RESEARCH REPORT SERIES
(Statistics #2019-07)

Synthetic Microdata for Establishment Surveys Under Informative Sampling

Hang J. Kim[1],
Joerg Drechsler[2],
Katherine J. Thompson[3]

[1]Division of Statistics and Data Science, University of Cincinnati; and Center for Statistical Research & Methodology, U.S. Census Bureau

[2]Institute for Employment Research, Nuremberg, Germany

[3]Economic Statistical Methods Division, U.S. Census Bureau

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: July 28, 2019

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. Professor Kim conducted this research as an ASA/NSF/Census Research Fellow.

Synthetic Microdata for Establishment Surveys under Informative Sampling

Hang J. Kim[1], Joerg Drechsler[2], and Katherine J. Thompson[3]

Abstract

Many agencies are currently investigating whether releasing synthetic microdata could be a viable dissemination strategy for highly sensitive data, such as business data, for which disclosure avoidance regulations otherwise prohibit the release of public use microdata. However, existing methods assume that the original data either cover the entire population or comprise a simple random sample from this population, which limits the application of these methods in the context of survey data with unequal survey weights. This paper discusses synthetic data generation under informative sampling. To utilize the design information in the survey weights, we rely on the pseudo likelihood approach when building a hierarchical Bayesian model to estimate the distribution of the finite population. Then, synthetic populations are randomly drawn from the estimated finite population density. We present the full conditional distributions of the Markov chain Monte Carlo algorithm for the posterior inference with the pseudo likelihood function. Using simulation studies, we show that the suggested synthetic data approach offers high utility for design-based and model-based analyses while offering a high level of disclosure protection. We apply the proposed method to a subset of the 2012 U.S. Economic Census and evaluate the results with utility metrics and disclosure avoidance metrics under data attacker scenarios commonly used for business data.

Keywords: disclosure risk; full synthesis; pseudo likelihood; survey weight; synthetic population

1 Introduction

Institutions and researchers often require access to detailed microdata, in addition to aggregate tabular data available to the public. Historically, agencies may selectively allow access to the *real data*

[1] Division of Statistics and Data Science, University of Cincinnati; and Center for Statistical Research & Methodology, U.S. Census Bureau

[2] Institute for Employment Research, Nuremberg, Germany

[3] Economic Statistical Methods Division, U.S. Census Bureau

Disclaimer: Any views expressed here are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied.

under strong security restrictions that respect mandated disclosure avoidance regulations. However, such access is strictly regulated and may not be convenient for many data users. Consequently, many programs produce public use microdata samples (PUMS), i.e., highly sanitized versions of real data. PUMS are often a subsample of the full survey or census with sensitive items perturbed using statistical disclosure limitation methods such as data swapping or response modification via random noise (Calvino, 2017).

However, PUMS are rarely a viable option for business data collections. First, many of the variables included in the business data are highly skewed, increasing the risk of re-identification for the largest units. As the number of items provided by the data increases, the identification risk for these outlying units in multiple dimensions sharply increases. Removing these units is not an option as these units are indispensable for accurate tabulations. Second, high correlations between the variables in the business data increase the attribute disclosure risks for some establishments. Third, sampling rates in business surveys with many large establishments having sampling weights of one are much higher than those typically found in household surveys. Thus, the inherent protection of most survey data – stemming from the fact that a potential attacker typically does not know whether the target record she is looking for is included in the sample – is substantially lower in the business survey data. Fourth, more information about businesses is publicly available, especially since larger businesses are often required by law to publish certain key statistics related to their economic activity and this information can be used to identify the units in the data. Finally, the incentives for re-identification are arguably higher, since other businesses will benefit from the knowledge gain if they successfully identify a competitor and learn about its sensitive information. For all these reasons, the practice of releasing PUMS for business data has been abandoned by the U.S. Census Bureau many years ago and disseminating *synthetic data* (Rubin, 1993) might be the only option for providing easily accessible data for general-purpose analyses. With the synthetic data approach, a model is fitted to the original data and (repeated) draws from the model are released instead of the original data.

To achieve a high level of utility, models used to generate the synthetic data should fulfill three requirements: (a) they should retain the multivariate relationships between items for data users that construct a regression or other micro-economic model, (b) they should enforce a predetermined set of edits, logical constraints developed by subject matter experts to check the original data and to ensure credibility in the output data, and (c) they should produce totals that closely correspond to the official statistics derived from the input data. With the first two requirements in mind, synthetic data approaches have been developed for government data: see, for example, Abowd et al. (2006), Hawala (2008), and Hu et al. (2018a) for household data; and Kinney et al. (2011), Drechsler (2011), and Kim et al. (2018) for business data. However, the existing methods assume that the source data covers the entire population or is a simple random sample from this population, which limits the application of these models in the context of survey data with unequal survey weights. Since survey data are typically collected using complex sampling designs, the assumption is unrealistic if the program is not a census and leads to biases in design based inferences, especially for finite population totals or means.

Our research was originally motivated by the U.S. Economic Census. Typical of many business surveys, the Economic Census comprises skewed multivariate populations that differ by sector and often by industry within sector. It is a stratified probability sample: the *measure of size* is used for stratification and many establishments are included with certainty, i.e., their probability of selection is one, whereas others are sampled with varying probabilities. This stratified simple random sampling is *informative*, i.e., the sampling distribution is not representative of the population distribution, so the sample design cannot be ignorable in finite population inference (Sugden and Smith, 1984). The informative sampling design can be partly addressed in data synthesis by including the measure of size variable as an additional predictor and releasing *synthetic samples* with the original survey weights. The motivation for including design variables is to make the sampling design ignorable when estimating the model parameters (Pfeffermann, 1993; Berg et al., 2016). From a practical perspective, this synthetic sample approach is non-optimal for a variety of reasons such as the measure of size variable being a potentially poor predictor of ancillary outcome variables, an unavailable measure of size variable, or institutional restrictions on releasing survey weights imposed when unit-identifying information may be encoded into the weights. Furthermore, as we show in our simulations, variance estimates based on the synthetic sample will be biased, as the efficiency of the stratified sampling design is not fully propagated during synthesis.

Thus, we propose a different strategy for incorporating sample design selection information into the synthesizers. Similar to Dong et al. (2014), we assume that the available data are probability samples selected using informative sampling designs. We introduce a method that generates *synthetic populations*, incorporating the survey weights into the models. The underlying idea is based on the pseudo likelihood approach (Godambe and Thompson, 1986; Pfeffermann, 1993; Fuller, 2009), which entails incorporating the sampling weights into the likelihood function. We incorporate the pseudo likelihood idea when building a hierarchical Bayesian model to estimate the distribution of the finite population. Then, the fully synthetic populations are generated by drawing repeatedly from the estimated finite population density. We derived the full conditional distributions of the Markov chain Monte Carlo algorithm for the posterior inference with the pseudo likelihood function. The inferential procedures for variance estimation with these synthetic populations and their validity are illustrated as part of the simulation study.

We compare our synthetic population approach with two approaches for generating synthetic samples. Generating synthetic populations instead of samples has several practical advantages: improved privacy protection by removing confidential information encoded in the survey weights, facilitating the use of standard statistical software for most analyses, and yielding simplified multiple imputation variance estimators. Furthermore, encoding sample design attributes into the modelling process improves the utility of the synthetic data for data users, increasing the similarity between the marginal total and their original-data counterpart.

The remainder of this paper is organized as follows. Section 2 presents our proposed model. Section 3 presents utility metrics and disclosure avoidance metrics under realistic data attacker scenarios in business data release. We explore the proposed synthetic data generation methods in Section 4 via a simulation study using a fictional population with characteristics common to establishment surveys such as skewed marginal distributions and highly correlated covariates. We apply the proposed methods to empirical data from selected industries in the 2012 U.S. Economic Census in Section 5, and conclude with general observations, recommendations, and ideas for future research in Section 6.

2 Methodology

We start this section by reviewing the business data synthesis proposed by Kim et al. (2018), which assumes the original data cover the entire finite population or they can be treated as a simple random sample (using sampling with replacement) from this population. We then present our extensions to account for informative sampling designs.

Before reviewing the methodology, we need to introduce some notation. Let $\boldsymbol{Y}_N$ be a $N \times p$ matrix comprising p variables from a finite population of size N. $\boldsymbol{Y}_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ denotes a probability sample of size n drawn from $\boldsymbol{Y}_N$, where $\boldsymbol{y}_i, i = 1, \ldots, n$, is a $(p+1) \times 1$ vector containing the p variables for unit i and a survey weight w_i, the reciprocal of the inclusion probability, which accounts for the sampling design. We assume that every unit in $\boldsymbol{Y}_N$ satisfies a set of predetermined edits, i.e., that all values in $\boldsymbol{Y}_N$ are placed within a feasible region $\mathcal{Y}$ formed by the edits. For example, the feasible region for the ratio of annual payroll to first quarter payroll will be bounded at the lower end by one and should be bounded at the upper end by some constant value that is larger than four. In reality, this delimited set of population values may not be the *true* values. For example, the population data may contain units whose observations legitimately fall outside the ratio edits, such as new businesses with exceptionally high start-up costs or failed businesses with exceptionally low profit margins. However, for modeling synthetic data, it is reasonable to generate values that conform with the edits, and to further assume that the units in $\boldsymbol{Y}_N$ are free from measurement or reporting errors. In practice, this may equate to using an edited and imputed dataset as input data. Hereafter, we refer to this consistent set of population values as the *true* values.

2.1 Methodology for simple random samples

The Dirichlet process (DP) Gaussian mixture model proposed for data synthesis in Kim et al. (2018) is a flexible tool for modeling irregularly shaped joint distributions subject to edits. The likelihood for the data under this model is written by $f(\boldsymbol{Y}_n|\Theta) = \prod_{i=1}^n f(\boldsymbol{y}_i|\Theta)$ with

$$f(\boldsymbol{y}_i|\Theta) = c_1(\mathcal{Y}, \Theta) \sum_{k=1}^{K} \eta_k \mathrm{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{y}_i \in \mathcal{Y}) \tag{1}$$

for $i = 1, \ldots, n$, where $\Theta = \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \eta_k) : k = 1, \ldots, K\}$ denotes the model parameters of the K mixture components, $\mathrm{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the normal distribution evaluated at $\boldsymbol{x}$, η_k is the weight of the k-th mixture component, $I(\cdot)$ is an indicator function which equals one if the statement is true and is zero otherwise, and $c(\mathcal{Y}, \Theta)$ is a normalizing constant such that $c_1(\mathcal{Y}, \Theta)^{-1} = \int_{\boldsymbol{x}\in\mathcal{Y}} \sum_{k=1}^{K} \eta_k \mathrm{N}(\log \boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \mathrm{d}\boldsymbol{x}$. Note that the log transformation is generally not required for Gaussian mixture modeling, but our applications use the log transformation to help account for the skewness of our data. Introducing membership indicators $\boldsymbol{z}_n = (z_1, \ldots, z_n)$, Equation (1) can be expressed by the following conditional distributions:

$$f(\boldsymbol{Y}_n|\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, \boldsymbol{z}_n) = \prod_{i=1}^{n} c_2(\mathcal{Y}, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})\ \mathrm{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) I(\boldsymbol{y}_i \in \mathcal{Y}), \tag{2}$$

$$f(\boldsymbol{z}_n|\{\eta_k\}) = \prod_{i=1}^{n} \eta_1^{I(z_i=1)} \cdots \eta_K^{I(z_i=K)}, \tag{3}$$

where $c_2(\mathcal{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma})^{-1} = \int_{\boldsymbol{x} \in \mathcal{Y}} \mathrm{N}(\log \boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{x}$. The stick-breaking representation of a truncated DP (Sethuraman 1994; Ishwaran and James 2001) is assumed as a flexible prior distribution for $\{\eta_k\}$,

$$\eta_k = \nu_k \prod_{g=1}^{k-1} (1 - \nu_g) \text{ for } k = 2, \ldots, K, \quad \eta_1 = \nu_1, \tag{4}$$

$$f(\nu_k|\alpha) \sim \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K-1, \quad \nu_K = 1, \tag{5}$$

$$f(\alpha) \sim \text{Gamma}(a_\alpha, b_\alpha), \tag{6}$$

where the prior mean of the DP concentration parameter α is a_α/b_α.

Given $\boldsymbol{Y}_n$, the synthetic sample $\widetilde{\boldsymbol{Y}}_n = \{\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n\}$ is generated from the posterior predictive distribution $f(\tilde{\boldsymbol{y}}_i|\Theta, \boldsymbol{Y}_N) = c_1(\mathcal{Y}, \{\Theta'_k\}) \sum_{k=1}^{K} \eta'_k \ \mathrm{N}(\log \tilde{\boldsymbol{y}}_i; \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k) I(\tilde{\boldsymbol{y}}_i \in \mathcal{Y})$ where η'_k (and similarly $\boldsymbol{\mu}'_k$ and $\boldsymbol{\Sigma}'_k$) denotes a random draw from the posterior distribution of η_k given $\boldsymbol{Y}_n$, i.e., $f(\eta_k|\boldsymbol{Y}_n)$. This sampling step to draw synthetic values is embedded in a Markov chain Monte Carlo (MCMC) algorithm whose other steps update the model parameters in Equations (2)–(6) (see Kim et al., 2018, for details).

It is often challenging to calculate the normalizing constant $c_2(\mathcal{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and evaluate the truncated normal distribution whose support $\mathcal{Y}$ is complex due to the edits. To avoid these calculations, the model uses a data augmentation step (Meng and Zaslavsky, 2002; O'Malley and Zaslavsky, 2008; Kim et al., 2014), which assumes that there is a hypothetical sample $\boldsymbol{Y}_{n_{\text{aug}}} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_{n_{\text{aug}}})$ of unknown size n_{aug} such that $\boldsymbol{y}_i \in \mathcal{Y}$ if $1 \leq i \leq n$ and $\boldsymbol{y}_i \notin \mathcal{Y}$ if $n+1 \leq i \leq n_{\text{aug}}$. As in O'Malley and Zaslavsky (2008) and Kim et al. (2014), we assume the likelihood of the augmented sample and the conditional distribution of its size as $f(\boldsymbol{Y}_{n_{\text{aug}}}|\Theta) = \prod_{i=1}^{n_{\text{aug}}} \sum_{k=1}^{K} \eta_k \ \mathrm{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $f(n_{\text{aug}} - n|n, \Theta, \mathcal{Y}) \sim \text{NegativeBinomial}\left(n, 1 - c_1(\mathcal{Y}, \Theta)^{-1}\right)$. This technique allows us to draw $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ from the closed form of the unconstrained multivariate normal distribution. By adopting the sampling method of O'Malley and Zaslavsky (2008), we do not need to evaluate the value of $c_2(\mathcal{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ when running the MCMC procedure.

