Distribution-Preserving Statistical Disclosure Limitation

Woodcock, Simon and Benedetto, Gary

September 2006

LONGITUDINAL EMPLOYER - HOUSEHOLD DYNAMICS

TECHNICAL PAPER NO. TP-2006-04

Distribution-Preserving Statistical Disclosure Limitation

Date : September 2006
Prepared by : Simon D. Woodcock and Gary Benedetto
Contact : U.S. Census Bureau, LEHD Program
FB 2138-3
4700 Silver Hill Rd.
Suitland, MD 20233 USA

This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant 5 R01 AG018854-02, and the Alfred P. Sloan Foundation. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau supports external researchers use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, Program Manager, U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. http://lehd.dsd.census.gov.

Distribution-Preserving Statistical Disclosure Limitation[1]

Simon D. Woodcock[2]
Simon Fraser University
simon_woodcock@sfu.ca

Gary Benedetto
US Census Bureau
University of Maryland
gary.linus.benedetto@census.gov

September 2006

[1] This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. This document is released to inform interested parties of research and to encourage discussion. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grants SES-9978093 and SES-0427889 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01 AG018854, and the Alfred P. Sloan Foundation. The views expressed on statistical, methodological, or technical issues are those of the authors and not necessarily those of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau supports external researchers' use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, Program Manager, U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. (Jeremy.S.Wu@census.gov http://lehd.dsd.census.gov).

[2] Correspondence to: Simon Woodcock, Department of Economics, Simon Fraser University, 8888 University Dr., Burnaby, BC V5A 1S6, Canada. We thank John Abowd, Sam Hawala, Jerry Reiter, Stephen Fienberg, Pascal Lavergne, Krishna Pendakur, members of the LEHD Program staff, and participants at the 2005 Meeting of the Canadian Econometric Study Group for helpful comments and suggestions. We also thank Bryan Richetti for providing SAS code used in the re-identification analysis. We gratefully acknowledge the financial support of National Science Foundation grants SES-0427889 and SES-0339191 awarded to Cornell University, and the Simon Fraser University President's Research Grant.

Abstract

One approach to limiting disclosure risk in public-use microdata is to release multiply-imputed, partially synthetic data sets. These are data on actual respondents, but with confidential data replaced by multiply-imputed synthetic values. A mis-specified imputation model can invalidate inferences because the distribution of synthetic data is completely determined by the model used to generate them. We present two practical methods of generating synthetic values when the imputer has only limited information about the true data generating process. One is applicable when the true likelihood is known up to a monotone transformation. The second requires only limited knowledge of the true likelihood, but nevertheless preserves the conditional distribution of the confidential data, up to sampling error, on arbitrary subdomains. Our method maximizes data utility and minimizes incremental disclosure risk up to posterior uncertainty in the imputation model and sampling error in the estimated transformation. We validate the approach with a simulation and application to a large linked employer-employee database.

Keywords: statistical disclosure limitation, confidentiality, privacy, multiple imputation, partially synthetic data

1 Introduction

Statistical agencies face two competing objectives when preparing data for public release. On the one hand, they endeavor to provide their users with high quality data. On the other hand, they must maintain the privacy of respondents. The trade-off between these objectives is very real because protecting privacy usually entails information loss (Duncan et al., 2001). Unless care is taken, measures to protect privacy can invalidate statistical inferences.

We present a practical method for protecting privacy in statistical databases that permits valid inferences about the population of interest. Our approach draws upon the established literature for multiple-imputation of missing data, and builds on recent research that applies multiple-imputation to the problem of statistical disclosure limitation. Our approach is to replace confidential data with synthetic values sampled from the posterior predictive distribution of an imputation model. This substantially limits the risk of identity disclosure. At the same time, it admits valid inferences using standard statistical methods and software. Furthermore, and this is the primary contribution of this paper, our method does not require complete knowledge of the joint distribution of the data, but nevertheless preserves the conditional distribution of the confidential data, up to sampling error, on arbitrary subdomains.

Traditional approaches to disclosure limitation include suppressing confidential data, aggregation, topcoding, adding noise, and swapping values between records (see e.g., Willenborg and de Waal (1996) or the appendix to Abowd and Woodcock (2001) for surveys). All of these have the potential to distort the joint distribution of the data, and may therefore invalidate inference. At best, valid inferences can be obtained using specialized software and methods, and/or when users are provided with detailed information about the methods used to limit disclosure risk. In practice, however, such detailed information cannot be released without compromising privacy.

An alternative that permits valid statistical inferences using standard software and methods is to release multiple synthetic data sets with the same joint distribution as the confidential database. Rubin (1993) suggests generating synthetic data through multiple imputation;[1] Fienberg (1994) suggests generating synthetic data by bootstrap methods.[2] Under either approach, the released data pose little or no disclosure risk because they are completely synthetic, i.e., contain no actual data on actual respondents. However, this approach requires knowledge, or a good estimate, of the joint distribution of the data. This is imprac-

[1] This proposal is developed more fully in Raghunathan et al. (2003). Reiter (2002) provides a simulation study, Reiter (2005c) discusses inference, and Reiter (2005b) provides an application.

[2] Fienberg et al. (1998) apply this method to categorical data; Fienberg and Makov (1998) use related concepts to develop a measure of disclosure risk

tical in many instances. A tractable alternative is to release data on actual respondents, but replace confidential data with synthetic values sampled from an estimate of the joint distribution of the confidential data conditional on disclosable data. Such data, which have become known as partially synthetic data, are the focus of this paper.

Kennickell (1997) pioneered the use of multiply-imputed, partially synthetic data in the Survey of Consumer Finances. Since that early work, several approaches have been suggested to generate the synthetic values. Abowd and Woodcock (2001) present a computationally tractable approximation to the joint distribution of the confidential data given disclosable data based on a sequence of regression models. They use this approximation to multiply-impute confidential values in linked employer-employee data. Little and Liu (2003) develop a parametric method, called SMIKe, to selectively multiply-impute discrete "key" variables that pose high disclosure risk. Reiter (2005d) presents a nonparametric method to multiply-impute synthetic values using classification and regression trees (CART).

Each of these approaches makes an important contribution, but all have limitations. SMIKe is only applicable to categorical key variables. CART, though data-driven and requiring little modeling input from the imputer, presents a sufficient computational burden to preclude applications involving many variables. And though Abowd and Woodcock (2001) demonstrate that regression-based methods perform well in practice, the regression models are subject to mis-specification when the true data generating process is unknown. This is the case considered here.

We present two methods to multiply-impute confidential data when the true likelihood is unknown. Both are predicated on the assumption that the data provider prefers to use simple, or otherwise convenient, imputation models to generate the synthetic values (e.g., regression models). We believe this assumption reflects reality at many statistical agencies. Our approach, therefore, is to apply a simple transformation to the confidential data that maps between their distribution and a distribution compatible with the imputation model, and apply an inverse transformation to the synthetic values. As we demonstrate through simulation and a large-scale application, this approach preserves important statistical properties of the confidential data, including higher moments, with low disclosure risk. Furthermore, it is easily applied in practical situations involving many variables and observations.

Our first method applies in the simplest possible case: when the likelihood is known up to a monotone transformation. In this case, generating synthetic values subject to a transformation (either known or estimated) is logically equivalent to direct synthesis, up to any uncertainty in an estimated transformation. This result is elementary and serves primarily to motivate our second method, which is more generally applicable. Here we apply a density-based transformation to the variable under imputation on an arbitrary collection of subdomains. The synthesis is performed using a convenient model on the transformed data, and then the synthetic values are returned to their natural scale via an inverse transformation. This preserves the distribution of the confidential data, up to uncertainty in the estimated transformation, on those subdomains. The density-based transformation is similar in spirit to the nonlinear data-fitting methods of Lin and Vonesh (1989) and Nusser et al. (1996), and the copula-based additive noise perturbation of Sarathy et al. (2002).

The remainder of the paper is organized as follows. To fix ideas, we introduce key concepts and a novel measure of data utility and disclosure risk for multiply-imputed, partially synthetic data in Section 2. Section 3 develops the transformation-based synthesis methods. Section 4 presents a simulation study, and in Section 5 we apply our method to a large longitudinal database on employers and employees. Section 6 concludes.

