
<td>Was spricht eigentlich gegen eine private Arbeitslosenversicherung?</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Authors</td>
<td>Title</td>
<td>Publication Date</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>27/2006</td>
<td>Wolff, J.</td>
<td>How does experience and job mobility determine wage gain in a transition and a non-transition economy?: the case of east and west Germany</td>
<td>12/06</td>
</tr>
<tr>
<td>28/2006</td>
<td>Blien, U., Kirchhof, K., Ludewig, O.</td>
<td>Agglomeration effects on labour demand</td>
<td>12/06</td>
</tr>
<tr>
<td>30/2006</td>
<td>Krug, G.</td>
<td>Kombilohn und Reziprozität in Beschäftigungsverhältnissen: eine Analyse im Rahmen des Matching-Ansatzes</td>
<td>12/06</td>
</tr>
<tr>
<td>1/2007</td>
<td>Moritz, M., Gröger, M.</td>
<td>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)</td>
<td>1/07</td>
</tr>
<tr>
<td>2/2007</td>
<td>Hampel, K., Kunz, M., Schanne, N., Wapler, R., Weyh, A.</td>
<td>Regional employment forecasts with spatial interdependencies</td>
<td>1/07</td>
</tr>
<tr>
<td>4/2007</td>
<td>Kristen, C., Granato, N.</td>
<td>The educational attainment of the second generation in Germany: social origins and ethnic inequality</td>
<td>1/07</td>
</tr>
<tr>
<td>5/2007</td>
<td>Jacob, M., Kleinert, C.</td>
<td>Does unemployment help or hinder becoming independent?: the role of employment status for leaving the parental home</td>
<td>1/07</td>
</tr>
<tr>
<td>6/2007</td>
<td>Konle-Seidl, R., Eichhorst, W., Grienberger-Zingerle, M.</td>
<td>Activation policies in Germany: from status protection to basic income support</td>
<td>1/07</td>
</tr>
<tr>
<td>7/2007</td>
<td>Lechner, M., Wunsch, C.</td>
<td>Are training programs more effective when unemployment is high?</td>
<td>2/07</td>
</tr>
<tr>
<td>10/2007</td>
<td>Bernhard, S., Kurz, K.</td>
<td>Familie und Arbeitsmarkt: eine Längsschnittstudie zum Einfluss beruflicher Unsicherheiten auf die Familienreform</td>
<td>2/07</td>
</tr>
<tr>
<td>11/2007</td>
<td>Drechsler, J., Dündler, A., Bender, S., Rässler, S., Zwick, T.</td>
<td>A new approach for disclosure control in the IAB Establishment Panel: multiple imputation for a better data access</td>
<td>2/07</td>
</tr>
<tr>
<td>15/2007</td>
<td>Blien, U.</td>
<td>Expensive and low-price places to live: regional price levels</td>
<td>4/07</td>
</tr>
<tr>
<td>No.</td>
<td>Authors</td>
<td>Title</td>
<td>Year</td>
</tr>
<tr>
<td>----------</td>
<td>----------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>16/2007</td>
<td>Gartner, H. Stüber, H. Wolf, K.</td>
<td>and the agglomeration wage differential in Western Germany</td>
<td></td>
</tr>
</tbody>
</table>

Stand: 29.6.2007