2.2 Extensions to account for informative sampling designs

Notice that the mixture model for the sampling distribution in Equation (1) does not account for an informative sampling design. With simple random sampling, the distribution of $\boldsymbol{y}_i$ in the sample as specified in (1) is similar to the population distribution of $\boldsymbol{y}_i$, and the sampling design is ignorable with respect to model-fitting and inference (Sugden and Smith, 1984). With complex sample designs such as stratified sampling or sampling with probability proportional to size, the design is informative, i.e., the distribution of $\boldsymbol{y}_i$ in the sample will differ from that in the population. The conditions under which the sampling design is ignorable are fairly restrictive and can often be difficult to validate (Pfeffermann, 1993). Ignorability requires analytical models incorporate design variables, which are available for all units in the population and were used to construct the probability sample; examples include stratum indicators in stratified sampling or measures of size in probability proportional to size sampling. Under the stratified simple random sampling without replacement design considered in this paper, strata boundaries are fixed, and sampling rates are constant within stratum h and different between strata, i.e., $\pi_{hi} = n_h/N_h$ for all $i \in h$. This sampling design is conditionally ignorable for model-based inference when *all* design variables are included in the model (Sugden and Smith, 1984; Pfeffermann, 1993).

To reflect the informative sampling design, we propose a strategy that incorporates the survey weights when generating the synthetic data. This idea builds on the pseudo likelihood approach (Godambe and Thompson, 1986; Pfeffermann, 1993; Fuller, 2009) which assumes a superpopulation model $\mathcal{L}(\Theta; \boldsymbol{Y}_N) = \prod_{i=1}^{N} f(\boldsymbol{y}_i|\Theta)$ with unknown parameters Θ. If we would have census data $\boldsymbol{Y}_N$, the maximum likelihood estimator of Θ is calculated by solving $U(\Theta) = \sum_{i=1}^{N} u(\boldsymbol{y}_i; \Theta) = 0$ where the estimating function is $u(\boldsymbol{y}_i; \Theta) = \mathrm{d} \log f(\boldsymbol{y}_i; \Theta_k)/\mathrm{d}\Theta$. When only the sample data $\boldsymbol{Y}_n$ are available, we solve the estimating equation $\hat{U}(\Theta) = \sum_{i=1}^{n} w_i \; u(\boldsymbol{y}_i; \Theta) = 0$ to obtain a consistent estimator $\hat{\Theta}$. This pseudo likelihood approach originally motivated from a superpopulation model perspective can be similarly used in Bayesian modeling or inference. Let i^* be a pseudo population unit which is close to a unsampled unit in the population and is represented by unit i included in the sample. Then, we define the following pseudo likelihood function:

$$\mathcal{L}(\Theta; \boldsymbol{Y}^{\text{ps}}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|\Theta) f(\boldsymbol{y}_i^*|\Theta)^{w_i - 1} \tag{7}$$

where a real value $w_i \geq 1$ is the survey weight of the sampled unit i, $f(\boldsymbol{y}_i^*|\Theta) = f(\boldsymbol{y}_i|\Theta)$ if $\boldsymbol{y}_i^* = \boldsymbol{y}_i$ and zero otherwise for i such that $w_i > 1$, and $f(\boldsymbol{y}_i^*|\Theta)^0$ is defined as one. Note that Equation (7) is not a legitimate *joint density* because its integral is not one, but it is served as a *likelihood function* of the finite population parameter Θ, approximated based on sample data $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ under informative sampling. In other words, Equation (7) is used to approximately estimate the finite population parameters given the sample data rather than describing the exact joint density of the sampled units $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and the pseudo units $\{\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*\}$ whose probabilistic characteristics as random variables are not precisely defined. When deriving the full conditional densities of model parameters for the fixed values of y_i, Equation (7) is simply proportional to $f(y_i|\Theta)^{w_i}$, but we define it in this form to emphasize that y_i^* is not a real random variable rather a hypothetical unit and its integral is not equal to one. We adopt this idea in the synthetic population generation with the DP Gaussian mixture model. Under the pseudo likelihood framework, the Bayesian model estimates the posterior distribution of Θ approximately capturing the distribution of the finite population. For example, a mean vector of a mixture component $\boldsymbol{\mu}_k$ is estimated using a weighted mean $\sum_{\{i:z_i=k\}} w_i \boldsymbol{y}_i / \sum_{\{i:z_i=k\}} w_i$ rather than a simple mean, and a mixture weight η_k is estimated by using a weighed quantity, $\sum_{\{i:z_i=k\}} w_i z_i$. The impact of the pseudo likelihood will be discussed in detail in the derivation of the full posterior distributions described below.

To complete the Bayesian model, we assume the following prior distributions, most of which are standard conjugate priors adopted for computational efficiency:

$$f(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) \sim \mathrm{N}\left(\boldsymbol{\mu}_0, \frac{1}{h_0}\boldsymbol{\Sigma}_k\right), \quad k = 1, \ldots, K, \tag{8}$$

$$f(\boldsymbol{\Sigma}_k|\boldsymbol{\Phi}) \sim \mathrm{InverseWishart}(\zeta_0, \boldsymbol{\Phi}), \tag{9}$$

where $\boldsymbol{\Phi}$ is a diagonal matrix with diagonal components ϕ_j, which follow $f(\phi_j) \sim \mathrm{Gamma}(a_\phi\ , b_\phi)$ for $j = 1, \ldots, p$. We set all hyperparameters to standard values recommended in the literature: $a_\alpha = b_\alpha = a_\phi = b_\phi = 0.25$, $\zeta_0 = p - 1$, $\boldsymbol{\mu} = \mathbf{0}$, and $h_0 = 1$. We refer readers to Dunson and Xing (2009) and Kim et al. (2014) for motivation and interpretation of the selected hyperparameter values; the choices had negligible impacts on the model fitting in our simulation study and the application to the Economic Census where the sample size is moderate to large. To implement data augmentation under this pseudo likelihood framework, we have the augmented sample $\boldsymbol{Y}_{n_{\text{aug}}}$, a stacked vector comprising the observed sample $\boldsymbol{Y}_n$ and the auxiliary units $\{\boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_{n_{\text{aug}}}\}$, and $\boldsymbol{w}_{n_{\text{aug}}} = \{w_1, \ldots, w_n, w_{n+1}\ , \ldots, w_{n_{\text{aug}}}\}$ such that $w_i = 1$ for $i \geq n + 1$, i.e., the hypothetical survey weights for the auxiliary units are one.

To generate synthetic populations based on the pseudo likelihood approach for the DP Gaussian mixture model, we can use a Gibbs sampler consisting of the MCMC steps described below. The derivation of the full conditional distributions for Steps 1, 2 and 5 are provided in Section 2.2.1. Steps 3 and 4 are identical to the MCMC steps for the standard DP stick-breaking representation. The posterior distributions illustrate how the pseudo likelihood function works. In Step 1, each $\boldsymbol{y}_i$ is weighted using its survey weight w_i to estimate the center and the spread of the mixture component $\boldsymbol{\mu}_{z_i}$ and $\boldsymbol{\Sigma}_{z_i}$ to which the unit belongs. In Step 2, the DP stick is divided with the length proportional to $N_k = \sum_{i=1}^{n_{\text{aug}}} I(z_i = k)\ w_i$ rather than $n_k = \sum_{i=1}^{n_{\text{aug}}} I(z_i = k)$, so the mixture weight η_k is determined by the estimated population size of the mixture component, N_k, not by the sample size in the component, n_k. Step 6 is the same as in Kim et al. (2014) except the survey weight of one is additionally recorded in Step (iii). To generate m synthetic populations, we incorporate an extra step every T/m iterations, where T is chosen large enough to ensure independence between the draws (see details below).

After running t_{burn} iterations for burn-in, we run the MCMC algorithm for T iterations.

Step 1. For each $k = 1, \ldots, K$, update $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ by drawing from

$$f(\boldsymbol{\Sigma}_k|\cdots) \sim \text{InverseWishart}\left(\zeta_k, \boldsymbol{\Phi}_k\right) \qquad \text{and} \qquad f(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \cdots) \sim \text{N}\left(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*\right)$$

where $N_k = \sum_{i=1}^{n_{\text{aug}}} I(z_i = k)\ w_i$, $\boldsymbol{s}_k = \sum_{i=1}^{n_{\text{aug}}} I(z_i = k) w_i \log \boldsymbol{y}_i$, $\boldsymbol{T}_k = \sum_{i=1}^{n_{\text{aug}}} I(z_i = k) w_i (\log \boldsymbol{y}_i - \boldsymbol{s}_k/N_k)(\log \boldsymbol{y}_i - \boldsymbol{s}_k/N_k)^\top$, $\zeta_k = \zeta_0 + N_k$, $\boldsymbol{\Phi}_k = \boldsymbol{\Phi} + \boldsymbol{T}_k + (N_k h_0)/(N_k + h_0)(\boldsymbol{s}_k/N_k - \boldsymbol{\mu}_0)(\boldsymbol{s}_k/N_k - \boldsymbol{\mu}_0)^\top$, $\boldsymbol{\mu}_k^* = (\boldsymbol{s}_k + h_0\boldsymbol{\mu}_0)/(N_k + h_0)$ and $\boldsymbol{\Sigma}_k^* = \boldsymbol{\Sigma}_k/(N_k + h_0)$.

Step 2. Draw ν_k from $f(\nu_k|\cdots) \sim \text{Beta}(1 + N_k, \alpha + \sum_{m=k+1}^{K} N_m)$ for $k = 1, \ldots, K-1$ and set $\nu_K = 1$. Then, $\eta_1 = \nu_1$ and $\eta_k = \nu_k \prod_{m=1}^{k-1}(1 - \nu_m)$ for $k = 2, \ldots, K$.

Step 3. Draw α from $f(\alpha|\cdots) \sim \text{Gamma}(a_\alpha + K - 1, b_\alpha - \log \eta_K)$.

Step 4. For $j = 1, \ldots, p$, draw ϕ_j from $f(\phi_j|\cdots) \sim \text{Gamma}(a_\phi + K\zeta_0/2,\ b_\phi + \sum_{k=1}^{K} \sigma_{k,j}^{-2}/2)$ where $\sigma_{k,j}^{-2}$ is the j-th diagonal element of $\boldsymbol{\Sigma}_k^{-1}$.

Step 5. For each $i = 1, \ldots, n$, draw z_i from $f(z_i|\cdots) \sim \text{Categorical}(\eta'_{i1}, \ldots, \eta'_{iK})$ where $\eta'_{ik} = \eta_k \text{ N}\left(\log \boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)/\{\sum_{g=1}^{K} \eta_g \text{ N}\left(\log \boldsymbol{y}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)\}$.

Step 6. Sample the auxiliary values for data augmentation. Let $c_{\text{pass}} = c_{\text{fail}} = 0$.

(a) Draw z' from the categorical distribution, $f(z'|\boldsymbol{\eta}) = \eta_1^{I(z'=1)} \cdots \eta_K^{I(z'=K)}$.

(b) Draw $\boldsymbol{y}'$ from $f(\boldsymbol{y}'|z', \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}) \sim \text{N}(\boldsymbol{\mu}_{z'}, \boldsymbol{\Sigma}_{z'})$.

(c) If $\boldsymbol{y}' \in \mathcal{Y}$, let $c_{\text{pass}} = c_{\text{pass}} + 1$.

If $\boldsymbol{y}' \notin \mathcal{Y}$, let $c_{\text{fail}} = c_{\text{fail}} + 1$, $\boldsymbol{y}_{n+c_{\text{fail}}} = \boldsymbol{y}'$, $z_{n+c_{\text{fail}}} = z'$, and $w_{n+c_{\text{fail}}} = 1$.

Repeat the process until $c_{\text{pass}} = n$ (or $c_{\text{fail}} = n$).

Synthesis. To draw m synthetic populations $\{\tilde{\boldsymbol{Y}}_N^{(1)}, \cdots, \tilde{\boldsymbol{Y}}_N^{(m)}\}$, we implement the following steps every T/m iteration. Let $c_{\text{synt}} = 0$.

(a) Draw z' from the categorical distribution, $f(z'|\boldsymbol{\eta}) = \eta_1^{I(z'=1)} \cdots \eta_K^{I(z_i=K)}$.

(b) Draw $\boldsymbol{y}'$ from $f(\boldsymbol{y}'|z', \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}) \sim \mathrm{N}(\boldsymbol{\mu}_{z'}, \boldsymbol{\Sigma}_{z'})$.

(c) If $\boldsymbol{y}' \in \mathcal{Y}$, let $c_{\text{synt}} = c_{\text{synt}} + 1$, $\tilde{\boldsymbol{y}}_{c_{\text{synt}}} = \boldsymbol{y}'$.

Repeat the process until $c_{\text{synt}} = N$.

2.2.1 Derivations of steps 1, 2 and 5 of the MCMC algorithm

Step 1

By extending the pseudo likelihood to utilize the membership indicator, we have

$$\prod_{i=1}^{n} f(\boldsymbol{y}_i|\Theta) f(\boldsymbol{y}_i^*|\Theta)^{w_i-1} = \prod_{i=1}^{n} \int f(\boldsymbol{y}_i, z_i|\Theta)\mathrm{d}\boldsymbol{z}_i \; f(\boldsymbol{y}_i^*, z_i^*|\Theta)^{w_i-1} \tag{10}$$

where $f(\boldsymbol{y}_i^*, z_i^*|\Theta) = f(\boldsymbol{y}_i, z_i|\Theta)$ if $(\boldsymbol{y}_i^*, z_i^*) = (\boldsymbol{y}_i, z_i)$ and zero elsewhere for all i such that $w_i > 1$, and $f(\boldsymbol{y}_i^*, z_i^*|\Theta)^0$ is defined as one.

The full conditional density of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ is proportional to $\mathcal{L}(\Theta; \boldsymbol{Y}^{\text{ps}})$, $f(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)$ and $f(\boldsymbol{\Sigma}_k|\boldsymbol{\Phi})$; and evaluated with all other unknowns fixed. Because the pseudo likelihood equation in (10) contains $f(\boldsymbol{y}_i^*, z_i^*|\Theta)$, the full conditional density of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ has a positive density only when $(\boldsymbol{y}_i^*, z_i^*) = (\boldsymbol{y}_i, z_i)$ for all i whose w_i is greater than one.