2 Background and Concepts

2.1 The Data Provider's Problem

Suppose the data provider has a database that consists of confidential microdata $\mathbf{Y}$ and disclosable microdata $\mathbf{X}$. Both $\mathbf{X}$ and $\mathbf{Y}$ may contain discrete and continuous elements. Let $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$ represent the database in question, and $F(\mathbf{D})$ its probability distribution.

The data provider wishes to release public microdata $\tilde{\mathbf{D}}$. The provider's competing objectives are to maximize *data utility* and minimize *disclosure risk*. Unfortunately, there is no universally agreed upon definition of data utility or disclosure risk. We follow Muralidhar and Sarathy (2003), and define data utility as the extent to which the released data $\tilde{\mathbf{D}}$ share the statistical properties of the confidential data $\mathbf{D}$. By this definition, data utility is maximized when $F(\tilde{\mathbf{D}}) = F(\mathbf{D})$. In this case, any statistical analysis performed on the released data gives exactly the same results as would have been obtained on the underlying confidential data. This definition is consistent with usual practice for assessing data utility, which is to compare the extent to which the released data and the confidential data yield similar inferences about quantities of substantive interest, e.g., moments (typically the first two), regression coefficients, and the like.

Following Muralidhar and Sarathy (2003) again, we define disclosure risk as the ability of a malicious data user (i.e., an intruder or snooper) to infer the value of a confidential datum. This includes both identity disclosure (i.e., inferring the identity of a respondent, when this is confidential), and attribute disclosure (i.e., inferring the value of a confidential variable). Muralidhar and Sarathy (2003), Duncan and Lambert (1986), and others argue that the relevant measure of disclosure risk is the incremental risk arising from data release. There is incremental disclosure risk if the data release provides information about the distribution of confidential microdata that cannot be inferred from the disclosable data alone. Hence incremental disclosure risk is minimized when $F(\mathbf{Y}|\tilde{\mathbf{D}}) = F(\mathbf{Y}|\mathbf{X})$. In practice, it can be difficult to determine whether this equality holds, however, and there remain few practical alternatives to measure disclosure risk.

Elliot (2001), Domingo-Ferrer and Torra (2003), Winkler (2004), Reiter (2005a), and others argue in favor of assessing disclosure risk through simulations that mimic the behavior of a malicious data user that seeks to compromise confidentiality. Such simulations, usually called re-identification experiments, use sophisticated record-linkage techniques to match records in the public microdata to a secondary data source – often the confidential data themselves. If a record in the public microdata is successfully matched to the same respondent's record in a secondary data source containing unique identifiers such as names or SSNs, the respondent is deemed "re-identified." This is usually considered an identity disclosure. Thus a useful measure of disclosure risk is given by the re-identification rate, i.e., the proportion of records in the public data that are re-identified via simulation.

2.2 Multiply-Imputed, Partially Synthetic Data

Partially synthetic data replaces confidential values $\mathbf{Y}$ with synthetic values $\tilde{\mathbf{Y}}.$ A partially synthetic data release is $\tilde{\mathbf{D}} = (\mathbf{X}, \tilde{\mathbf{Y}})$. Data utility is maximized when $F(\mathbf{X}, \tilde{\mathbf{Y}}) = F(\mathbf{X}, \mathbf{Y})$. There is no incremental disclosure risk when $F(\mathbf{Y}|\mathbf{X}, \tilde{\mathbf{Y}}) = F(\mathbf{Y}|\mathbf{X})$. Muralidhar and Sarathy (2003) show that incremental disclosure risk is minimized when $\tilde{\mathbf{Y}}$ is generated by sampling from $F(\mathbf{Y}|\mathbf{X})$. The synthetic values can be generated by various methods, including sampling from a smoothed estimate of $F(\mathbf{Y}|\mathbf{X})$, as proposed by Fienberg (1994); or multiple-imputation, as proposed by Rubin (1993). We adopt the latter approach.

Multiply-imputed, partially synthetic (MIPS) data are based on a parametric imputation model for the confidential data conditional on disclosable data. This is defined by a likelihood $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are unknown parameters. Synthetic values are sampled from the posterior predictive distribution of the imputation model:

$$p\left(\tilde{\mathbf{Y}}|\mathbf{X}, \mathbf{Y}\right) = \int p\left(\tilde{\mathbf{Y}}|\mathbf{X}, \boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})\, d\boldsymbol{\theta}. \tag{1}$$

Relating MIPS data to the Muralidhar and Sarathy (2003) definitions of data utility and disclosure risk is somewhat awkward because their definitions do not acknowledge uncertainty about the joint distribution of the data.[3] However, it is possible to make some progress. We

[3] They implicitly assume that the likelihood is known and there is no posterior uncertainty.

see from equation (1) that the distribution of synthetic values $p(\tilde{\mathbf{Y}}|\mathbf{Y}, \mathbf{X})$ depends on $\mathbf{Y}$ only via the posterior distribution of $\boldsymbol{\theta}$. It follows that when there is no posterior uncertainty, i.e., when the distribution of $\boldsymbol{\theta}$ is known, MIPS data maximize data utility and minimize disclosure risk. That is, if there was no posterior uncertainty we could sample from the predictive distribution[4]

$$p\left(\tilde{\mathbf{Y}}|\mathbf{X}\right) = \int p\left(\tilde{\mathbf{Y}}|\mathbf{X}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} = p\left(\mathbf{Y}|\mathbf{X}\right). \tag{2}$$

This equality implies that when there is no posterior uncertainty, MIPS data maximize data utility and minimize disclosure risk. This motivates a new notion of data utility and disclosure risk applicable to multiply-imputed, partially synthetic data. We say that MIPS data maximize data utility and minimize disclosure risk *up to posterior uncertainty.*

2.3 Inference

The main virtue of multiple-imputation is that it yields valid statistical inferences. It is well known that this requires multiple draws from the posterior predictive distribution. We refer to a particular draw from the posterior predictive distribution as a partially synthetic data implicate, $\tilde{\mathbf{Y}}^m$. Valid statistical inferences require that the data provider release multiple implicates: $\tilde{\mathbf{D}}^m = \left(\mathbf{X}, \tilde{\mathbf{Y}}^m\right)$ for $m = 1, 2, ..., M$.

Suppose that with access to the confidential data $\mathbf{D}$, users would base inference about a scalar population quantity Q on a sample statistic q with asymptotic distribution $(Q - q) \overset{a}{\sim} N\left(0, V\right)$. Obtaining valid inferences from MIPS data is straightforward. The user computes the sample statistic q^m on each partially synthetic data implicate. Let v^m denote the sampling variance of q^m. Estimates from the M implicates are combined using the statistics:

$$\bar{q} = \frac{1}{M}\sum_{m=1}^{M} q^m, \quad b = \frac{1}{M-1}\sum_{m=1}^{M}\left(q^m - \bar{q}\right)^2, \quad \bar{v} = \frac{1}{M}\sum_{m=1}^{M} v^m. \tag{3}$$

Reiter (2003) shows that inferences based on $\bar{q}$ are valid for Q, and that an unbiased estimator of the variance of $\bar{q}$ is $T = M^{-1}b + \bar{v}$. These combining rules differ slightly from those for multiply-imputed missing data (e.g., Rubin, 1987) because in MIPS data the "non-response" mechanism (i.e., the decision to impute a confidential value) is non-stochastic.[5]

[4] If the distribution of $\boldsymbol{\theta}$ were degenerate and its value known, we could sample synthetic values from $p\left(\tilde{\mathbf{Y}}|\mathbf{X}, \boldsymbol{\theta}\right)$, which achieves the same result.

[5] Reiter (2004) presents combining rules for the case where multiple imputation is used both for missing data imputation and disclosure limitation.

2.4 Specifying the Likelihood

In practice, the most challenging aspect of generating MIPS data is specifying the joint likelihood $p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}\right)$. This is particularly true when there are many confidential variables, when some are continuous and others are discrete, and when relationships among variables are complex. This is the usual situation in practical applications. It is therefore advantageous to specify the joint likelihood as a sequence of univariate conditional likelihoods.