Then, for the k-th mixture component, the full conditional density of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ is written by

$$
\begin{aligned}
&f(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \cdots) = f(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \boldsymbol{\Phi}, \{\boldsymbol{y}_i(=\boldsymbol{y}_i^*), z_i(=z_i^*), w_i : i = 1, \ldots, n_{\text{aug}}\}) \\
&\propto \left[\prod_{i \in A_k} |\boldsymbol{\Sigma}_k|^{-w_i/2} \exp\left\{-\frac{1}{2} w_i (\log \boldsymbol{y}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{y}_i - \boldsymbol{\mu}_k)\right\}\right] \\
&\quad \cdot |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^\top h_0 \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)\right\} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*)^\top \boldsymbol{\Sigma}_k^{*-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*) \\
&\quad \cdot |\boldsymbol{\Sigma}_k|^{-(\zeta_0 + p + 1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\boldsymbol{\Phi} \boldsymbol{\Sigma}_k)\right\}.
\end{aligned}
$$

where A_k is the index set of all units $i = 1, \ldots, n_{\text{aug}}$ such that $z_i = k$,

$$
N_k = \sum_{i \in A_k} w_i, \quad S_k = \sum_{i \in A_k} w_i \log \boldsymbol{y}_i, \quad \boldsymbol{T}_k = \sum_{i \in A_k} w_i \left(\log \boldsymbol{y}_i - \frac{S_k}{N_k}\right)\left(\log \boldsymbol{y}_i - \frac{S_k}{N_k}\right)^\top
$$

$$
\boldsymbol{\mu}_k^* = \frac{N_k \frac{S_k}{N_k} + h_0 \boldsymbol{\mu}_0}{N_k + h_0}, \quad \text{and} \quad \boldsymbol{\Sigma}_k^* = \frac{1}{N_k + h_0} \boldsymbol{\Sigma}_k.
$$

The terms inside the exponents can be rewritten as

$$
\begin{aligned}
&\sum_{i \in A_k} w_i (\log \boldsymbol{y}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\log \boldsymbol{y}_i - \boldsymbol{\mu}_k) + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^\top h_0 \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0) \\
&= \sum_{i \in A_k} w_i (\log \boldsymbol{y}_i - S_k/N_k)^\top \boldsymbol{\Sigma}_k^{-1} (\log \boldsymbol{y}_i - S_k/N_k) + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*)^\top \boldsymbol{\Sigma}_k^{*-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*) \\
&\qquad - \left(\frac{N_k \frac{S_k}{N_k} + h_0 \boldsymbol{\mu}_0}{N_k + h_0}\right)^\top (N_k + h_0) \boldsymbol{\Sigma}_k^{-1} \frac{N_k \frac{S_k}{N_k} + h_0 \boldsymbol{\mu}_0}{N_k + h_0} + \left(\frac{S_k}{N_k}\right)^\top N_k \boldsymbol{\Sigma}_k^{-1} \left(\frac{S_k}{N_k}\right) + \boldsymbol{\mu}_0^\top h_0 \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_0 \\
&= (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*)^\top \boldsymbol{\Sigma}_k^{*-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*) + \sum_{i \in A_k} w_i (\log \boldsymbol{y}_i - S_k/N_k)^\top \boldsymbol{\Sigma}_k^{-1} (\log \boldsymbol{y}_i - S_k/N_k) \\
&\quad + \frac{N_k h_0}{N_k + h_0} (S_k/N_k - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_k^{-1} (S_k/N_k - \boldsymbol{\mu}_0).
\end{aligned}
$$

Therefore,

$$f(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\cdots) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\right)^\top \boldsymbol{\Sigma}_k^{*-1}\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\right)\right\}$$

$$\cdot\ |\boldsymbol{\Sigma}|^{-(\zeta_0+N_k+p+1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Phi} + \boldsymbol{T}_k + \frac{N_k h_0}{N_k + h_0}\left(\frac{S_k}{N_k} - \boldsymbol{\mu}_0\right)^\top\left(\frac{S_k}{N_k} - \boldsymbol{\mu}_0\right)\right\}\boldsymbol{\Sigma}_k^{-1}\right]$$

$$\text{where} \qquad \boldsymbol{T}_k = \sum_{i\in A_k} w_i\left(\log \boldsymbol{y}_i - \frac{S_k}{N_k}\right)^\top\left(\log \boldsymbol{y}_i - \frac{S_k}{N_k},\right)$$

and the first term is the kernel of a normal distribution and the second term is that of the inverse-Wishart distribution.

Therefore, $f(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\cdots) = f(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k,\cdots)f(\boldsymbol{\Sigma}_k|\cdots)$ where

$$f(\boldsymbol{\Sigma}_k|\cdots) \sim \mathrm{InverseWishart}\left(\zeta_0 + N_k, \boldsymbol{\Phi} + \boldsymbol{T}_k + \frac{N_k h_0}{N_k + h_0}\left(\frac{S_k}{N_k} - \boldsymbol{\mu}_0\right)\left(\frac{S_k}{N_k} - \boldsymbol{\mu}_0\right)^\top\right)$$

$$f(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k,\cdots) \sim \mathrm{N}\left(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*\right)$$

Step 2

Similarly to Step 1, the full conditional density of $\{\nu_k\}$ has a positive density when $z_i = z_i^*$ for all i. Therefore,

$$f(\nu_1,\ldots,\nu_{K-1}|\cdots) = f(\nu_1,\ldots,\nu_{K-1}|\alpha, \{z_i(=z_i^*), w_i : i = 1,\ldots,n_{\mathrm{aug}}\})$$

$$\propto \prod_{i=1}^{n_{\mathrm{aug}}} \eta_1^{w_i I(z_i=1)} \cdots \eta_K^{w_i I(z_i=K)} \prod_{k=1}^{K-1}(1-\nu_k)^{\alpha-1}$$

$$\propto \nu_1^{\sum_{\{i:z_i=1\}} w_i}(1-\nu_1)^{\sum_{g=2}^{K}\sum_{\{i:z_i=g\}} w_i}\ \nu_2^{\sum_{\{i:z_i=2\}} w_i}(1-\nu_2)^{\sum_{g=3}^{K}\sum_{\{i:z_i=g\}} w_i}$$

$$\cdots\ \nu_{K-1}^{\sum_{\{i:z_i=K-1\}} w_i}(1-\nu_{K-1})^{\sum_{\{i:z_i=K\}} w_i}\prod_{k=1}^{K-1}(1-\nu_k)^{\alpha-1}$$

$$\propto \prod_{k=1}^{K-1}\nu_k^{\sum_{\{i:z_i=k\}} w_i}(1-\nu_k)^{\alpha+\sum_{g=k+1}^{K}\sum_{\{i:z_i=g\}} w_i-1}$$

Therefore, for $k=1,\dots,K-1$,

$$f(\nu_k|\cdots)\sim \text{Beta}\left(1+N_k,\alpha+\sum_{g=k+1}^{K}N_g\right)\qquad \text{where}\quad N_k=\sum_{\{i:z_i=k\}}w_i.$$

Step 5

Because the pseudo likelihood has a positive density when $z_i=z_i^*$ for all units i whose w_i is greater than one, we need to update z_i and z_i^* simultaneously. To do this, we draw z_i from its full conditional density $f(z_i|\cdots)=f(z_i|\boldsymbol{y}_i,\Theta,\boldsymbol{\eta})\propto f(\boldsymbol{y}_i|z_i)f(z_i|\boldsymbol{\eta})$ which is written by

$$\Pr(z_i=k|\cdots)=\frac{\eta_k\ N(\boldsymbol{y}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{g=1}^{K}\eta_g\ N(\boldsymbol{y}_i;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)}.$$

Then, the value of z_i^* is automatically updated by $z_i^*=z_i$.

2.3 Variance estimation for multiply imputed synthetic datasets

Similar to multiple imputation for nonresponse, the release of multiple synthetic datasets enables the analyst to properly account for the extra uncertainty from the synthesis process. In this section we review the inferential procedures for the different synthesis approaches.

Let θ be a parameter of interest, such as a population mean, total, or correlation coefficient, and $\hat{\theta}^{(l)}$ be its estimate constructed from the l-th synthetic data, $l=1,\dots,m$. For synthetic samples $\{\tilde{\boldsymbol{Y}}_n^{(1)}, \ldots, \tilde{\boldsymbol{Y}}_n^{(m)}\}$, the following quantities are required for inference: $\hat{\theta} = \sum_{l=1}^m \hat{\theta}^{(l)}/m$, $b_m = \sum_{l=1}^m (\hat{\theta}^{(l)} - \hat{\theta})^2/(m-1)$, and $\bar{u}_m = \sum_{l=1}^m u^{(l)}/m$ where $\hat{\theta}^{(l)}$ is an estimate given the l-th synthetic sample and $u^{(l)}$ denotes the estimate for the sampling variance of $\hat{\theta}^{(l)}$. For synthetic populations $\{\tilde{\boldsymbol{Y}}_N^{(1)}, \ldots, \tilde{\boldsymbol{Y}}_N^{(m)}\}$, the sampling variance is zero, i.e., $u^{(l)} = 0$. Hence the required quantities are $\hat{\theta} = \sum_{l=1}^m \hat{\theta}^{(l)}/m$ and the between-variance $B_m = \sum_{l=1}^m (\hat{\theta}^{(l)} - \hat{\theta})^2/(m-1)$ where $\hat{\theta}^{(l)}$ is an estimate given the l-th synthetic population. Independent of the synthesis strategy, the final point estimate for θ is always $\hat{\theta}$. However, the estimator for the variance will differ as we review below.

2.3.1 Partial synthesis

With partial synthesis as proposed by Little (1993), only some of the variables from the original data are replaced with synthetic values, i.e., the synthetic data are generated from the *conditional* predictive distribution given some observed values, $f(\tilde{\boldsymbol{y}}_{i,a}|\boldsymbol{y}_{i,-a}, \Theta)$ where index a denotes the set of synthesized variables, $-a$ denotes the variables that remain unchanged, and Θ contains the parameters governing the distribution of $f(\boldsymbol{y}_{i,a}|\boldsymbol{y}_{i,-a}, \Theta)$. Statistical agencies can release $\{\tilde{\boldsymbol{y}}_{i,a}\}$ only or $\{\tilde{\boldsymbol{y}}_{i,a}, \boldsymbol{y}_{i,-a}\}$ if $\boldsymbol{y}_{i,-a}$ is considered safe to be released. For partially synthetic samples, Reiter (2003) suggested estimating the variance of $\hat{\theta}$ as

$$\widehat{\mathrm{V}}_{\text{part}} = \frac{1}{m} b_m + \bar{u}_m. \tag{11}$$

Inference of θ is made using a t-distribution with degrees of freedom $\nu_p = (m-1)\{1 + \bar{u}_m/(b_m/m)\}^2$.

2.3.2 Full synthesis

The fully synthetic data approach as proposed by Rubin (1993) is closely related to the idea of multiple imputation for nonresponse. All records not included in the original sample are treated as missing data, which are multiply imputed to generate synthetic populations. Simple random samples from these populations are generated in the second step to be released to the public. To illustrate, we assume the original data comprises a sample of size n from a population of size N (Drechsler, 2011).

If some variables X_N for the whole population are available in data synthesis, the synthetic datasets can be generated in two steps. First, construct m imputed populations by multiply imputing the unsampled units $\tilde{Y}_{N-n}$ independently from the posterior predictive distribution $f(Y_{N-n}|X_N, Y_n)$ given the sampled units Y_n and X_N. Second, take a simple random sample from each imputed population and release them to the public. For these fully synthetic samples, Raghunathan et al. (2003) suggested estimating the variance as

$$\widehat{\mathrm{V}}_{\text{full}} = \left(1 + \frac{1}{m}\right) b_m - \bar{u}_m. \tag{12}$$

If no X variables are available as in our case, the synthetic data can be generated by drawing from the *unconditional* predictive distribution, $f(\tilde{Y}_{N-n}|Y_n) = \int f(\tilde{Y}_{N-n}|\Theta) f(\Theta|Y_n) \mathrm{d}\Theta$. In this case, the variance estimator in (12) is biased if all records in Y_N (i.e., not only Y_{N-n} but also Y_n) are synthesized, but the bias will be negligible if the sampling rate of the original data is small; see Drechsler (2018) for further discussion. We note that an alternative variance estimator for fully synthetic data has been proposed by Raab et al. (2017). This estimator, which can be applied if no X variable has been used and all records have been synthesized, only depends on $\bar{u}_m$. This is attractive since it allows to estimate the variance even if only one synthetic copy has been released. However, it is not clear how to apply this estimator if the synthetic data comprise the entire population as in our case since $\bar{u}_m$ should be zero in such a situation. For more discussion regarding the advantages and disadvantages of the different variance estimators for fully synthetic data, we refer the reader to Drechsler (2018).

When generating synthetic populations following the procedures in Section 2.2, we un-complex the sampling process. Thus, the uncertainty introduced by the sampling process is already reflected in the spread between the synthetic populations. The variance estimator for synthetic populations is introduced by Raghunathan et al. (2003, Section 4.1) as

$$\widehat{\mathrm{V}}_{\mathrm{pop}} = \left(1 + \frac{1}{m}\right) B_m. \tag{13}$$

In the equation, the quantity B_m/m estimates the uncertainty introduced by releasing a *finite* number m of synthetic datasets. In a hypothetical scenario where an infinite number of synthetic populations are released, i.e., $m \to \infty$, the uncertainty in estimating B_∞ – measured by B_m/m – disappears, and the model uncertainty in the predictive distribution is measured by B_∞.

3 Evaluation Criteria: Analytical Validity and Disclosure Risk

Any disclosure limitation method needs to strike a balance between preserving analytical validity and offering a sufficient level of data protection. In this section, we review the criteria that we use to evaluate our proposed methods along these two dimensions.

3.1 Analytical validity criteria

Given that the synthetic data are proposed as an alternative to PUMS, we put forward a general-purpose view of synthetic data utility. Most synthetic data evaluations in the literature (Hu et al., 2018b; Kim et al., 2018; Snoke et al., 2018) assume that the typical data user is primarily interested in modeling and assessing relationships between variables. This is certainly true of the economists and other academic researchers that request access to the U.S. Census microdata. However, valid design-based inferences for finite population totals are equally important, when it comes to business data. Since most establishment surveys are designed to measure economic activity, main users such as government policymakers, individual businesses, trade associations, and economic development agencies want to estimate the finite population parameters to understand the current state of different economic sectors and to make informed decisions. Furthermore, ratios of industry totals provide useful and important insights into the current economy. For example, in industries where the majority of establishments operate year-round, the ratio of annual payroll to 1st quarter payroll should be approximately four. Consequently, smaller values of the industry-average ratios would be indicative of economic decline over the year within the industry, whereas larger ratios would be indicative of economic growth. Large ratios of total sales to total payroll can be indicative of industry growth, occasionally booms. Another example is that wage per employee ratios are used in both economic and demographic analyses to assess labor conditions within industries. Under this broader view of synthetic data designed to reach a larger class of potential data users, we look at two dimensions of analytical validity: the analytical completeness of the data in terms of accurate estimates for the marginal totals, which would in turn provide accurate ratios, and the analytical validity when modeling multivariate relationships between the variables included in the data. For the former, we compare estimated totals based on the synthetic data to their input data counterparts. For the latter, we compare regression parameters from various regression models.

3.2 Disclosure risk criteria

The literature on risk measures for fully synthetic datasets is sparse. Re-identification experiments such as those considered in Drechsler and Reiter (2008) and Reiter and Mitra (2009) are not meaningful for this context since all records are drawn from a model, breaking the link between original and synthetic records. Reiter et al. (2014) propose some risk measures for fully synthetic data, however the computational requirement is so demanding that the measures are difficult to compute in realistic applications such as the ones we consider in this paper.

With many economic data collections, it may not be difficult to identify the largest units in certain industries in the dataset using the attributes available and some common knowledge of the industry. Indeed, excluding these units would severely bias the totals. That said, their reported attributes are not always publicly available and in fact are often protected by mandate, e.g., Title 13. Hence, risks related to attribute disclosure are typically more relevant than risks of re-identification. Thus, we focus on a simple measure of attribute disclosure, which is well established in the business data context: the distance between the true value for some sensitive attribute of the largest unit in the original database and the estimate a potential attacker, possibly with background knowledge, might obtain from the synthetic data.

Obviously, concerns about privacy protection are not restricted to the largest units. However, the smaller businesses tend to report similar values especially when rounding is employed, so that the individual businesses are less identifiable than their larger counterparts, especially in a survey sampling context.

To define our measure of attribute disclosure, we need to introduce some notation. Let L_j be the largest value in the input data for item (attribute) j and let $\hat{L}_j$ be the attacker's estimate for this value defined below. We use the absolute relative difference $ARD_j = |\hat{L}_j - L_j|/L_j$ for each item as the measure of disclosure risk. We consider two attacker scenarios:

> Scenario 1: A data attacker has access to the synthetic data only. In this case, we assume the attacker would use the average of the largest values from each of the multiple synthetic data sets, i.e., $\hat{L}_j = \sum_{l=1}^{m} \max_i(\tilde{y}_{ij}^{(l)})/m$, where $\tilde{y}_{ij}^{(l)}$ represents the j-th item for unit i in the l-th synthetic dataset.
>
> Scenario 2: The attacker knows that he is the second largest unit in the database and uses the synthetic data to bound the value for the largest establishment. Let S_j be the true value of the second largest unit in the original input data, $\hat{T}_j$ be the estimated population total for item j obtained from the synthetic data, and $\hat{L}_j = \hat{T}_j - S_j$ be the attacker's estimate of the largest value. The approach can easily be extended to allow for collaborators. Let C_j be the sum of the true values provided by the collaborator businesses. The estimate will then be $\hat{L}_j = \hat{T}_j - S_j - C_j$. This scenario is consistent with the p-percent rule often used by federal

statistical agencies to assess disclosure risk in magnitude tables (see FCSM, 2005, Chapter IV). These metrics are limited to univariate analyses. One could imagine that an attacker has advanced knowledge of an atypical unit, for example, a small or medium sized business with an unusually high wage-to-employee ratio for a specific industry, which is at higher risk of attribute disclosure. However, we emphasize that whether the data are perfectly protected to be released is not the goal of this paper to establish. The focus is on evaluating the relative performance of different synthesis strategies using risk measures that are illustrative of the high level of protection offered through data synthesis.

3.3 Evaluating synthetic data approaches

If the calculated values of the suggested analytical validity measure and disclosure risk measures are similar in a comparative study, the following discussion points should be additionally considered. First, with partially synthetic data re-identificaion is still possible, i.e., a record in the synthetic data is identified as a real population unit, which will typically also lead to an increase in the risk of attribute disclosure. As pointed out above, re-identification risk is not meaningful in fully synthetic data because all records in the synthetic data are drawn from a model, breaking the link between original units and synthetic units. Therefore, if the calculated values of the suggested analytical validity and disclosure risk measures are similar for fully and partially synthetic data, the fully synthetic datasets should be arguably preferred. Second, there are some studies suggesting that the release of the original survey weights itself may increase the disclosure risk (for example, Willenborg and De Waal, 1996; Fienberg, 2010; Cox et al., 2011). In this spirit, an agency may prefer the release of synthetic populations over the release of synthetic samples with original survey weights attached. Lastly, synthetic populations can be favored over synthetic samples in terms of users' convenience. With synthetic samples with the original survey weights, the users need to estimate the sampling variance accounting for the complex sampling design used for the original data, which may or may not be easy in practice. However, the release of the synthetic populations facilitates the use of standard statistical software for most analyses since only the between-variance in the analysis outcomes for the multiple populations needs to be calculated.

4 Simulation Study

We conduct a repeated simulation study to assess the statistical properties of the synthetic data models discussed in Section 2. The simulation consists of four steps: (a) generating a finite population with distributional properties commonly found in business data, (b) drawing random samples from this population using a stratified design representative of sampling designs typically employed in business data contexts, (c) synthesizing the data using different strategies discussed below, and (d) evaluating the synthesized data in terms of their analytical validity and disclosure protection. Steps (b)–(d) are repeated 400 times to enable us to assess the validity of the obtained inferences. In the following, we discuss each step in more detail.

4.1 Step 1: Generating the population

The simulated population has N=100,000 records of $\boldsymbol{y}_i$, each record consisting of four survey variables $\boldsymbol{y}_{i,\text{surv}} = (y_1, y_2, y_3, y_4)$ and a measure of size variable s_i. We assume the survey variables $\boldsymbol{y}_{i,\text{surv}}$ follow a truncated normal mixture distribution with $K^* = 3$ components on the log scale under edit constraints whose distribution is written by

$$\begin{aligned} f(\boldsymbol{y}_{i,\text{surv}} | \{\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*\}, z_i^*) &= \ c_2(\mathcal{Y}, \boldsymbol{\mu}_{z_i^*}^*, \boldsymbol{\Sigma}_{z_i^*}^*)\ \text{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_{z_i^*}, \boldsymbol{\Sigma}_{z_i^*}) I(\boldsymbol{y}_i \in \mathcal{Y}) \\ f(z_i^*) &\sim \text{Categorical}(\eta_1^*, \cdots, \eta_{K^*}^*). \end{aligned}$$

The star symbols in the parameters emphasize that these simulation values do not necessarily exactly match to the parameter estimates in Section 2.2 because the synthetic model assumes a Dirichlet process mixture model, while this simulation data assumes a finite number of normal mixture components. We generate the simulated population $\boldsymbol{Y}_N$ by integrating over the following steps:

Start with $i = 1$.

1. Generate z_i^* from Categorical$(\eta_1^*, \eta_2^*, \eta_3^*)$ where the simulation mixture weights are set as $(\eta_1^*, \eta_2^*, \eta_3^*) = (0.25, 0.6, 0.15)$.

2. Draw $\log \tilde{\boldsymbol{y}}$ from $\mathrm{N}(\boldsymbol{\mu}_{z_i^*}, \boldsymbol{\Sigma}_{z_i^*})$ where the simulation parameters used are $\boldsymbol{\mu}_1 = (2, 2, 3, 2)^\top, \boldsymbol{\mu}_2 = (6, 5, 6, 6)^\top, \boldsymbol{\mu}_3 = (8, 7, 9, 10)^\top$, and

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2.0 & 1.3 & 0.8 & 0.8 \\ 1.3 & 1.0 & 0.6 & 0.7 \\ 0.8 & 0.6 & 1.0 & 0.8 \\ 0.8 & 0.7 & 0.8 & 1.0 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.5 & -0.3 & 0.5 & 0.8 \\ -0.3 & 1.5 & 0.57 & 0.7 \\ 0.5 & 0.57 & 1.2 & 0.64 \\ 0.8 & 0.7 & 0.64 & 1.0 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_3 = \begin{bmatrix} 2.0 & 1.3 & 0.9 & 1.5 \\ 1.3 & 1.2 & 0.5 & 0.7 \\ 0.9 & 0.5 & 0.7 & 0.6 \\ 1.5 & 0.7 & 0.6 & 1.7 \end{bmatrix}.$$

3. Check if $\tilde{\boldsymbol{y}} \in \mathcal{Y}$, i.e., it satisfies the ratio edits assumed as

	y_1/y_2	y_1/y_3	y_1/y_4	y_2/y_3	y_2/y_4	y_3/y_4
Lower bound	0.367	0.082	0.007	0.006	0.007	0.007
Upper bound	148.41	20.09	148.41	2.72	148.41	148.41

   If $\tilde{\boldsymbol{y}}$ satisfies all ratio edits, let $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}$ and let $i = i + 1$. Otherwise, go to Step 2 and draw $\tilde{\boldsymbol{y}}$ again.

4. Repeat Steps 2 and 3 until $i \leq N$.

Figure 1 shows the bivariate distributions for all survey variables where the solid lines indicate ratio edits imposed when generating the true population as well as at the synthesis stage. the imposed ratio edits, the model parameters for generating the survey variables, and the measure of size variable.

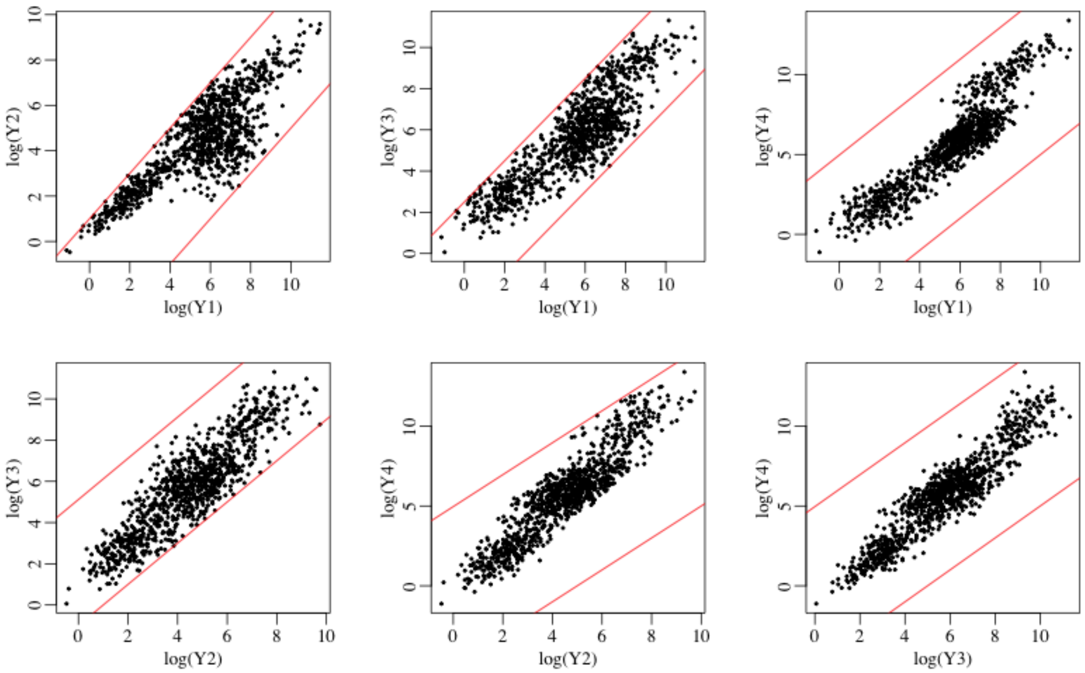


Figure 1: Bivariate distributions of the simulated population displayed in the log scale. The solid lines indicate ratio edits.

Table 1: Correlation coefficients in the simulated population.

	Y_1	Y_2	Y_3	Y_4
Y_1	1.00	0.76	0.66	0.80
Y_2	0.76	1.00	0.62	0.58
Y_3	0.66	0.62	1.00	0.62
Y_4	0.80	0.58	0.62	1.00

The ratio edits $Y_j/Y_{j'} \leq U_{jj'}$, i.e., $\log Y_j \leq \log U_{jj'} + \log Y_{j'}$ on the log scale, are satisfied by all units in the simulated population for all $j \neq j'$.

Table 1 presents the correlations between the survey variables. This simulated population resembles a *typical* economic population, with highly skewed marginal distributions and strong association among variables.

4.2 Step 2: Drawing random samples from the population

Economic surveys are often designed to produce reliable estimates of population totals, and stratified sampling designs are commonly applied to improve the efficiency of the estimates and ensure sufficient sample sizes in subdomains of interest.

In this simulation study, the measure of size variables were randomly generated by $s_i = \exp(q_i) + \varepsilon_i$ where $q_i \sim \mathrm{N}(\log y_{i1}, 1)$ and $\varepsilon_i \sim \mathrm{N}(0, (\log y_{i1} - \min_j y_{j1})^2/10000)$. The table below shows the five number summary of the simulated measure of size variables.

Min.	1st Q.	Median	Mean	3rd Q.	Max.
0.007	52.9	346.8	2972.2	1399.0	2376532.0

We assume that the measure of size variable is only used for the sampling design and would not be released to the public. We employ stratified simple random sampling without replacement, using Neyman allocation, which is typical of many business surveys. Following the survey design principles outlined in Smith (2013, Chapter 5, pp. 165–201), the Neyman allocation minimizes the overall sample size for a fixed coefficient of variation on the measure of size variable, using the same variable for stratification and allocation. With skewed populations, a small number of units will account for a large proportion of the population total. To minimize bias in the estimated total, the largest cases are included in the sample with certainty, i.e., their probability of selection equals one.

For the simulation, the certainty strata boundary was determined with the Lavallée-Hidiroglou stratification algorithm (Lavallée and Hidiroglou, 1988) and noncertainty stratum boundaries were determined via the *cum-root-f* rule (Dalenius and Hodges Jr, 1959). Setting the target coefficient of variation for the measure of size variable s to 0.02 and assuming Neyman allocation (equal costs per unit) yielded a sampling design comprising five strata and a fixed sample size of $n = 6,300$. Table 2 presents the sampling fractions (n_h/N_h) and the realized sample sizes (n_h) by stratum.

This stratified design is typical of economic surveys: it includes *large* establishments with certainty (stratum 1), it samples *medium sized* establishments with high sampling rates (stratum 2), and

Table 2: Stratum sizes and sampling rates for the simulation study.

Stratum	1	2	3	4	5
n_h	1410	476	109	286	4019
n_h/N_h	1.000	0.235	0.062	0.061	0.045

it samples the remaining cases with very low sampling rates (strata 3 through 5). Note that the sampled establishments with the largest survey weights will contribute proportionally little to the overall tabulated totals, typical of business survey designs (Thompson and Oliver, 2012).

4.3 Step 3: Synthesizing the data

We evaluate three synthesis methods: our proposed synthetic *population* method (denoted by PopSyn hereafter) introduced in Section 2.2 and two synthetic *sample* methods based on the likelihood function in Equation (1), one which completely ignores the sampling design (denoted by SampleSyn1 hereafter) and the other using the measure of size variable as an additional predictor (denoted by SampleSyn2 hereafter).

SampleSyn1 is the original DP Gaussian mixture synthesis model of Kim et al. (2018), which only uses the survey variables, i.e., $\boldsymbol{Y}_n = \{\boldsymbol{y}_{i,\text{surv}} : i = 1, \ldots, n\}$. This method draws the multiple synthetic samples $\widetilde{\boldsymbol{Y}}_n^{(l)} = \{\tilde{\boldsymbol{y}}_1^{(l)}, \tilde{\boldsymbol{y}}_2^{(l)}, \cdots, \tilde{\boldsymbol{y}}_n^{(l)}\}$, $l = 1, \ldots, m$, independently from the posterior predictive distribution $f(\tilde{\boldsymbol{y}}_i|\boldsymbol{Y}_n) = \int f(\tilde{\boldsymbol{y}}_i|\Theta)f(\Theta|\boldsymbol{Y}_n)\mathrm{d}\Theta$. Since this method does not reflect any sampling information, $f(\Theta|\boldsymbol{Y}_n)$ estimates the model parameters of the joint distribution of the sampled units, rather than that of the population units.