If we write $\mathbf{Y} = [\mathbf{y}_1 \;\; \mathbf{y}_2 \;\; \cdots \;\; \mathbf{y}_K]$ and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \;\; \boldsymbol{\theta}_2 \;\; \cdots \;\; \boldsymbol{\theta}_K]$, we can use the factorization

$$p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}\right) = p_1\left(\mathbf{y}_1|\mathbf{X},\boldsymbol{\theta}_1\right) p_2\left(\mathbf{y}_2|\mathbf{X},\mathbf{y}_1,\boldsymbol{\theta}_2\right) \cdots p_K\left(\mathbf{y}_K|\mathbf{X},\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_{K-1},\boldsymbol{\theta}_K\right) \tag{4}$$

and specify a univariate likelihood for each $\mathbf{y}_k$.[6] This allows us to sequentially generate synthetic values $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1 \;\; \tilde{\mathbf{y}}_2 \;\; \cdots \;\; \tilde{\mathbf{y}}_K]$ using univariate conditional models. That is, sample $\tilde{\mathbf{y}}_1$ from the posterior predictive distribution of $\mathbf{y}_1$ given $\mathbf{X}$, then $\tilde{\mathbf{y}}_2$ from the posterior predictive distribution of $\mathbf{y}_2$ given $\mathbf{X}$ and $\tilde{\mathbf{y}}_1$, and so on. The joint posterior predictive density from which the synthetic values are sampled is:

$$\begin{aligned} p\left(\tilde{\mathbf{Y}}|\mathbf{X},\mathbf{Y}\right) \;\; = \;\; & \int p_1\left(\tilde{\mathbf{y}}_1|\mathbf{X},\boldsymbol{\theta}_1\right) p_2\left(\tilde{\mathbf{y}}_2|\mathbf{X},\tilde{\mathbf{y}}_1,\boldsymbol{\theta}_2\right) \cdots p_K\left(\tilde{\mathbf{y}}_K|\mathbf{X},\tilde{\mathbf{y}}_1,\tilde{\mathbf{y}}_2,...,\tilde{\mathbf{y}}_{K-1},\boldsymbol{\theta}_K\right) \\ & \times p_1\left(\boldsymbol{\theta}_1|\mathbf{X},\mathbf{y}_1\right) p_2\left(\boldsymbol{\theta}_2|\mathbf{X},\mathbf{y}_1,\mathbf{y}_2\right) \cdots p_K\left(\boldsymbol{\theta}_K|\mathbf{X},\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_{K-1},\mathbf{y}_K\right) d\boldsymbol{\theta}. (5) \end{aligned}$$

The sequential approach is very flexible. It is straightforward to accommodate continuous and discrete variables by specifying an appropriate likelihood for each $\mathbf{y}_k$. Likewise, it is possible to preserve complex relationships between variables through conditional dependence. Furthermore, as we saw in Section 2.2, the synthetic values maximize data utility and minimize disclosure risk, up to posterior uncertainty, when the true likelihood is known. A simple example serves to illustrate.

Example 1 (Normal linear regression) *Suppose the data provider wishes to generate synthetic values of confidential variable $\mathbf{y}_k$ conditional on a subset of variables in the database, $\mathbf{W} \subseteq \mathbf{D}$. Suppose further that*

$$\mathbf{y}_k \left| \left(\mathbf{W},\boldsymbol{\beta}_k,\sigma_k^2\right) \right. \sim N\left(\mathbf{W}\boldsymbol{\beta}_k,\sigma_k^2\mathbf{I}\right). \tag{6}$$

[6] An alternative, proposed by Abowd and Woodcock (2001) and based on the Sequential Regression Multivariate Imputation (SRMI) algorithm of Raghunathan et al. (2001), is to approximate $p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}\right)$ by a sequence of regression models. This is an iterative procedure, consisting of L rounds of synthesis. In each round, synthetic values are drawn sequentially for each $\mathbf{y}_k$, conditional on $\mathbf{X}$ and the most recently-drawn synthetic values for all other confidential variables. They define each univariate likelihoods using an appropriate generalized linear model.

The corresponding likelihood is that of the normal linear regression model. Synthetic values are easy to generate under the usual uninformative prior for $\boldsymbol{\beta}_k$ and σ_k^2. For each posterior draw $\left(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2\right)$, sample $\tilde{\mathbf{y}}_k$ from the normal distribution with conditional mean $\mathbf{W}\hat{\boldsymbol{\beta}}_k$ and variance $\hat{\sigma}_k^2$. The synthetic data have the same conditional distribution as the confidential data, up to posterior uncertainty about parameters $(\boldsymbol{\beta}_k, \sigma_k^2)$.

3 Data Utility When the Likelihood is Unknown

In most practical applications, the true likelihood of the confidential data given disclosable data is unknown. Mis-specifying $p\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}\right)$ necessarily compromises data utility.[7] To see this, note that the distribution of the synthetic values is completely determined by the posterior predictive distribution of the imputation model. If this differs from the distribution of the confidential data, data utility is compromised because $F\left(\mathbf{Y}, \mathbf{X}\right) \neq F\left(\tilde{\mathbf{Y}}, \mathbf{X}\right)$. Consider Example 1 and suppose the true conditional distribution of $\mathbf{y}_k$ is not given by (6). If the regression model defined by (6) is used to generate the synthetic values, the synthetic values will have a normal distribution conditional on $\mathbf{W}$, regardless of the true conditional distribution of $\mathbf{y}_k$. Thus any departure from normality in the conditional distribution of $\mathbf{y}_k$ induces mis-specification of the form $F\left(\tilde{\mathbf{y}}_k, \mathbf{W}|\boldsymbol{\beta}_k, \sigma_k^2\right) \neq F\left(\mathbf{y}_k, \mathbf{W}|\boldsymbol{\theta}_k\right)$.

This example demonstrates the important trade-off between data utility and simplicity of the imputation model. All else equal, the data provider will prefer to use simple models to generate the synthetic values, e.g., regression models or alternatives that impose little computational burden and are easy to interpret. However any mis-specification that arises from simplification of the likelihood compromises data utility. In this section, we develop two practical solutions. The first is applicable when the true likelihood is known up to a monotone transformation. The second is more generally applicable and preserves the conditional distribution of the confidential data, up to sampling error, on an arbitrary collection of subdomains.

3.1 Likelihood Known Up To a Monotone Transformation

In some practical applications, the likelihood is known up to a monotone transformation. For instance, many economic variables have highly skewed distributions. Subject to a monotone transformation (such as the natural logarithm) the conditional distribution is often well approximated by a normal distribution.

[7] Mis-specifying the likelihood may also affect disclosure risk, but there is no particular reason to expect it will increase.

Suppose the data provider seeks to generate synthetic values of a continuous variable $\mathbf{y}_k$. We assume there exists a $\mathbf{z}_k = g\left(\mathbf{y}_k\right)$ such that

$$\mathbf{z}_k \left|\left(\mathbf{W}, \boldsymbol{\theta}_k\right) \sim p_{z|W}\left(\mathbf{z}_k|\mathbf{W}, \boldsymbol{\theta}_k\right)\right. \tag{7}$$

for some subset $\mathbf{W}$ of the database $\mathbf{D}$, and where it is convenient to sample synthetic values from the posterior predictive distribution defined by the likelihood $p_{z|W}$. Assume further that g is monotone and bijective, so that g^{-1} exists, and is differentiable. Then we have the elementary result

$$p_{y|W}\left(\mathbf{y}_k|\mathbf{W}, \boldsymbol{\theta}_k\right) = p_{z|W}\left(g\left(\mathbf{y}_k\right)|\mathbf{W}, \boldsymbol{\theta}_k\right)\left|\frac{d}{d\mathbf{y}_k} g\left(\mathbf{y}_k\right)\right| \tag{8}$$

for those $\mathbf{y}_k = g^{-1}\left(\mathbf{z}_k\right)$ such that $p_{z|W}\left(\mathbf{z}_k|\mathbf{W}, \boldsymbol{\theta}_k\right) > 0$. This result is useful when it is difficult to sample directly from the posterior predictive distribution defined by the likelihood $p_{y|W}$, but easy to sample from the predictive distribution defined by $p_{z|W}$. This case is frequently encountered in practice, for instance the log-normal example given previously.

Typically the transformation g will be unknown. In principle, it can be estimated, e.g., the Box-Cox transformation. Of course any estimate $\hat{g}$ contains sampling error. Valid inferences based on (3) require that the imputation method is *proper* in the sense of Rubin (1987), i.e., propagates model uncertainty across the implicates. Hence it is critical to introduce between-implicate variation in the estimated transformation. This is easily accomplished, for instance by estimating g on an approximate Bayesian bootstrap sample of observations in each implicate.