SampleSyn2 implements the methodology of Kim et al. (2018) including the design variable s_i in addition to the survey variables, i.e., $\boldsymbol{Y}_n = \{(\boldsymbol{y}_{i,\text{surv}}, s_i) : i = 1, \ldots, n\}$. The motivation for including the design variable is to make the sampling design ignorable when estimating the model parameters (Pfeffermann, 1993; Berg et al., 2016). After fitting the Gaussian mixture model in (1) to $\{\boldsymbol{y}_i = (\boldsymbol{y}_{i,\text{surv}}, s_i) : i = 1, \ldots, n\}$, partially synthetic samples for the survey variables are drawn from the conditional posterior predictive distribution given the design variable, $f(\tilde{\boldsymbol{y}}_{i,\text{surv}}|s_i, \Theta)f(\Theta|\boldsymbol{Y}_n)$.

This synthesis strategy assumes that the agency would disseminate the synthetic survey variables together with original design weights, which are still valid given that the data were generated conditional on the original measure of size variable s_i.

For PopSyn, we use the suggested method described in Section 2.2 to generate synthetic populations of size $N = 100{,}000$. For each synthesis method we create $m = 10$ synthetic datasets by drawing every 200th iteration of a MCMC chain after a burn-in period of 2,000 iterations. Using trace plots and autocorrelation plots, we confirmed that the burn-in period is sufficiently long to ensure convergence to the stationary distribution and the draws from the chains are sufficiently far apart to ensure independence between the multiple synthetic datasets.

4.4 Step 4: Evaluating the analytical validity and the risk of disclosure

As discussed in Section 3.1, we evaluate two dimensions of analytical validity. First, for the design-based measure of economic activity, we estimate the population totals for all survey variables included in the simulation study. Second, as an example of an econometric model, we look at the regression coefficients in the following linear regression model:

$$\log Y_1 = \beta_0 + \beta_1 \log Y_2 + \beta_2 \log Y_3 + \beta_3 \log Y_4 + \varepsilon.$$

Regardless of any assumptions related to the distribution of the random error ε, we can compare the least square estimate computed from each synthetic dataset.

Two questions arise in this context: (i) Should the design weights be used when estimating the regression coefficients in the linear regression model? (ii) How should we estimate the variance of the estimated totals?

Starting with the first question, we compute the weighted least square estimate (WLS) for the synthetic samples (SampleSyn1 and SampleSyn2), whereas the ordinary least square estimate (OLS) is used for the synthetic populations. Note, however, that the SampleSyn1 approach, implicitly assuming simple random sampling, has constant weights for all synthesized units, leading to identical results for WLS and OLS.

The answer to the second question depends on the synthesis strategy. The outputs of SampleSyn1 are not associated with the original survey weights, so the estimate for the sampling variance $u^{(l)}$ for the l-th synthetic data set is computed assuming simple random sampling. Because the synthetic units in this approach do not match with any original units or original survey weights, the outputs are considered as fully synthetic samples and Equation (12) is used for variance estimation. Since the outputs of SampleSyn2 are partially synthetic samples where the original measure of size variables s_i are linked to the synthetic survey variables $\tilde{\boldsymbol{y}}_{i,\text{surv}}$, we use the variance estimation equation in (11) when estimating variances based on SampleSyn2. For this design, the sampling design has to be taken into account when computing the within variance $\bar{u}_m$. Note that stratum indicators are typically not released due to confidentiality concerns under currently employed data access regimes. Because we cannot anticipate whether this would change if synthetic data would be released instead, we consider two alternative design-based variance estimators for $u^{(l)}$ for SampleSyn2. If stratum indicators can be included in the partially synthetic samples along with the design weights, the stratified sampling design-based variance estimator for a total $\hat{V}(\hat{t})_{\text{str}} = \sum_h (1 - n_h/N_h) N_h^2 s_h^2 / n_h$ calculated from the l-th synthetic sample can be used for $u^{(l)}$, where n_h and N_h are the within-strata sample and population sizes, respectively, and s_h^2 is the sample-based within-strata variance. This estimator appropriately accounts for finite-population sampling and differing strata means. However, if such design information is deemed sensitive, then the more conservative Hansen-Hurwitz variance estimator $\hat{V}(\hat{t})_{\text{HH}} = 1/n \sum_i (w_i y_i - \hat{y})^2$ can be used, where $n = \sum_h n_h$ and $\hat{y} = \sum_i w_i y_i$ calculated from each synthetic sample. These two variance estimators represent best and worst-case scenarios for inference. We note that other design-based variance estimators could be used if the strata indicators are deemed sensitive. For example, surveys could provide design-effects for selected items (Kish and Frankel, 1974), generalized variance functions (Valliant, 1987), or more aggregated standard error computation units that allow replicate variance estimation for a stratified sample, e.g., U.S. Census Bureau (2006). If these strategies would be employed, we would expect the inferential properties of the totals to lie somewhere between the two scenarios. For the fully synthetic populations from PopSyn, we can simply use the variance equation in (13). As the variance of the estimated total only depends on the variance between the synthetic datasets B_m, the analyst no longer needs to worry about the original sampling design.

For the risk evaluations we compute the absolute relative difference measure under Scenarios 1 and 2 discussed in Section 3.2. As pointed out above, the whole process of sampling from the population, generating three different versions of synthetic data, and evaluating the generated data is repeated 400 times.

4.5 Results

4.5.1 Analytical validity

Table 3 contains the results for the estimated totals. In the table, θ denotes the true population totals, while $\hat{\theta}$ denotes their estimates from the different synthetic datasets. Summarizing over 400 repeated simulations, Table 3 presents the empirical mean $\mathrm{E}(\hat{\theta})$, the empirical variance $\mathrm{V}(\hat{\theta})$, the mean squared error $\mathrm{MSE}(\hat{\theta})$, and the empirical mean of the variance estimates $\mathrm{E}\{\hat{\mathrm{V}}(\hat{\theta})\}$ for each synthesis design.

The results from Table 3 can be summarized as follows. First, the synthetic generator implemented with SampleSyn1 (originally proposed by Kim et al., 2018) results in extremely positively biased estimates of the population totals, since it fails to incorporate informative sampling design features. Second, incorporating the survey design features in the synthesis models yields estimates of the population totals for synthetic samples (SampleSyn2) and synthetic populations (PopSyn) that are within 4% of the corresponding population totals for all variables. A possible source of the deviation in SampleSyn2 is the relatively limited design information contained in the measure of size

Table 3: Inference for population totals based on the different synthesis strategies (unit: 10^6 for point estimates and 10^{12} for MSE and variance estimates). For SampleSyn2, $\mathrm{E}\{\hat{\mathrm{V}}(\hat{\theta})\}_{\mathrm{str}}$ computes the within-variance $u^{(l)}$ using the variance estimate appropriate for stratified simple random sampling without replacement, while $\mathrm{E}\{\hat{\mathrm{V}}(\hat{\theta})\}_{\mathrm{HH}}$ uses the Hansen-Hurwitz variance estimator.

		Y_1	Y_2	Y_3	Y_4
Simulation Population	θ	177.9	44.7	220.9	716.8
SampleSyn1	$\mathrm{E}(\hat{\theta})$	1445.3	218.6	1008.6	5819.9
	$\mathrm{MSE}(\hat{\theta})$	1838487.3	32096.1	717691.3	34985282.8
	$\mathrm{V}(\hat{\theta})$	232893.9	1885.6	97514.8	8967013.0
	$\mathrm{E}\{\hat{\mathrm{V}}(\hat{\theta})\}$	18187.9	148.2	5986.0	570918.9
SampleSyn2	$\mathrm{E}(\hat{\theta})$	184.0	46.1	223.0	745.3
	$\mathrm{MSE}(\hat{\theta})$	68.3	9.1	68.2	1550.0
	$\mathrm{V}(\hat{\theta})$	31.1	7.2	64.2	739.7
	$\mathrm{E}\{\hat{\mathrm{V}}(\hat{\theta})\}_{\mathrm{str}}$	12.7	1.3	34.5	599.2
	$\mathrm{E}\{\hat{\mathrm{V}}(\hat{\theta})\}_{\mathrm{HH}}$	21.6	1.6	43.3	849.7
PopSyn	$\mathrm{E}(\hat{\theta})$	178.7	46.6	228.1	685.0
	$\mathrm{MSE}(\hat{\theta})$	35.8	8.2	111.3	1598.9
	$\mathrm{V}(\hat{\theta})$	35.3	4.8	60.7	588.4
	$\mathrm{E}\{\hat{\mathrm{V}}(\hat{\theta})\}$	46.5	5.4	53.2	569.8

variable compared to the stratum indicator which contains the full design information used in the stratified sampling. The deviation in PopSyn can be explained by the fact that the pseudo likelihood function is an *approximation* to the true population likelihood. Third, SampleSyn2 consistently underestimates the variances for all variables, while the variance estimates for PopSyn are close to their expected values.

The substantial underestimation of the true variance for SampleSyn2 can be explained if we note that the synthesis model only partially accounts for the sampling design. Including the measure of size variable s_i in the synthesis model helps correct the bias in the point estimates, since the sampling design becomes less informative once we condition on s_i. However, the synthesis model does not exploit the efficiency gains from stratified sampling, i.e., the synthesis models still assume that the original sample was generated using simple random sampling. Hence, the uncertainty introduced at the synthesis stage is larger than if the sampling design were properly reflected. As a consequence, the variance estimate, which assumes full efficiency both at the sampling as well as at the synthesis stage, underestimates the true variance of the point estimates with the synthetic data. To overcome this problem, separate synthesis models would have to be run in each stratum, which typically is prohibitive due to small sample sizes within strata. Reiter et al. (2006) studied a similar problem in the context of multiple imputation inference under complex sampling designs. They suggested two alternatives to ensure approximately valid inferences: including stratum indicators in the imputation model or using hierarchical models with strata defining the second level of the model. Both options are difficult to implement for the Gaussian mixture synthesis model, which assumes that all variables are continuous.

Table 4 presents the results for the linear regression model. In this table, $\theta = \beta_j$ for $j = 1, 2, 3, 4$ are the regression coefficients for the simulated population and $\hat{\theta}$ denotes their estimates for the three synthesis strategies. Because the analysis model is mis-specified (the true data generating process is built on a normal mixture model, not a multivariate normal distribution or a linear model), we do not report results regarding the estimated variances from the synthetic data, since the variance estimates for the regression coefficients would be biased even for the original input data $\boldsymbol{Y}_n$. However, we still think that this type of evaluation is important, since it mimics a scenario typically encountered in practice. The analysis model will rarely match the data generating process exactly, but the results from the synthetic data should nevertheless be close to the results from the original data.

The results in Table 4 can be summarized as follows. Entirely disregarding the design information in SampleSyn1 results in biased estimates for all parameters. For β_1, the estimate even has a wrong sign. PopSyn results in smallest mean squared errors for three of the four parameters, while the measure for β_2 is smallest for SampleSyn2. However, the differences between the two methods are generally small.

Table 4: (Weighted) least square estimates of $\mathrm{E}(\log Y_1) = \beta_0 + \beta_1 \log Y_2 + \beta_2 \log Y_3 + \beta_3 \log Y_4$ for the different synthesis approaches (unit: 10^{-2} for MSE).

		β_0	β_1	β_2	β_3
Population	θ	0.91	-0.08	0.19	0.69
SampleSyn1	$\mathrm{E}(\hat{\theta})$	0.80	0.03	0.22	0.60
	$\mathrm{MSE}(\hat{\theta})$	1.29	1.12	0.15	0.89
SampleSyn2	$\mathrm{E}(\hat{\theta})$	0.88	-0.06	0.19	0.68
	$\mathrm{MSE}(\hat{\theta})$	0.32	0.24	0.03	0.24
PopSyn	$\mathrm{E}(\hat{\theta})$	0.90	-0.11	0.22	0.69
	$\mathrm{MSE}(\hat{\theta})$	0.28	0.14	0.10	0.04

Table 5: Quartiles of absolute relative differences, ARD= $|\hat{L} - L|/L$, across the 400 repeated simulations assuming that an attacker has access to the synthetic data only. Larger values of the absolute relative difference (ARD) measure indicate higher levels of protection.

		Y_1	Y_2	Y_3	Y_4
SampleSyn1	Q1	1.77	0.40	1.72	0.81
	Median	2.13	0.54	1.96	1.01
	Q3	2.70	0.74	2.35	1.34
SampleSyn2	Q1	0.30	0.04	1.03	0.11
	Median	0.41	0.08	1.16	0.21
	Q3	0.62	0.13	1.33	0.31
PopSyn	Q1	0.69	1.73	8.87	0.18
	Median	2.21	5.59	17.77	0.60
	Q3	4.16	12.43	28.35	2.30
Simulation Truth (unit: 1000)	L	621.8	91.7	152.7	2318.2

4.5.2 Disclosure risks

Table 5 compares the disclosure risks of the synthesis approaches under Scenario 1 (described in Section 3.2) assuming that an attacker has access to the synthetic data only. Note that larger values of the absolute relative difference (ARD) measure indicate higher levels of protection. The table shows that the ARD values are always lowest for the SampleSyn2 approach, while the PopSyn approach typically results in the largest values; only the first quartiles of Y_1 and Y_4 and the median of Y_4 are largest for SampleSyn1.

We also compute ARDs under Scenario 2 assuming that the attacker has access to the synthetic data and the second largest value of the studied item. This scenario results in extremely high values of ARD for all methods and variables. For example, the minimum value of ARD for Y_1 is greater than 250, i.e., the difference between the largest value and its estimate is 250 times bigger than the largest value. With our simulated data, which does not contain any extremely large establishments accounting for more than, say 50%, of the population total, this attacker strategy is not threatening. However, we acknowledge that risks from this approach could be higher in practice, if the attacker can stratify the data by other known attributes such as industry.

Combining the analytical validity results and the risk measures in a risk-utility framework, we note that although SampleSyn1 outperforms SampleSyn2 in terms of data protection, the price for this extra protection seems not justified due to the unacceptably low levels of analytical validity (as shown in Table 3). On the other hand, PopSyn performs at least as well as and often better than SampleSyn2 from a utility perspective while offering substantially improved data protection, at least based on the ARD measure. Therefore, we can conclude that PopSyn should be preferred over the other two methods, especially if we keep in mind that it will also simplify the analysis task since the sampling design information no longer needs to be taken into account.

5 Empirical Study

5.1 Background

In this section, we apply the SampleSyn1, SampleSyn2, and PopSyn generators to empirical data from the 2012 Economic Census. The synthetic data are generated from the Economic Census *final* data – edited and imputed in the production stage – from three different industries, each representing a different economic sector and highlighting different features of the census universe.

The Economic Census is the U.S. Government's official five-year measure of American business and the economy. The Economic Census is not a complete census of establishments. For the 2012 collection, the majority of the sectors employed stratified systematic samples: all multi-unit enterprises and large single-unit establishments were included in the sample with certainty in strata defined by 6- or 8-digit NAICS industry and state depending on the industry; and the remaining single-unit establishments were sampled, under the restriction that survey weights never exceed 20. The manufacturing and mining sectors employed cut-off samples for the 2012 Economic Census, the construction sector employed a stratified PPS sample (Pareto sampling), and the wholesale trade sector included all establishments with certainty (census). The Economic Census collects a core set of *general statistics* items from each establishment. Examples include total receipts or shipments ("Sales"), annual payroll ("AnnPay"), and the number of employees in the first quarter ("Emp1Q"). With the exception of the construction sector, all trade areas construct a complete universe of general statistics values by using administrative data in place of respondent data for unsampled units. In addition, the Economic Census collects industry-specific information such as the revenue obtained from product sales. The sector and industry specific items are estimated from the sampled units. Table 6 presents key features of our input data. Establishments in the studied transportation industry (488330) tend to be homogeneous in terms of the frame measure of size, regardless of geographic location. In contrast, the establishments in the studied insurance industry (524210) and the services industry (544110) vary in size and hence have the wider range of weights. The services industry establishments are either taxable or tax-exempt, with operating expenses ("OperExp") collected from the tax-exempt establishments. The editing and imputation parameters for the transportation and insurance industries used in the production stage are developed at the national industry level. In the services industry, the taxable and tax-exempt establishments are further split into two separate domains, resulting in 4 *imputation cells* total for the editing and imputation. In this particular service industry, there was insufficient data in one imputation cell for modeling, and these data were dropped from the modeling and analysis.

Table 6: Input data in the empirical study. The study variables are Annual Payroll, 1st Quarter Payroll, 1st Quarter Employment, Sales/Receipts, and Operating Expenses. [†]Editing/imputation performed separately for taxable and tax-exempt establishments (two categories per establishment type). [‡]Operating expenses are collected from tax-exempt only establishments (one imputation cell).

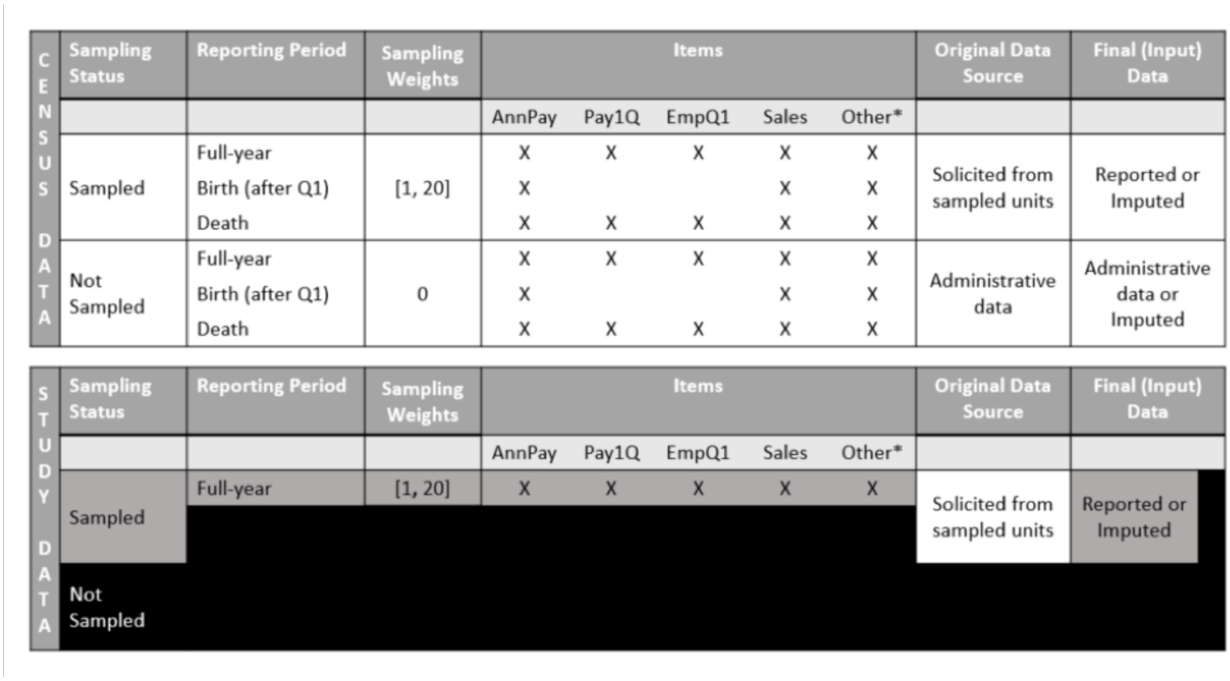
Industry (sector label)	488330	524210	544110
Sector	Transportation	Insurance	Services
Description	Navigational services to shipping	Insurance agencies and brokers	Offices of lawyers
Total sampled units	545	29,682	37,670
Proportion of certainty units	0.64	0.51	0.51
Range of survey weights	1.00 – 1.50	1.00 – 10.00	1.00 – 20.00
Number of imputation cells	1	1	3[†]
Study variables			
Annual Payroll	x	x	x
1st Quarter Payroll	x	x	x
1st Quarter Employment	x	x	x
Sales/Receipts	x	x	x
Operating Expenses	–	–	x[‡]

Unfortunately, the original frame values of sampling measure of size were not available in our historic files. Instead, we created an artificial size variable for the SampleSyn2 applications as the geometric mean of annual payroll and sales. This is a crude approximation of the actual frame measure of size and is loosely patterned after methods used by annual business surveys at the Census Bureau, which create a hybrid measure of size for each unit on the frame.

To create input data for the data synthesis of this application study, we constructed probability samples from the complete collection by dropping all unsampled units. See Figure 2 for the depiction of the input data creation process. The upper table represents the (complete) census data where all units from the frame are included. There is no solicitation of units that were not selected in the sampling process (denoted “Not Sampled”); the four core variables are imputed with administrative data. Other variables (denoted as Other*) may be solicited from sampled units or will be imputed using some form of industry regression model for unsampled units. The lower table depicts the input data used to generate the synthetic data for this application study and comprises only sampled units that were in business for the calendar year (2012). The Economic Census final data have undergone all production edits and contain imputed values. A few units in each industry were dropped (less than 1% per industry) if the final data did not satisfy all applicable ratio edits. This could occur in selected cases when an analyst verified that the reported data were correct and bypassed the edits. The operating expenses variable is included in the synthetic data generators for the services industry, as it is included in an explicit ratio edit and therefore provides important bounding for other associated variables. However, the results for operating expenses are omitted in the subsequent sections, as they do not exhibit different patterns from other variables in the industry.

In the analyses below, we treat the Horvitz-Thompson estimates from the input data as the *truth.* These estimates are not equivalent to their corresponding published totals. Besides dropping the unsampled units, we restricted the input data to *full-year reporters*, establishments that were in business for the entire calendar year. We use the production ratio edits employed by the 2012 Economic Census in our synthesis models, which are designed for the full-year reporters. We included an additional set of imputation-cell level range edits on annual payroll, 1st quarter employment, and sales, obtaining the upper limits from the maximum value of each item in the input data, multiplied by a small growth factor. The range edits, not used in production, help prevent the generators from creating impossibly large values, e.g., a synthetic unit-level value of sales larger than the gross domestic product.

This empirical analysis is not necessarily a good representation of Economic Census processing. However, it is an excellent representation of conditions for synthesizing economic data from a sample survey. By modeling synthetic data from the final data used for tabulation, we hope to produce datasets that will have similar properties in both model-based and design-based inferences as the original data without the associated disclosure risks. For many areas of research and analysis, this will be sufficient if the synthetic distributions are not overly distorted. For others, the synthetic data will be sufficient for testing programs, and researchers can request special access to the program microdata or a validation server if it exists.



CENSUS DATA	Sampling Status	Reporting Period	Sampling Weights	Items					Original Data Source	Final (Input) Data
				AnnPay	Pay1Q	EmpQ1	Sales	Other*		
	Sampled	Full-year	[1, 20]	X	X	X	X	X	Solicited from sampled units	Reported or Imputed
		Birth (after Q1)		X			X	X		
		Death		X	X	X	X	X		
	Not Sampled	Full-year	0	X	X	X	X	X	Administrative data	Administrative data or Imputed
		Birth (after Q1)		X			X	X		
		Death		X	X	X	X	X		

STUDY DATA	Sampling Status	Reporting Period	Sampling Weights	Items					Original Data Source	Final (Input) Data
				AnnPay	Pay1Q	EmpQ1	Sales	Other*		
	Sampled	Full-year	[1, 20]	X	X	X	X	X	Solicited from sampled units	Reported or Imputed
	Not Sampled									

Figure 2: Depiction of the process creating input data used for synthetic data generation in the empirical study. The upper table represents the (complete) census data in the studied industries. The lower table represents the final input data for this empirical study. AnnPay denotes annual payroll; Pay1Q, 1st quarter payroll; Emp1Q, 1st quarter employment; Sales, sales/receipts; and Other*, other variables.

As done at the U.S. Census Bureau, we assume that *all* establishment values in the datasets used for tabulations are protected, regardless of whether they are the originally reported values, administrative data substitutions, or imputed values. The protected values can be inconsistent with the supplied edits, assuming that an analyst has validated them.

5.2 Results

Within each industry, we conducted repeated experiments for each synthetic data method (SampleSyn1, SampleSyn2, and PopSyn), with a different random seed for each MCMC chain. In each trial, each synthetic method creates m=10 synthetic datasets by drawing every 200th iteration of a MCMC chain after 2000 iterations of a burn-in period. We decided to repeat the entire synthesis process ten times, to rule out the possibility that some of the findings might be the result of exceptionally *lucky* or *unlucky* draws from the synthesis model.

5.2.1 Analytical validity

In Table 7, we present summary statistics on the ratios of estimated synthetic data totals to the corresponding *true* values (θ) from the input data. For each synthetic method, we obtained the estimated total ($\hat{\theta}_r$) from each independent MCMC run ($r = 1, 2, \dots, 10$) by combining $m = 10$ synthetic datasets. The table reports the average of the ratios across the ten runs. The numbers in parentheses below the averages are the minimum and maximum values across the ten runs. Regardless of industry or item, there is very little variability in the estimates between the ten trials of the synthetic data generator, with the possible exception of the transportation industry.

We see similar patterns in Table 7 as in the simulation study. Despite imposing limits on unit-level item values via range edits, the estimated totals with SampleSyn1 are overestimated, with the degree of overestimation increasing as the survey weights become more variable, i.e., the highest in the service industry. In contrast, the estimated totals with SampleSyn2 and PopSyn approximate their

Table 7: Averages of the ratios of estimated synthetic data totals to corresponding true values across the ten trials. A ratio close to one indicates high utility of the synthetic dataset. The numbers in the parentheses are the minimum and maximum values of the ratios across 10 independent trials.

Industry	Method	AnnPay	Pay1Q	Emp1Q	Sales
Transportation	SampleSyn1	1.24 (1.16,1.36)	1.16 (1.08,1.27)	1.26 (1.17,1.34)	1.35 (1.26,1.43)
	SampleSyn2	0.98 (0.96,1.00)	0.93 (0.91,0.96)	1.05 (1.02,1.08)	1.03 (0.99,1.06)
	PopSyn	1.00 (0.96,1.05)	0.95 (0.91,1.00)	1.04 (1.00,1.07)	1.04 (0.99,1.10)
Insurance	SampleSyn1	2.54 (2.49,2.60)	2.51 (2.46,2.58)	1.95 (1.93,1.97)	2.46 (2.43,2.50)
	SampleSyn2	1.02 (1.02,1.03)	1.01 (1.00,1.02)	0.97 (0.96,1.00)	0.98 (0.96,0.98)
	PopSyn	1.04 (1.02,1.05)	1.03 (1.00,1.06)	1.00 (0.99,1.01)	1.05 (1.02,1.06)
Services	SampleSyn1	3.25 (3.19,3.44)	3.17 (3.10,3.40)	2.72 (2.67,2.83)	3.06 (2.98,3.32)
	SampleSyn2	1.04 (1.02,1.04)	1.02 (1.01,1.03)	1.03 (1.02,1.03)	0.98 (0.97,0.99)
	PopSyn	0.99 (0.94,1.04)	0.99 (0.93,1.03)	1.03 (0.98,1.07)	1.01 (0.94,1.06)

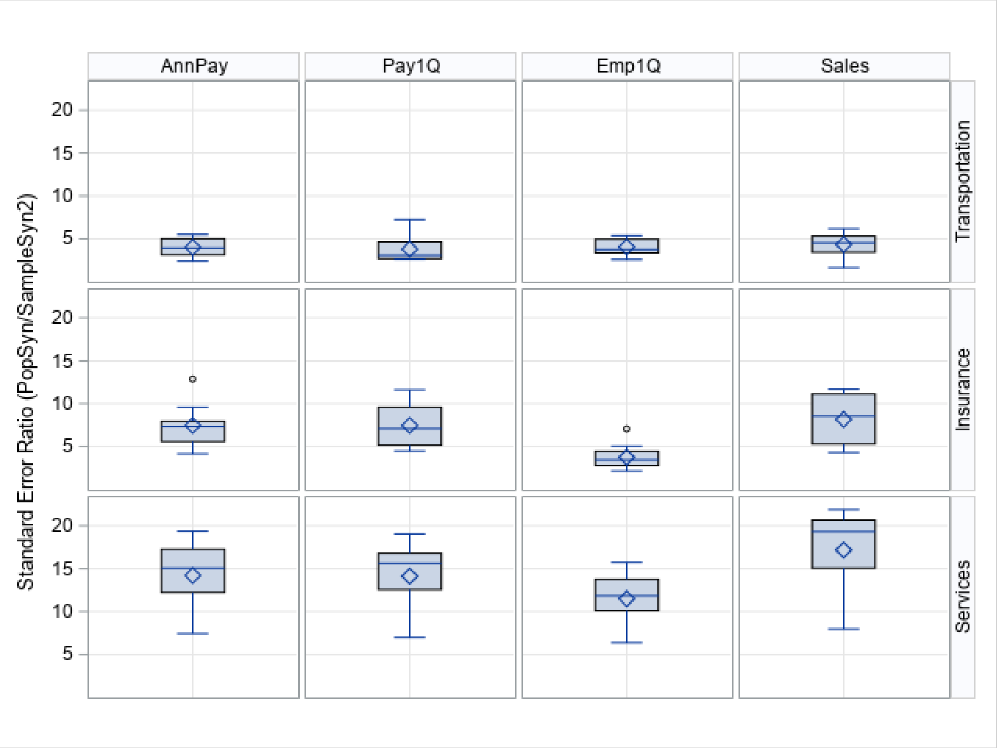


Figure 3: Ratios of the standard errors of PopSyn to those of SampleSyn2 by item within industry. Distributions are obtained from the 10 independent trials within method.

input data counterparts. Clearly, SampleSyn1 is inadequate for producing viable totals.