Whether g is known or estimated, we can proceed by sampling synthetic values of $\mathbf{z}_k$ from the posterior predictive distribution

$$p_{\tilde{z}|W}\left(\tilde{\mathbf{z}}_k|\mathbf{W}, \mathbf{z}_k\right) = \int p_{z|W}\left(\tilde{\mathbf{z}}_k|\mathbf{W}, \boldsymbol{\theta}_k\right) p\left(\boldsymbol{\theta}_k|\mathbf{W}, \mathbf{z}_k\right) d\boldsymbol{\theta}_k \tag{9}$$

and define the synthetic values $\tilde{\mathbf{y}}_k = g^{-1}\left(\tilde{\mathbf{z}}_k\right)$ or $\tilde{\mathbf{y}}_k = \hat{g}^{-1}\left(\tilde{\mathbf{z}}_k\right)$, as appropriate. For known

g, the synthetic values are distributed according to

$$
\begin{aligned}
p_{\tilde{y}|W}\left(\tilde{\mathbf{y}}_k|\mathbf{W},\mathbf{z}_k\right) &= p_{\tilde{z}|W}\left(g\left(\tilde{\mathbf{y}}_k\right)|\mathbf{W},\mathbf{z}_k\right)\left|\frac{d}{d\tilde{\mathbf{y}}_k}g\left(\tilde{\mathbf{y}}_k\right)\right| \\
&= \int p_{z|W}\left(g\left(\tilde{\mathbf{y}}_k\right)|\mathbf{W},\boldsymbol{\theta}_k\right)p\left(\boldsymbol{\theta}_k|\mathbf{W},\mathbf{z}_k\right)d\boldsymbol{\theta}_k\left|\frac{d}{d\tilde{\mathbf{y}}_k}g\left(\tilde{\mathbf{y}}_k\right)\right| \\
&= \int p_{y|W}\left(\tilde{\mathbf{y}}_k|\mathbf{W},\boldsymbol{\theta}_k\right)\left|\frac{d}{d\tilde{\mathbf{z}}_k}g^{-1}\left(\tilde{\mathbf{z}}_k\right)\right|p\left(\boldsymbol{\theta}_k|\mathbf{W},\mathbf{z}_k\right)d\boldsymbol{\theta}_k\left|\frac{d}{d\tilde{\mathbf{y}}_k}g\left(\tilde{\mathbf{y}}_k\right)\right| \\
&= \int p_{y|W}\left(\tilde{\mathbf{y}}_k|\mathbf{W},\boldsymbol{\theta}_k\right)p\left(\boldsymbol{\theta}_k|\mathbf{W},\mathbf{z}_k\right)d\boldsymbol{\theta}_k \qquad (10)
\end{aligned}
$$

which is equivalent to the distribution we would have obtained had we synthesized $\mathbf{y}_k$ directly. When g is estimated, this equivalence only holds up to sampling error in $\hat{g}$. Equality (10) implies that partially synthetic data generated this way maximize data utility and minimize disclosure risk, up to posterior uncertainty in the imputation model and sampling error in the estimated transformation.

3.2 A Density-Based Transformation

In many situations, it is unlikely that a simple parametric transformation like the Box-Cox will satisfactorily map the distribution of $\mathbf{y}_k$ into a distribution from which it is convenient to sample synthetic values, e.g., that defined by a regression model. The leading example is when the conditional distribution of $\mathbf{y}_k$ is multi-modal. In this section, we develop a flexible method to generate synthetic values using simple imputation models that preserves the conditional distribution of the confidential variable on an arbitrary collection of subdomains.

Suppose the data provider wishes to generate synthetic values of a continuous variable $\mathbf{y}_k$ conditional on $\mathbf{W}$. Define an arbitrary partition of the conditioning data $\mathbf{W} = [\mathbf{W}_1 \;\; \mathbf{W}_2]$. In principle, either $\mathbf{W}_1$ or $\mathbf{W}_2$ may be empty. The data provider seeks to generate synthetic values using a convenient model based on

$$
\mathbf{z}_k\left|\left(\mathbf{W}_1,\mathbf{W_2}=\mathbf{w_2},\boldsymbol{\theta}_k\right)\right. \sim p_{z|W}\left(\mathbf{z}_k|\mathbf{W}_1,\mathbf{W_2}=\mathbf{w_2},\boldsymbol{\theta}_k\right) \qquad (11)
$$

where $\mathbf{z}_k$ is some transformation of $\mathbf{y}_k$ defined on the subdomain $\mathbf{W}_2 = \mathbf{w}_2$. We now define this transformation.

Let $\hat{K}$ denote a smoothed estimate of the cumulative distribution of $\mathbf{y}_k$ on the subdomain $\mathbf{W}_2 = \mathbf{w}_2$. For instance, the integrated kernel density $\hat{K}\left(\mathbf{y}_k|\mathbf{W}_2=\mathbf{w}_2\right) \approx F_{y|W_2=w_2}\left(\mathbf{y}_k|\mathbf{W}_2=\mathbf{w}_2\right)$, where $F_{y|W_2=w_2}$ is the marginal cdf of $\mathbf{y}_k$ on the subdomain $\mathbf{W}_2 = \mathbf{w}_2$. Let $P_{z|W_2=w_2}$ denote the cumulative distribution function associated with the likelihood $p_{z|W}$ obtained by averaging over $\mathbf{W}_1$. Now define the transformation

$$\mathbf{z}_k \equiv P^{-1}_{z|W_2=w_2}\left(\hat{K}\left(\mathbf{y}_k|\mathbf{W}_2 = \mathbf{w}_2\right)\right). \tag{12}$$

We see that $\mathbf{z}_k \sim P_{z|W_2=w_2}$ by construction. Let $\tilde{\mathbf{z}}_k$ denote synthetic values sampled from the posterior predictive distribution:

$$p_{\tilde{z}|W}\left(\tilde{\mathbf{z}}_k|\mathbf{W}_1, \mathbf{W_2} = \mathbf{w_2}, \mathbf{z}_k\right) = \int p_{z|W}\left(\tilde{\mathbf{z}}_k|\mathbf{W}_1, \mathbf{W_2} = \mathbf{w_2}, \boldsymbol{\theta}_k\right) p\left(\boldsymbol{\theta}_k|\mathbf{W}_1, \mathbf{W_2} = \mathbf{w_2}, \mathbf{z}_k\right) d\boldsymbol{\theta}_k. \tag{13}$$

The synthetic values $\tilde{\mathbf{y}}_k$ are defined by the inverse transformation

$$\tilde{\mathbf{y}}_k = \hat{K}^{-1}\left(P_{\tilde{z}|W_2=w_2}\left(\tilde{\mathbf{z}}_k\right)\right) \tag{14}$$

where $P_{\tilde{z}|W_2=w_2}$ is the cumulative distribution function associated with the predictive distribution $p_{\tilde{z}|W}$, again obtained by averaging over $\mathbf{W}_1$.[8] By construction, the synthetic values are distributed according to $\tilde{\mathbf{y}}_k \sim \hat{K}\left(\mathbf{y}_k|\mathbf{W}_2 = \mathbf{w}_2\right)$. Hence the synthetic values preserve the distribution of the true confidential values, up to sampling error in $\hat{K}$, on the subdomain $\mathbf{W}_2 = \mathbf{w}_2$. This procedure can be repeated for each subdomain defined by $\mathbf{W}_2$, preserving the distribution of $\mathbf{y}_k|\mathbf{W}_2$ up to sampling error in $\hat{K}$. This procedure maximizes data utility and minimizes disclosure risk, up to posterior uncertainty in the imputation model and sampling error in the estimated transformation, on these subdomains.

Because the transformation (12) and inverse transformation (14) are monotone, they preserve monotone and rank-order relationships between $\mathbf{y}_k$ and $\mathbf{W}_1$. As demonstrated by the simulation and empirical application that follow, many other features of the joint distribution of $\mathbf{y}_k$ and $\mathbf{W}_1$ are also preserved in the synthetic data.