The simulation study in the previous section demonstrated that the true variance of the estimated totals is approximately similar for SampleSyn2 and PopSyn, but SampleSyn2 underestimates these variances, while the variance estimates of PopSyn are close to the empirical variances. If similar patterns hold in the empirical applications, we would expect that the estimated variances for PopSyn would generally be larger than the estimated variances for SampleSyn2. Figure 3 presents box plots of the standard error ratios $(\widehat{SE}_{\text{PopSyn}}/\widehat{SE}_{\text{SampleSyn2}})$ within industry by item. We used a stratified simple random sampling variance estimate for SampleSyn2 (denoted $\hat{V}(\hat{\theta})_{\text{str}}$ in Section 4.5.1), assuming that any released synthetic data would include instructions on design-appropriate within sample variance components. Consistent with the simulation results, the variance estimates of PopSyn tend to be much larger than those of SampleSyn2. These effects are especially notable

Table 8: Regression slope estimates (industry-average ratios) by industry and method, averaged over 10 independent trials.

Industry	Method	AnnPay/Emp1Q	AnnPay/Pay1Q	Sales/AnnPay
Transportation	Truth	68.60	4.12	3.64
	SampleSyn1	67.38	4.36	3.97
	SampleSyn2	64.62	4.34	3.84
	PopSyn	66.40	4.33	3.81
Insurance	Truth	59.63	3.91	2.82
	SampleSyn1	77.61	3.96	2.74
	SampleSyn2	62.56	3.97	2.71
	PopSyn	61.65	3.97	2.85
Services	Truth	77.40	4.37	2.80
	SampleSyn1	92.65	4.48	2.63
	SampleSyn2	78.00	4.42	2.65
	PopSyn	74.56	4.40	2.84

in the Services industry, which uses different edits by imputation cell and has the largest spread in survey weights. Assuming the findings from the simulation studies hold, i.e., the true variances of the point estimates are approximately similar for both synthesis approaches and the variance estimate of PopSyn is close to the true value, this application study does provide some evidence that the precision of the total estimate from SampleSyn2 is likewise overestimated.

Next, we turn to two model-based utility metrics. First, we consider a linear regression model, $Y_i = \beta X_i + \varepsilon_i$ where $\varepsilon_i \sim (0, X_i\sigma^2)$. Under this model, the least square estimate is $\hat{\beta} = \sum_i Y_i / \sum_i X_i$, i.e., the industry-average ratio. Survey weights are included in the regression estimates for the SampleSyn2 measures. Table 8 presents the regression parameters for three ratio tests. These industry-average ratios represent useful economic measures, as described earlier in Section 3.1. The results can be summarized as follows. In the transportation industry, little sampling is performed. Hence, the sampling design is not strongly informative and the results generally are very similar with all synthetic data methods although the upward bias for the Sales to AnnPay ratio is a little larger for SampleSyn1. For the insurance and services industries, SampleSyn1 shows substantial differences to the values from the original data (Truth) for the ratio of AnnPay to Emp1Q. For the

Table 9: Regression parameters for multiple regression model, $E\{\log(\text{AnnPay})\} = \beta_0 + \beta_1 \log(\text{EmpQ1}) + \beta_2 \log(\text{Sales})$, averaged over 10 trials. Coefficients whose estimates are different from the true values by at least 5% of the true values are marked with * symbols.

Industry	Method	β_0	β_1	β_2
Transportation	Truth	1.52	0.41	0.52
	SampleSyn1	1.77*	0.48*	0.47*
	SampleSyn2	1.53	0.45*	0.50
	PopSyn	1.62*	0.44*	0.50
Insurance	Truth	0.67	0.66	0.56
	SampleSyn1	0.65	0.60*	0.59*
	SampleSyn2	0.79*	0.64	0.55
	PopSyn	0.67	0.65	0.56
Services	Truth	0.87	0.68	0.53
	SampleSyn1	1.01*	0.67	0.53
	SampleSyn2	0.82*	0.62*	0.55
	PopSyn	0.87	0.68	0.53

ratio of AnnPay to Pay1Q all synthesis methods perform very similarly, which can be explained by the strong degree of association between the two variables. Finally, for the Sales to AnnPay ratio, SampleSyn1 and SampleSyn2 show similar amounts of differences to the original values. The results for SampleSyn2 might be improved with a better measure of unit size. Overall, PopSyn shows good approximations to the industry-average ratios.

Finally, we consider the regression of log(AnnPay) on log(EmpQ1) and log(Sales), obtaining all regression parameters with OLS and WLS incorporating survey weights as appropriate. Table 9 presents the regression parameters, averaged over 10 independent trials. Note that this model is not – to our knowledge – an indicative of any meaningful economic phenomenon; instead, it is meant to illustrate the type of exploratory analysis that could be performed by a curious data user. The results in Table 9 illustrate the preservation of multivariate relationships in PopSyn synthetic data but show marginal performances of SampleSyn1 and SampleSyn2.

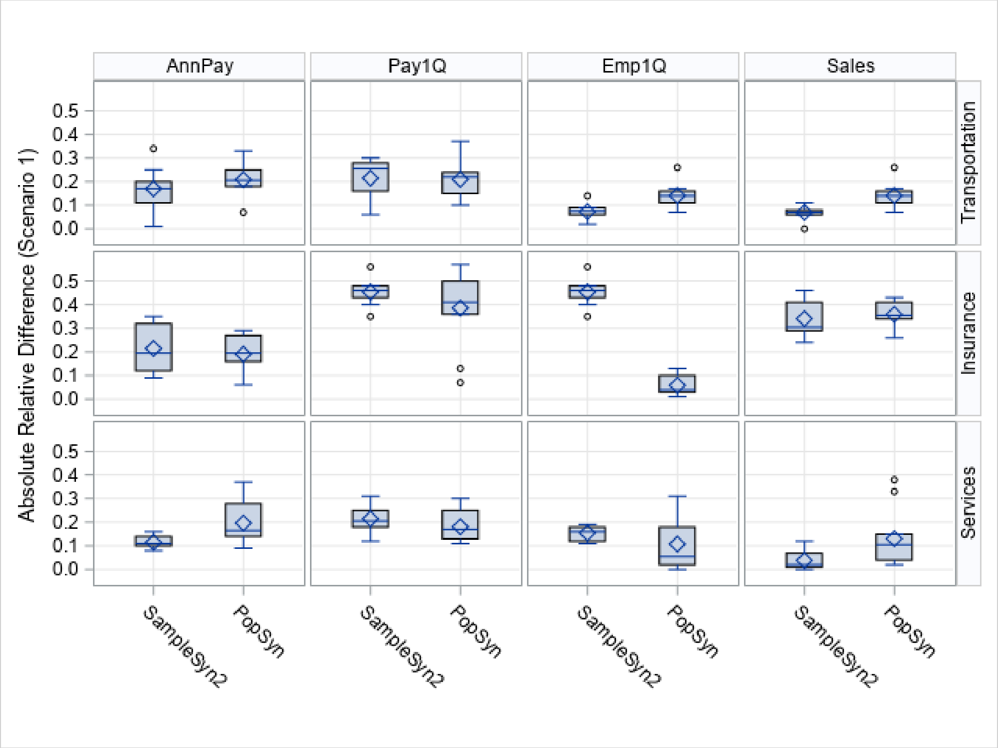


Figure 4: Box plots of the absolute relative difference of the largest estimated value, ARD= $|\hat{L}-L|/L$, between the synthetic data and the true value by industry and item. Larger values of the ARD measure indicate higher levels of protection.

5.2.2 Disclosure risk

As a whole, the SampleSyn2 and PopSyn total estimates were very close to the *true* values. This proximity could have implications on the disclosure risk (sensitivity) of the synthetic data. Figure 4 presents box plots of the absolute relative difference, ARD= $|\hat{L} - L|/L$, of the four items collected in all industries under Scenario 1 (the attacker has access to the synthetic data only). We also computed the ARD of each total under Scenario 2 (the attacker has access to the synthetic data and the value from the 2nd largest establishment plus c collaborators, where $c = 0$, 1, or 2). However, the estimates under this scenario were not sensitive, with all ARDs well above 100%, so we do not consider the scenario hereafter.

Figure 4 provides a few interesting insights. In the transportation industry, the differences of ARDs between SampleSyn2 and PopSyn are nonnegligible for all items but the first quarter payroll (Pay1Q). ARD values are above 0.05 for all four variables in all ten trials for PopSyn but only for seven out of ten trials for SampleSyn2. Hence, PopSyn offers a higher level of protection for the transportation industry. In the insurance and services industries, we expected that SampleSyn2 would result in very small ARD values for annual payroll (AnnPay), Pay1Q, and annual sales (Sales), as the constructed size variable is the geometric mean of the first two variables, and Pay1Q is strongly related to AnnPay. We do not see strong evidence of this in the insurance industry, but do see these effects in the services industry, which is the most finely stratified, has the most variable weights, and imposes the finest-level edit restrictions. Notice the extremely high disclosure risk in Sales from SampleSyn2, with ARD values below 0.05 in seven of the ten SampleSyn2 trials. Generating a fully synthetic population (PopSyn) generally reduces the sensitivity of the largest values for AnnPay, Pay1Q, and Sales over their partially synthetic sample counterparts (SampleSyn2). Again, this makes sense given that the size variable is a strong predictor of these three items, whereas the sampling weight used in PopSyn is not strongly related to any items.

Some attention should be paid to the ARD values of Emp1Q in both the insurance and services industries. As expected, the estimate of the largest unit's values of Emp1Q from SampleSyn2 is not too precise due to the weak association between the size variable included in the SampleSyn2 model and the number of employees. While this appears to be a sensible explanation for the low sensitivity of the SampleSyn2 employment estimates, it is not sufficient to explain the extremely high precision of the corresponding PopSyn estimates, especially in the insurance industry. Both insurance and service industries contain a small set of certainty units that have similar large values of employment (at least 1,000 employees more than the 99th percentile value in the industry) with relatively small ratios of AnnPay to Sales distinct from other establishments. In other words, this set of a few establishments has large workforces paid at the low end of the industry rate with average profit ratios. In PopSyn, these few large establishments form a single mixture component with a small variance $\mathbf{\Sigma}_k$, which in turn generates some records that are very similar to the original establishments contained in this mixture component. The fact that the establishment with the largest Emp1Q value happens to be among this group of homogeneous establishments results in an accurate estimate of the largest value of Emp1Q. We present a more descriptive explanation, along with a sensitivity analysis in the insurance industry, in Section 5.2.3. The effects of this set of isolated observations are less pronounced in the services industry because the synthetic data are generated separately by imputation cell, so that the *largest* employment observation in the fully assembled industry data could be obtained from one of three different imputation cell populations.

We make three general observations with respect to this phenomenon. First, such clustering is not unusual in economic data sets. For example, a small subgroup of businesses could have large expenditures on structural improvements while maintaining more typical levels of payroll, employment, and sales. Second, the effects of the isolated clusters on the fully synthetic populations should decline as more variables are included in the synthesis, as it becomes less likely that several units are very homogeneous in all the attributes while at the same time being far enough from the rest of the data so that they will always end up in the same mixture component. Lastly, these results demonstrate that one cannot offer an *ex-ante* guarantee that the data will be fully protected, even with full synthesis.

5.2.3 Sensitivity analysis in the insurance industry

In Section 5.2.2, we describe the effects of clustered values on the studied synthetic data generator. Because of the regulations, we cannot provide scatter plots of the real Economic Census data used for the synthesis. Instead, Figure 5 illustrates this phenomenon, using a fictional multivariate distribution of three variables (employment, sales, and inventories). The sales and inventories variables are collected in thousands of dollars: sales values range from 1,345 to 270,244; inventories values range from 1,223 to 273,031; and the correlation between the two items is 0.74. In contrast, employment is collected in units, and its values range from 4 to 410, with correlation coefficients of 0.57 to sales and 0.37 to inventories. In this figure, the total number of employees are plotted against the ratio of sales to inventories.

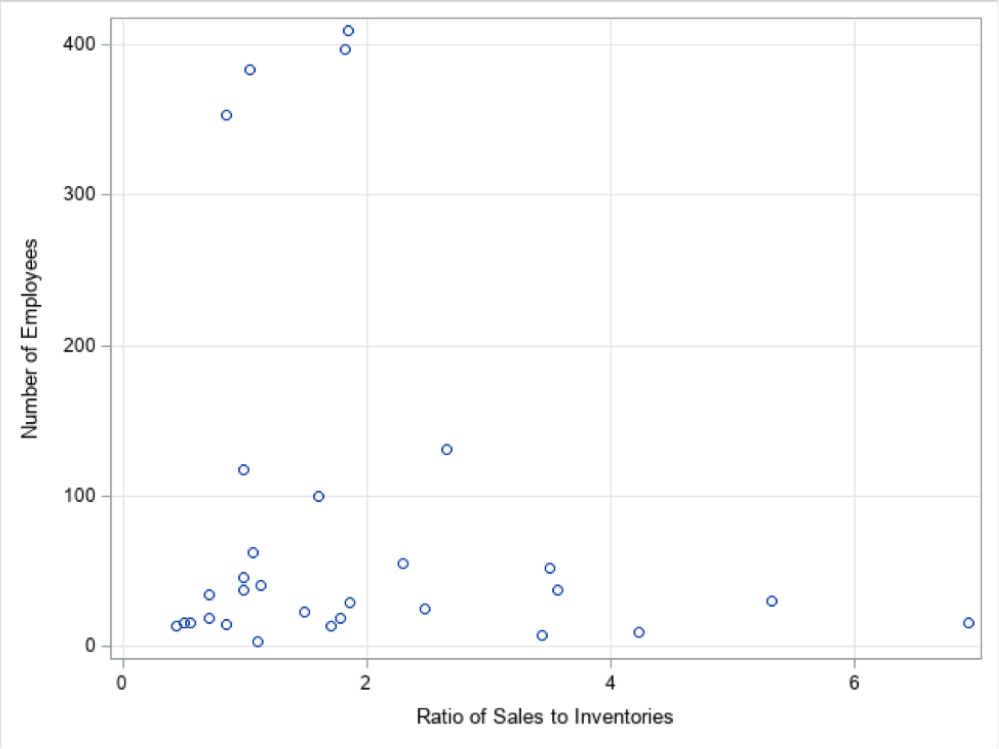


Figure 5: Fictional multivariate distribution of employment and inventories to illustrate the effects of clustered values in Section 5.2.2.