As in Section 3.1, the transformation defined here is estimated and therefore contains sampling error. Once again, care must be taken to introduce between-implicate variation in the estimated transformation if valid inferences are to be obtained using equation (3). This is easily accomplished, for instance by estimating $\hat{K}$ on an approximate Bayesian bootstrap sample of observations on each subdomain of $\mathbf{W}_2$ in each implicate.

In principle, the partition of $\mathbf{W}$ into $\mathbf{W}_1$ and $\mathbf{W}_2$ is arbitrary. There is a trade-off, however. Increasing the number of variables in $\mathbf{W}_2$ preserves more dimensions of the distribution of $\mathbf{y}|\mathbf{W}$. However, it also reduces the number of observations in each subdomain,[9] thereby

[8] In the case where $P_{\tilde{Z}|W_2=w_2}$ is unknown, it can be estimated, for instance the integrated kernel density of the synthetic values $\tilde{\mathbf{z}}_k$ on the subdomain $\mathbf{W}_2 = \mathbf{w}_2$.

[9] E.g., the subdomains defined by the cross-classification of race by sex will contain more observations than the subdomains defined by the cross-classification of race by sex by age.

reducing the precision of the estimated distribution $\hat{K}$ and of the synthesis model. It also increases computational burden, since the density and synthesis model are estimated on each subdomain.

Example 2 (Normal linear regression) *Suppose the data provider wishes to use a normal linear regression model to generate synthetic values of* $\mathbf{y}_k$ *conditional on* $\mathbf{W}$*, but the distribution of* $\mathbf{y}_k|\mathbf{W}$ *is not normal. Let* $\mathbf{W}_1$ *be the set of continuous variables in* $\mathbf{W}$*, and* $\mathbf{W}_2$ *the set of categorical variables. Define the subdomains* $\mathbf{w}_2$ *according to the cells of the cross-classification of variables in* $\mathbf{W}_2$*. On each subdomain, estimate the integrated kernel density* $\hat{K}\left(\mathbf{y}_k|\mathbf{W}_2=\mathbf{w}_2\right)$ *on an approximate Bayesian bootstrap sample of observations. Define the transformed values* $\mathbf{z}_k=\Phi^{-1}\left(\hat{K}\left(\mathbf{y}_k|\mathbf{W}_2=\mathbf{w}_2\right)\right)$*, where* Φ *denotes the standard normal CDF. Then* $\mathbf{z}_k\sim N\left(0,1\right)$ *on each subdomain, by construction. Synthetic values* $\tilde{\mathbf{z}}_k$ *are sampled from the posterior predictive distribution defined by the normal linear regression of* $\mathbf{z}_k$ *on* $\mathbf{W}_1$ *under an uninformative prior. Averaged over* $\mathbf{W}_1$*, the synthetic values* $\tilde{\mathbf{z}}_k$ *have an approximately standard normal distribution. Define the inverse transformation* $\tilde{\mathbf{y}}_k=\hat{K}^{-1}\left(\Phi\left(\tilde{\mathbf{z}}_k\right)\right)$*. The synthetic and confidential values are identically distributed (up to sampling error) on the subdomain* $\mathbf{W}_2=\mathbf{w}_2$*, i.e.,* $\tilde{\mathbf{y}}_k\sim\hat{K}\left(\mathbf{y}_k|\mathbf{W}_2=\mathbf{w}_2\right)\approx F_{y|W_2=w_2}\left(\mathbf{y}_k|\mathbf{W}_2=\mathbf{w}_2\right)$*.*

3.2.1 Extension: Longitudinal Data

In longitudinal data, we frequently have repeated measurements on confidential variables. Preserving time series properties in the synthetic data necessitates conditioning the synthesis on multiple elements of the time series. For instance, if we denote the period t measurement of $\mathbf{y}_k$ by $\mathbf{y}_{k,t}$, we typically need to condition its synthesis on $\mathbf{y}_{k,t-1}$, $\mathbf{y}_{k,t+1}$, etc. to preserve the time series properties of $\mathbf{y}_k$. If we apply the density-based transformation to $\mathbf{y}_{k,t}$, we must treat other elements of the time series likewise. That is, apply the density-based transformation to $y_{k,t}$ and other elements of the time series that will be used to condition the imputation. The transformed values $\mathbf{z}_{k,t-1},\mathbf{z}_{k,t+1}$, etc. can be included in $\mathbf{W}_1$ and the synthesis proceeds as before.

4 Simulation

We illustrate and evaluate the synthesis methods described above with a brief simulation. We simulate 5,000 confidential databases, each comprising 10,000 observations on six variables. Of the six variables, we treat three as disclosable and three as confidential. We generate three partially synthetic implicates of each simulated database, as described below, to assess the quality and disclosure risk of the partially synthetic data.

The disclosable variables are defined as follows. The first, denoted g, takes value one or two with equal probability. We refer to g as an observation's *group.* The other disclosable variables are x_1 and x_2, independently distributed $N(0,1)$ and rounded to the nearest integer on $[-2,2]$.

The confidential variables are defined as follows. We begin by defining

$$\begin{aligned} z_1 &= 3g + \left(g^{1/2}/3\right) x_1 + \left(g^{1/2}/3\right) x_2 + \varepsilon_1 && (15) \\ z_2 &= 3g + \left(g^{1/2}/4\right) x_1 + \left(g^{1/2}/4\right) x_2 + \left(g^{1/2}/4\right) z_1 + \varepsilon_2 && (16) \\ z_3 &= x_1 - (g/2)^{1/2} x_2 + \varepsilon_3 && (17) \end{aligned}$$

where the "errors" are independently distributed $\varepsilon_1 \sim N(0, g/9), \varepsilon_2 \sim N(0, g/16)$, and $\varepsilon_3 \sim N(0, g/2)$. We define the confidential variables as $y_1 = \exp(z_1)$, $y_2 = \exp(z_2)$, and $y_3 = F_{y_3|g}^{-1}\left(F_{z_3|g}(z_3)\right)$ where $F_{y_3|g}$ is the cdf of a 70 : 30 mixture of a $N(g, g^2)$ and a $N(3g, g^2/4)$, and $F_{z_3|g}$ is the cdf of z_3 conditional on g.[10] Conditional on the observation's group g, the distributions of y_1 and y_2 are highly skewed and that of y_3 is bimodal. Subject to the monotone transformations $z_1 = \ln(y1)$, $z_2 = \ln(y_2)$, and $z_3 = F_{z_3|g}^{-1}\left(F_{y_{3|g}}(y_3)\right)$, however, they have normal conditional distributions in each group.

Because equation (17) implies that the distribution of y_3 depends only on x_1, x_2, and g, we synthesize this variable independently of y_1 and y_2. Equations (15) and (16) imply dependence between y_1 and y_2, so we synthesize these variables sequentially. We synthesize y_1 first, conditional on g, x_1, and x_2, and then synthesize y_2, conditional on g, x_1, x_2, and y_1.

To synthesize y_3, we follow the procedure outlined in Example 2 exactly, with $\mathbf{W}_1 = \{x_1, x_2\}$ and $\mathbf{W}_2 = g$. We synthesize y_1 and y_2 under two scenarios. In each scenario, we let $\mathbf{W}_2 = g$ for both variables, and let $\mathbf{W}_1 = \{x_1, x_2\}$ for y_1, and $\mathbf{W}_1 = \{x_1, x_2, y_1\}$ for y_2. In the first scenario, we presume the transformation that maps between (y_1, y_2) and (z_1, z_2) is known and synthesize these variables as described in Section 3.1. That is, we sequentially generate synthetic values following Example 1, after applying the exact (logarithmic) transformation to y_1 and y_2. In the second scenario, we presume the transformation is unknown and sequentially generate the synthetic values using the density-based transformation, following Example 2 for each variable. Under both scenarios, when synthesizing y_2 we apply the relevant transformation to y_1 on the right-hand side of the regression model (like the extension to longitudinal data in Section 3.2.1). Generating synthetic values under these two scenarios allows us to assess the information loss due to ignorance of the transformation.

We present several measures of synthetic data quality for group $g = 1$ in Tables 1 through

[10] Note that $z_3|g \sim N(0, 1+g)$. Likewise, $z_1|g \sim N(3g, g/3)$ and $z_2|g \sim N\left(3g\left[1 + g^{1/2}/4\right], [g/4]\left[1 + g^{1/2}/3\right]\right)$.

3. Results for group $g = 2$ are qualitatively similar, and are appendicized for brevity. The synthetic data replicate the statistical properties of the confidential data with considerable accuracy. The exact transformation does a better job of preserving the distribution of the confidential data than the density-based transformation does, but on net, the gains to knowing the transformation are rather small.

Table 1 reports the first four moments and selected percentiles of the univariate distribution of each confidential variable in the simulated true and synthetic data. By all measures, the distribution of synthetic data based on the exact (logarithmic) transformation is virtually identical to that of the true data. This is also true of synthetic values of y_3 generated using the density-based transformation. The distribution of synthetic values of y_1 and y_2 generated using the density-based transformation match the first two moments of the true data very closely, but are slightly less skewed and have somewhat thinner tails. For the most part, however, these discrepancies are small and are accompanied by larger standard errors than in the true data.

Table 2 presents product-moment and rank-order correlations. Correlations based on synthetic values computed using the exact transformation are indistinguishable from the true correlations to three decimal places. Rank-order correlations in the partially synthetic data computed using the density-based transformation are likewise indistinguishable from the true data. This is to be expected, since the density-based transformation preserves rank-order relationships. Product-moment correlations in these synthetic data are also very close to those in the true data, typically matching them to at least two decimal places.

To further assess the quality of the partially synthetic data, Table 3 presents estimated coefficients from the regression of $\ln(y_2)$ on $x_1, x_2, \ln(y_1)$ and an intercept. The estimated coefficients in the synthetic data correspond very closely to those obtained on the true data, with only minor discrepancies arising in the third decimal place. Model fit, as measured by root-MSE, is slightly worse in the synthetic data, which is to be expected.

We undertake a very conservative analysis of disclosure risk. In each simulation, we begin by averaging the synthetic values of each confidential variable across the three implicates. Then, in each of the 50 cells of the cross-classification of the disclosable variables $(g \times x_1 \times x_2)$, we compute the Mahalanobis distance between each synthetic record and each confidential record. The closest confidential record to each synthetic record constitutes a match. If a synthetic record is matched to its confidential source record, the record is deemed re-identified. If the synthetic record is matched to any other confidential record, the record is deemed not to have been re-identified. Our measure of disclosure risk is the re-identification rate in each cell: the proportion of records that are re-identified.

We argue that this provides a conservative measure of disclosure risk for two reasons.

First, it presumes that a malicious user knows which synthetic records in each implicate correspond to the same respondent. This is necessary for the intruder to average the synthetic values across implicates. Second, it presumes that the intruder has the maximum possible information available to re-identify records in the synthetic data: the confidential data themselves.[11]

The overall re-identification rate is very low, averaging 0.5 percent over the 5000 simulations with a standard deviation of 0.1 percent. Table 4 presents re-identification rates by cell of the cross-classification of disclosable variables. There is considerable variation in the re-identification rate across cells. This corresponds closely to the inverse of cell size. Re-identification is most common in the smallest cells. The four smallest cells average slightly more than 22 observations, and here the re-identification rate is about 4.75 percent. This is only slightly larger than the inverse of the cell size. That is, on average about 1.05 records in 22 are re-identified in the smallest cells. Re-identification is least common in the largest cell: 0.14 percent in the cell averaging 733 observations. Again, this is only slightly larger than the inverse of the cell size. In fact, on average 1.02 records are re-identified in each cell. Note that if synthetic records were randomly matched to confidential records, the expected number of re-identifications per cell is one.[12] Thus the partially synthetic data provide extremely good disclosure protection, with re-identification rates approaching the lower bound implied by purely random matching.

5 Application

We apply the density-based transformation of Section 3.2 to synthesize earnings and date of birth in the Longitudinal Employer-Household Dynamics (LEHD) Program database. The LEHD data are administrative. They are based on the universe of quarterly employment records collected by state agencies to administer the Unemployment Insurance (UI) system. The LEHD database integrates the UI employment reports with a variety of internal Census Bureau data sources to attach individual and business characteristics to the administrative records. See Abowd et al. (2006) for a detailed description of the LEHD data. We select a simple random sample of individuals employed in one state (whose identity is confidential) between 1990 and 2001.[13] The sample contains about 30 million quarterly employment

[11] Implicitly, we also assume that the intruder knows the disclosable variables are unperturbed. In attempting to re-identify records, the intruder therefore requires exact agreement on the disclosable variables.

[12] Domingo-Ferrer and Torra (2003) show that if two files contain n records on the same set of n respondents, the probability of correctly re-identifying exactly r respondents using a random matching strategy is $p(r) = \frac{1}{r!}\sum_{v=0}^{n-r}(-1)^{v}/v!$. It follows that the expected value of r is 1 for any n, or equivalently, the probability that a randomly selected record is re-identified is $1/n$.

[13] We cannot disclose the sampling rate for confidentiality reasons.

records on about 1 million individuals.

5.1 Synthesis Details

We synthesize date of birth and earnings sequentially, with earnings following date of birth. For each variable, the synthesis procedure follows Example 2.

Date of birth is integer-valued and reported with daily detail. Earnings are reported quarterly in dollars. We treat both distributions as continuous. We truncate the right tail of the distribution of earnings at \$1 million per quarter (the 99th percentile is less than \$40,000). This is primarily because the application is illustrative and we wish to facilitate computation – the truncated observations likely require a distinct synthesis model.

To synthesize date of birth, the conditioning set $\mathbf{W}_2$ includes sex, race, county of residence, and several indicators for missing data. To synthesize earnings, we define $\mathbf{W}_2$ as sex, race, full-time status, an indicator for foreign birth, major SIC division of the employer, and several indicators for missing data.

We estimate the integrated kernel density of earnings and birth date on an approximate Bayesian bootstrap sample of observations in each cell of the cross-classification of variables in $\mathbf{W}_2$ that contains sufficient data. We use an ad hoc rule to define "sufficient data": at least ten times as many observations as conditioning variables in the imputation model ($\mathbf{W}_1$). We collapse cells with insufficient data, in which case we add main effects for the collapsed cells to $\mathbf{W}_1$.[14] Following Example 2, we use the estimated distribution to transform the variable under synthesis so it has a standard normal distribution in each cell. For synthesizing earnings, we apply a similar transformation to up to two leads and lags of earnings at the same employer (where these exist).

For synthesizing date of birth, $\mathbf{W}_1$ includes an indicator for foreign birth, a quartic in years of education, annual summaries of earnings and quarters worked, the proportion of employment spells that were full-time, the proportion of employment spells in each major SIC division and county, the mean and variance of (log) firm size and payroll in the individual's employment history, and individual and firm main effects from an auxiliary regression of annualized earnings on various observable characteristics of workers and firms.[15] To synthesize earnings, we define $\mathbf{W}_1$ to include up to two transformed leads and lags of earnings at the same employer (where these exist), a quartic in education, a quartic in labor force experience

[14] The cross-classification of variables in $\mathbf{W}_2$ defines over 100,000 cells for each variable. Most of these contain little or no data, which necessitates collapsing many cells. Although only about ten percent of observations are in collapsed cells, the collapse reduces the number of cells below 1000 for date of birth, and below 3000 for earnings. Cell sizes vary between approximately 1500 and 1.4 million observations. The median cell size is approximately 3150 observations.

[15] See Abowd et al. (2003) for details on this auxiliary regression.

(which is a function of age),[16] main effects for county of residence and county of employment, main effects for non-employment in each year of the sample, the employer's (log) employment and payroll, main effects for year and quarter, and individual and firm main effects estimated in an auxiliary regression of annualized earnings on observable characteristics of workers and firms.

We use a normal linear regression model with uninformative prior to generate the synthetic values of the transformed variables. We estimate a separate regression model for each variable on each subdomain defined by $\mathbf{W}_2$. Since there are a large number of variables in the conditioning set $\mathbf{W}_1$, and since many of these are highly colinear, we apply a simple model selection subroutine to improve precision of the estimated posterior distribution of regression coefficients, as follows. On each subdomain, we estimate a candidate regression on all elements of $\mathbf{W}_1$. Only those variables that meet the Schwarz (1978) criterion are retained. We then estimate the final imputation model on the reduced set of conditioning variables.