Notice the distinct set of four large establishments in the upper left hand corner, which will be distinctly represented in a mixture component with a small value of $\mathbf{\Sigma}_k$. The PopSyn synthesizer will always generate some records that are very similar to the establishments contained in this mixture component. Because the unit with the largest value of employment is contained in this group of homogeneous units, we postulate that there is a very high probability of consistently generating a synthetic unit that matches the largest unit in the original data (in terms of employment) on all items. To verify this hypothesis, we modified the four largest values of EMP1Q in the input dataset for the insurance industry, replacing each original value with their mean, i.e., the average of the four largest values, and leaving the remaining items in these four records untouched. After modifying the input data (one of four variables in four of approximately 29,000 observations), we ran 10 independent trials using the PopSyn method, with a different random seed for each MCMC chain, essentially repeating the procedure described in Section 5.1. Figure 6 presents summary results over 10 independent trials, with the upper row presenting ratios of the PopSyn synthetic totals to the true totals (utility measure) and the lower row presenting Scenario 1 privacy loss for each studied variable, where privacy loss is defined as one minus absolute relative difference.

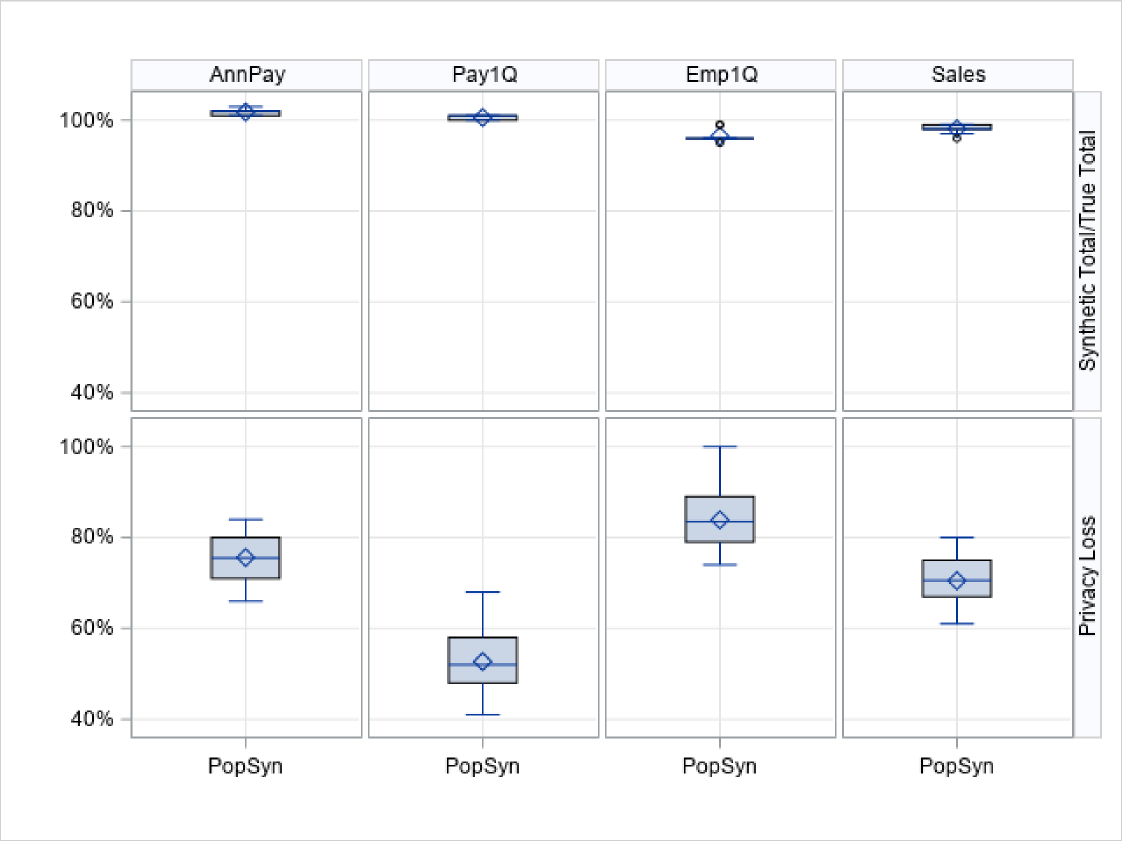


Figure 6: Utility (upper row) and privacy metrics (lower row) obtained from modified insurance data by macroaggregation.

As expected, the synthetic data totals remain very close to the original data totals, and the privacy loss for annual payroll, 1st quarter payroll, and sales are – as expected – unchanged. However, the privacy loss for 1st quarter employment is greatly improved, with no degradation in corresponding utility. Indeed, the ARD values for 1st quarter employment were greater than 0.10 in nine of the 10 independent trials. Note that we do not recommend modifying the input data as a general practice for resolving overly accurate synthetic data estimates. This investigative approach, however, did provide strong evidence in favor of our hypothesis and – in this case – yielded synthetic populations with high utility and low disclosure risk in this specific situation. Note that the differences in regression parameters for all studied models was trivial and are not reported.

6 Conclusion

In most National Statistical Institutes, access to confidential business data is strictly regulated to reduce the risk of an unauthorized disclosure and providing alternative microdata such as PUMS is generally discouraged with business data. Consequently, disseminating fully synthetic data instead of the original data is an appealing strategy to facilitate access to the data. Our paper focuses on providing useful *general-purpose* synthetic data under informative sampling designs. As we illustrate through a simulation study followed by an empirical application using selected industries from the 2012 U.S. Economic Census, synthetic data approaches previously proposed in the literature (SampleSyn1 in our study), treating the original sample as a simple random sample from the population, will introduce substantial bias in inference, as they fail to account for the informative sampling design. To overcome this problem, we propose a pseudo likelihood approach and present the necessary adjustments for the Gibbs sampler employed to generate the synthetic data. We treat the revised Gibbs sampler as a *super-population* synthesizer and draw finite synthetic populations. In our simulations and application, the synthetic population produced approximately the same benchmark totals as the corresponding original data, retained the multivariate relationships between items, and conformed to a predetermined set of edits.

The pseudo likelihood approach proposed in this paper offers three important advantages. First, it is straightforward for statistical agencies to implement, as it only requires incorporating the sampling weights into the likelihood function. A commonly suggested strategy for dealing with informative sampling designs in a strictly model-based framework is to include the entire set of design variables in the model as additional predictors (SampleSyn2 in our study). This approach assumes that all design variables are available to the data synthesizer, which is not always the case as in the empirical application presented in Section 5. Second, the approach should also be applicable in the general context of multiple imputation for nonresponse for complex survey designs. The multiple imputation approach has been criticized repeatedly for ignoring the complex sampling designs typically encountered when dealing with survey data (Fay, 1996; Kim et al., 2006, among others). We are planning to evaluate whether the pseudo likelihood approach can be a suitable way to address this criticism. Third, users do not need to consider the sampling design of the original survey data because they are provided with synthetic *populations*.

The empirical data application of this paper further reveals that even fully synthetic data are not free from disclosure risk. This is an important finding. It seems prudent to assume that risks are generally negligible for fully synthetic data, as the data cannot be linked to any original records. However, even though risks of re-identification are typically trivial for fully synthetic data, the risk of attribute disclosure may not be. Indeed, attribute disclosure is often a large concern with business populations. Although the inclusion of the largest establishments (or enterprises) in sample surveys is generally not considered as confidential, specific attributes of these units are expected to be protected in the disseminated data products. Attribute risk appears to be amplified with the release of masked or synthetic microdata, as an atypical unit in a tabulation cell could potentially provide information on many related items. The insurance industry application provided in Section 5 illustrates this phenomenon: any attacker can accurately estimate the number of employees for the largest unit in this industry from the synthetic population data generated by the pseudo likelihood model and could subsequently make reasonable inferences about the same unit's values of annual payroll, 1st quarter payroll, and total sales.

This finding highlights the common dilemma in data dissemination: there is always a trade-off between preserving the information in the data and the risk of disclosure. The flexibility of the DP Gaussian mixture model used in our considered synthetic data generators ensures that the analytical validity is high even when looking at small subsets of the particular set of data. However, any utility measure becomes a risk measure when applied at a very detailed level. This tradeoff needs to be respected in the development and release of synthetic data.

We propose a synthetic data generator that fuses design-based and model-based methods, adopting the design-based perspective of a superpopulation from which finite populations are randomly drawn while using model-based methods to generate viable posterior predictive populations. Sample design features are embedded within the generating models. Variance estimation with the multiple synthetic populations is straightforward and easy to implement in practice, allowing for valid inference in many commonly-used settings. We did identify data conditions that can lead to increased sensitivity in the synthetic data, and this sensitivity should be checked prior to synthetic data release. Developing methods for dealing with this problem is another interesting area for future research.

Certainly, the proposed synthesis approach, which accommodates many special features of business data distributions, is an improvement over more standard synthetic data generators such as synthpop (Nowok et al., 2016). Given the utility requirement that the synthetic data produce *similar* benchmark totals as the corresponding original data statistics, it might be possible to release these tabulations in place of noise-infused totals as obtained using differential privacy or other formal privacy frameworks. Of course, that is another area of future research, as is the effect of releasing synthetic microdata or macrodata on the overall privacy budget.

Acknowledgments. This research was supported by the American Statistical Association, National Science Foundation and the U.S. Census Bureau through the ASA/NSF/Census fellowship to Kim (YA132318AE0011). Any views expressed here are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied.

References

Abowd, J. M., Stinson, M., and Benedetto, G. (2006), "Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project," Technical Report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program, 2006 (Available from https://ecommons.cornell.edu/handle/1813/43929).

Berg, E., Kim, J.-K., and Skinner, C. (2016), "Imputation under informative sampling," *Journal of Survey Statistics and Methodology*, 4, 436–462.

Calvino, A. (2017), "A simple method for limiting disclosure in continuous microdata based on principal component analysis," *Journal of Official Statistics*, 33, 15–41.

Cox, L. H., Karr, A. F., and Kinney, S. K. (2011), "Risk-utility paradigms for statistical disclosure limitation: how to think, but not how to act," *International Statistical Review*, 79, 160–183.

Dalenius, T. and Hodges Jr, J. L. (1959), "Minimum variance stratification," *Journal of the American Statistical Association*, 54, 88–101.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014), "A nonparametric method to generate synthetic populations to adjust for complex sampling design features," *Survey Methodology*, 40, 29–46.

Drechsler, J. (2011), *Synthetic datasets for statistical disclosure control: theory and implementation*, vol. 201, Springer Science & Business Media.

— (2018), "Some clarifications regarding fully synthetic data," in *International Conference on Privacy in Statistical Databases*, Springer, pp. 109–121.

Drechsler, J. and Reiter, J. P. (2008), "Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data," in *International Conference on Privacy in Statistical Databases*, Springer, Berlin, Heidelberg, pp. 227–238.

Dunson, D. B. and Xing, C. (2009), "Nonparametric Bayes modeling of multivariate categorical data," *Journal of the American Statistical Association*, 104, 1042–1051.

Fay, R. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, 91, 490–498.

FCSM (2005), "Report on Statistical Disclosure Control Limitation Methodology," Federal Committee on Statistical Methodology Working Paper 22 (Available from https://nces.ed.gov/fcsm/pdf/spwp22.pdf).

Fienberg, S. E. (2010), "The relevance or irrelevance of weights for confidentiality and statistical analyses," *Journal of Privacy and Confidentiality*, 1, 183–195.

Fuller, W. A. (2009), *Sampling Statistics*, Wiley, Hoboken.

Godambe, V. and Thompson, M. E. (1986), "Parameters of superpopulation and survey population: their relationships and estimation," *International Statistical Review*, 54, 127–138.

Hawala, S. (2008), "Producing partially synthetic data to avoid disclosure," in *Proceedings of the Joint Statistical Meetings*, American Statistical Association, Alexandria, VA.

Hu, J., Reiter, J. P., and Wang, Q. (2018a), "Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data," *Bayesian Analysis*, 13, 183–200.

— (2018b), "Disclosure risk evaluation for fully synthetic categorical data," in *International Conference on Privacy in Statistical Databases*, Springer, pp. 185–199.

Kim, H. J., Reiter, J. P., and Karr, A. F. (2018), "Simultaneous edit-imputation and disclosure limitation for business establishment data," *Journal of Applied Statistics*, 45, 63–82.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014), "Multiple imputation of missing or faulty values under linear constraints," *Journal of Business & Economic Statistics*, 32, 375–386.

Kim, J., Brick, J., Fuller, W., and Kalton, G. (2006), "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68, 509–521.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011), "Towards unrestricted public use business microdata: The synthetic longitudinal business database," *International Statistical Review*, 79, 362–384.

Kish, L. and Frankel, M. R. (1974), "Inference from complex samples," *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 1–22.

Lavallée, P. and Hidiroglou, M. A. (1988), "On the stratification of skewed populations," *Survey Methodology*, 14, 33–43.

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

Meng, X.-L. and Zaslavsky, A. M. (2002), "Single observation unbiased priors," *The Annals of Statistics*, 30, 1345–1375.

Nowok, B., Raab, G., and Dibben, C. (2016), “synthpop: Bespoke Creation of Synthetic Data in R,” *Journal of Statistical Software*, 74, 1–26.

O’Malley, A. J. and Zaslavsky, A. M. (2008), “Domain-level covariance analysis for multilevel survey data with structured nonresponse,” *Journal of the American Statistical Association*, 103, 1405–1418.

Pfeffermann, D. (1993), “The role of sampling weights when modeling survey data,” *International Statistical Review*, 61, 317–337.

Raab, G. M., Nowok, B., and Dibben, C. (2017), “Practical data synthesis for large samples,” *Journal of Privacy and Confidentiality*, 7, 67–97.

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. (2003), “Inference for partially synthetic, public use microdata sets,” *Survey Methodology*, 29, 181–188.

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

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006), “The importance of modeling the sampling design in multiple imputation for missing data,” *Survey Methodology*, 32, 143–149.

Reiter, J. P., Wang, Q., and Zhang, B. (2014), “Bayesian estimation of disclosure risks for multiply imputed, synthetic data,” *Journal of Privacy and Confidentiality*, 6, 17–33.

Rubin, D. B. (1993), “Statistical disclosure limitation,” *Journal of Official Statistics*, 9, 461–468.

Smith, P. (2013), “Sampling and Estimation for Business Surveys,” in *Designing and Conducting*

Business Surveys, eds. Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D. K., John Wiley & Sons, Ltd, chap. 5, pp. 165–218.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A. (2018), "General and specific utility measures for synthetic data," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 663–688.

Sugden, R. and Smith, T. (1984), "Ignorable and informative designs in survey sampling inference," *Biometrika*, 71, 495–506.

Thompson, K. J. and Oliver, B. E. (2012), "Response rates in business surveys: Going beyond the usual performance measure," *Journal of Official Statistics*, 28, 221–237.

U.S. Census Bureau (2006), "Current Population Survey Design and Methodology," Technical Paper 66 (Available from https://www.census.gov/prod/2006pubs/tp-66.pdf).

Valliant, R. (1987), "Generalized variance functions in stratified two-stage sampling," *Journal of the American Statistical Association*, 82, 499–508.

Willenborg, L. and De Waal, T. (1996), *Statistical disclosure control in practice*, New York: Springer-Verlag.