We sample synthetic values from the posterior predictive distribution of the synthesis model subject to two restrictions. We restrict the parameter draw to lie within three standard deviations of the posterior mode, and restrict the synthetic values to lie within one standard deviation of the true value on the variable's natural scale. We then invert the density-based transformation, returning the synthetic values to their natural scale.

5.1.1 Results

We do not attempt to assess re-identification rates in the partially synthetic data, because synthesizing only these two variables is almost certainly insufficient to prevent re-identification.[17] Our discussion therefore focuses on the quality of the synthetic data.

Table 5 reports moments and percentiles of the marginal distributions of true and synthetic age and earnings. For both variables, the distribution of the synthetic variables match the confidential data very closely, though the synthetic distributions exhibit slightly lower dispersion, are slightly more symmetric, and have slightly thinner tails. This suggests some slight reversion to the mean. This tendency is also apparent on subdomains defined by the cross-classification of sex and race. For brevity, moments on these subdomains are appendicized. Plots of the estimated marginal densities are more illustrative. Figure 1 plots the estimated marginal density of age by race (additional plots by race and sex are also appen-

[16] In the first period that an individual appears in the data, her (initial) potential experience is calculated as the maximum of age minus years of education minus 6, and zero. In each subsequent quarter that the individual is employed, experience accumulates by 0.25.

[17] That is, the large number of unsynthesized variables on the file will be sufficient to re-identify many observations.

dicized). In each cell, the densities match very closely and reproduce the multiple modes of the age distribution. The synthetic densities are slightly more concentrated around the mean, however, particularly in the smaller cells.[18] Figure 2 plots the estimated marginal density of true and synthetic earnings between \$1 and \$40,000 by race (recall the 99th percentile of the distribution is less than \$40,000). Again, they are very similar in every case. The only notable discrepancy occurs for values below \$1000. These outcomes are slightly under-represented in the synthetic data.

Table 6 presents product-moment correlations in the true and synthetic data. For the synthesized variables, correlations between true and synthetic values are very high (0.82 for age, 0.96 for earnings). Correlations between age/earnings and other items in the database are replicated almost exactly in the synthetic data. Table 7 presents rank-order correlations. Again, rank-order correlations between true and synthetic values are very high (0.81 for age, 0.88 for earnings). The rank-order correlations between synthetic values and other items in the database closely reflect those for the true data, although the correlation between age and earnings is slightly attenuated (0.33 in the true data, 0.26 in the synthetic data).

Table 8 presents time series correlations of earnings in the true and synthetic data. Rank-order correlations are considerably stronger than product-moment correlations for all time periods. The rank-order correlations are slightly attenuated in the synthetic data, and product-moment correlations are slightly amplified. Overall, however, the synthetic data faithfully reproduce the time series properties of earnings.

We close our analysis of the synthetic data by considering a regression model of substantive economic interest. The model predicts the natural logarithm of quarterly earnings based on individual and employer characteristics for a sample of men employed full time.[19] This is a very well-studied specification. Coefficient estimates from the true and synthetic data are presented in Table 9. On the whole, the true and synthetic data yield very similar inferences. In particular, the experience profile is virtually identical in the two databases. The only notable discrepancies are in the education profile, which has the correct slope but is shifted downward by approximately 0.02 log points, and several of the industry main effects. As in the simulation exercise, model fit (as measured by root-MSE) is slightly worse in the synthetic data than the true data, as we would expect.

[18] This is due to cell collapsing. Observations in the smallest sex × race cells are more likely to be subject to cell collapse along these dimensions.

[19] Note the estimated specification differs from the synthesis model for earnings. In particular, it is based on log earnings, instead of the density-based transformation of earnings. Furthermore, the estimated specification includes main effects for foreign birth and the employer's industry, whereas the synthesis model is fully in interacted with these variables, and excludes leads and lags of earnings, main effects for county of residence and employment, main effects for non-employment in each year, and individual and firm main effects.

6 Conclusion

Statistical disclosure limitation methods promise high quality microdata with low disclosure risk. Among existing disclosure limitation methods, multiply-imputed partially synthetic data strike a compelling balance between these competing objectives. Indeed, the main virtue of this approach is that it preserves the ability of users to obtain valid statistical inferences about a population of interest. Furthermore, as we argue herein, MIPS data maximize data quality and minimize disclosure risk, up to posterior uncertainty in the imputation model. Our simulation supports this assertion, with simulated re-identification rates approaching the lower bound implied by random matching, while preserving the conditional distribution of confidential variables on pre-specified subdomains. Our application to LEHD data demonstrates the feasibility of our approach in large scale applications, and further illustrates the high quality of the partially synthetic data.

Like all model-based disclosure limitation methods, however, the quality of MIPS data depends on correctly specifying the imputation model used to generate the partially synthetic data. Our transformation-based methods address one form of mis-specification that arises when the joint distribution of the confidential data conditional on disclosable data is unknown. However, other forms of mis-specification are possible. In particular, MIPS data will only preserve multivariate relationships that are present in the imputation model. To preserve *all* multivariate relationships in the partially synthetic data requires, in principle, that the imputation model conditions on "everything." Of course, this is not possible in practice. We saw evidence of this in our application to LEHD data, where it was necessary to collapse some subdomains on which we sought to preserve the conditional distribution of age and earnings, and to reduce the number of conditioning variables in the imputation regressions though model selection. Further research is required to determine optimal methods for reducing the dimensionality of the synthesis problem.

It is important that data providers recognize and advertise the limitations of partially synthetic data they release. In particular, the model used to generate the MIPS data will make them well suited to some analyses and poorly suited to others. Data providers must therefore accompany any release of MIPS data with sufficient information for users to determine whether the MIPS data are appropriate for their analysis.

References

Abowd, J. M., P. Lengermann, and K. McKinney (2003, March). The measurement of human capital in the U.S. economy. Cornell University Working Paper.

Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2006). The LEHD infrastructure files and the creation of the quarterly workforce indicators. LEHD Technical Paper No. TP-2006-1.

Abowd, J. M. and S. D. Woodcock (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. I. Lane, J. J. Theeuwes, and L. V. Zayatz (Eds.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Chapter 10, pp. 215–278. North-Holland.

Domingo-Ferrer, J. and V. Torra (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing 13*, 343–354.

Duncan, G. T., S. A. Keller-McNulty, and S. L. Stokes (2001). Disclosure risk vs. data utility: The r-u confidentiality map. National Institute of Statistical Sciences Technical Report No. 121.

Duncan, G. T. and D. Lambert (1986). Disclosure-limited data dissemination. *J. American Statistical Association 81*(393), 10–18.

Elliot, M. (2001). Disclosure risk assessment. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. M. Zayatz (Eds.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Chapter 4, pp. 75–90. North-Holland.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Carnegie Mellon University Department of Statistics Technical Report No. 611.

Fienberg, S. E. and U. E. Makov (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics 14*(4), 385–397.

Fienberg, S. E., U. E. Makov, and R. J. Steele (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics 14*(4), 485–502.

Kennickell, A. B. (1997, November). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. SCF Working Paper.

Lin, L. I.-K. and E. F. Vonesh (1989). An empirical nonlinear data-fitting approach for transforming data to normality. *The American Statistician 43*(4), 237–243.

Little, R. and F. Liu (2003). Selective multiple imputation of keys for statistical disclosure control in microdata. The University of Michigan Department of Biostatistics Working Paper Series.

Muralidhar, K. and R. Sarathy (2003). A theoretical basis for perturbation methods. *Statistics and Computing 13*, 329–335.

Nusser, S., A. Carriquiry, K. Dodd, and W. Fuller (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association 91*(436), 1440–1449.

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

Raghunathan, T. E., J. M. Lepkowski, J. V. Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*(1), 85–95.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics 18*(4), 531–544.

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

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

Reiter, J. P. (2005a). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association 100*(472), 1103–1112.

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

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

Reiter, J. P. (2005d). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics 21*, 441–465.

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

Rubin, D. B. (1993). Discussion of statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Sarathy, R., K. Muralidhar, and R. Parsa (2002). Perturbing nonnormal confidential attributes: The copula approach. *Management Science 48*(12), 1613–1627.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Willenborg, L. and T. de Waal (1996). *Statistical Disclosure Control in Practice*. Springer-Verlag.

Winkler, W. E. (2004). Re-identification methods for masked microdata. In J. Doming-Ferrer and V. Torra (Eds.), *Privacy in Statistical Databases*, pp. 216–230. Springer. Lecture Notes in Computer Science 3050.

TABLE 1

Marginal Distribution of Simulated True and Synthetic Values in Group *g* = 1

	y_1			y_2			y_3	
	True	Synthetic (Exact Transform)	Synthetic (Density-based Transform)	True	Synthetic (Exact Transform)	Synthetic (Density-based Transform)	True	Synthetic (Density-based Transform)
Moments								
Mean	23.8	23.8	23.7	49.3	49.4	49.2	1.60	1.60
	(0.21)	(0.26)	(0.24)	(0.21)	(1.59)	(0.46)	(0.03)	(0.03)
Standard Deviation	14.9	14.9	14.4	28.5	28.6	27.8	1.30	1.30
	(0.31)	(0.33)	(0.35)	(0.54)	(1.08)	(0.65)	(0.01)	(0.02)
Skewness	1.93	1.93	1.69	1.72	1.72	1.53	-0.12	-0.10
	(0.16)	(0.10)	(0.16)	(0.12)	(0.08)	(0.13)	(0.03)	(0.04)
Kurtosis	6.59	6.61	4.87	4.99	5.01	3.77	-0.81	-0.82
	(1.91)	(1.18)	(1.58)	(1.15)	(0.75)	(1.07)	(0.04)	(0.05)
Percentiles								
1st	5.28	5.28	4.62	12.2	12.3	11.2	-1.21	-1.18
	(0.16)	(0.13)	(0.20)	(0.32)	(0.46)	(0.42)	(0.07)	(0.08)
5th	7.72	7.72	7.55	17.4	17.4	17.1	-0.51	-0.51
	(0.13)	(0.13)	(0.16)	(0.27)	(0.59)	(0.33)	(0.04)	(0.05)
50th	20.1	20.1	20.3	42.5	42.7	42.9	1.57	1.58
	(0.21)	(0.22)	(0.26)	(0.42)	(1.37)	(0.50)	(0.04)	(0.05)
95th	52.3	52.3	51.4	104	105	103	3.57	3.58
	(0.90)	(0.89)	(1.09)	(1.63)	(3.66)	(2.06)	(0.03)	(0.03)
99th	76.5	76.5	73.2	148	149	143	4.01	4.03
	(2.23)	(1.89)	(2.66)	(3.92)	(5.71)	(4.96)	(0.04)	(0.04)

Notes: Main entry in each column is the sample mean of the statistic in 5000 simulations. Simulated standard errors are in parentheses. In each simulation, statistics based on synthetic data are computed in each synthetic implicate. Statistics are averaged over implicates before computing the mean and standard error over simulations.

TABLE 2

Correlations in Simulated True and Synthetic Data for Group g = 1

	Product-Moment Correlations				Rank-Order Correlations			
	x_1	x_2	y_1	y_2	x_1	x_2	y_1	y_2
True Data								
x_2	0.000				0.000			
y_1	0.533	0.533			0.567	0.567		
y_2	0.576	0.576	0.774		0.606	0.607	0.794	
y_3	0.706	-0.497	0.112	0.121	0.700	-0.487	0.119	0.126
Synthetic Data, Exact Transform								
y_1	0.533	0.533			0.567	0.567		
y_2	0.576	0.576	0.774		0.606	0.607	0.794	
Synthetic Data, Density-Based Transform								
y_1	0.540	0.540			0.567	0.567		
y_2	0.582	0.582	0.781		0.606	0.606	0.794	
y_3	0.706	-0.497	0.114	0.122	0.699	-0.487	0.118	0.126

Notes: Entry in each column is the sample mean of the statistic in 5000 simulations. Simulated standard errors are available on request. All standard errors are less than 0.016. In each simulation, statistics based on synthetic data are computed in each of three synthetic implicates. Statistics are averaged over implicates before computing the mean and standard error over simulations.

TABLE 3

Estimated Regression Coefficients in Simulated True and Synthetic Data, Group *g* = 1

	True	Synthetic (Exact Transform)	Synthetic (Density-Based Transform)
Intercept	3.00	3.00	3.00
	(0.032)	(0.048)	(0.040)
x_1	0.250	0.250	0.252
	(0.005)	(0.007)	(0.007)
x_2	0.250	0.250	0.252
	(0.005)	(0.007)	(0.007)
$\ln(y_1)$	0.250	0.250	0.249
	(0.011)	(0.012)	(0.013)
RMSE	0.250	0.250	0.255
	(0.003)	(0.003)	(0.004)
Number of Observations	5000	5000	5000

Notes: Dependent variable is $\ln(y_2)$. Main entry in each column is the sample mean of the statistic in 5000 simulations. Simulated standard errors are in parentheses. In each simulation, statistics based on synthetic data are computed in each of three synthetic implicates. Statistics are averaged over implicates before computing the mean and standard error over simulations.

TABLE 4

Simulated Re-identification Rates by Cell in Group *g* = 1

Value of x_1	Value of x_2				
	-2	-1	0	1	2
Synthetic Data, Density-Based Transform					
-2	0.047 (0.047) N = 22.3	0.013 (0.012) N = 80.7	0.008 (0.008) N = 128	0.013 (0.012) N = 80.5	0.048 (0.048) N = 22.4
-1	0.013 (0.013) N = 80.7	0.003 (0.003) N = 292	0.002 (0.002) N = 463	0.003 (0.003) N = 292	0.013 (0.013) N = 80.7
0	0.008 (0.008) N = 128	0.002 (0.002) N = 463	0.001 (0.001) N = 733	0.002 (0.002) N = 463	0.008 (0.008) N = 128
1	0.013 (0.013) N = 81.0	0.003 (0.003) N = 292	0.002 (0.002) N = 463	0.003 (0.003) N = 293	0.013 (0.013) N = 80.7
2	0.048 (0.047) N = 22.4	0.012 (0.012) N = 80.7	0.008 (0.008) N = 128	0.012 (0.012) N = 80.8	0.047 (0.046) N = 22.4
Synthetic Data, Exact Transform					
-2	0.046 (0.047)	0.013 (0.013)	0.008 (0.008)	0.013 (0.012)	0.049 (0.047)
-1	0.013 (0.013)	0.003 (0.003)	0.002 (0.002)	0.003 (0.003)	0.013 (0.012)
0	0.008 (0.008)	0.002 (0.002)	0.001 (0.001)	0.002 (0.002)	0.008 (0.008)
1	0.012 (0.012)	0.003 (0.004)	0.002 (0.002)	0.003 (0.003)	0.013 (0.012)
2	0.048 (0.047)	0.012 (0.012)	0.008 (0.008)	0.012 (0.013)	0.047 (0.046)

Notes: First entry in each cell is the average re-identification rate in that cell in 5000 simulations. The second entry, in parentheses, is the standard deviation of the re-identification rate in that cell in 5000 simulations. The third entry in the top panel is the average number of observations in that cell in 5000 simulations. Both panels are based on the same simulated data.

TABLE 5
Marginal Distribution of True and Synthetic Values

	Age on Jan. 1, 1990			Quarterly Employment Earnings		
	True Value	Synthetic Value	Between-Implicate Std. Dev	True Value	Synthetic Value	Between-Implicate Std. Dev
Moments						
Mean	29.7	29.7	(0.12)	6,731	6,715	(18.1)
Standard Deviation	16.3	15.8	(1.42)	14,024	13,291	(696)
Skewness	0.53	0.48	(0.02)	31.4	30.3	(0.11)
Kurtosis	-0.10	-0.05	(0.04)	1,722	1,678	(9.19)
Percentiles						
1st	1.43	1.10	(0.22)	39.0	1.00	(0.00)
5th	6.78	6.86	(0.10)	160	185	(0.00)
50th	28.0	28.3	(0.19)	4,546	4,523	(20.6)
95th	59.8	58.1	(0.18)	18,218	18,354	(23.7)
99th	72.1	71.0	(0.16)	38,200	37,304	(31.4)
Number of Observations		1,286,444			29,991,540	

Source: Authors' calculations based on the LEHD database.

Notes: Statistics in the columns labeled “Synthetic Value” are averaged over three synthetic data implicates. The distribution of true and synthetic earnings is truncated at one and one million dollars.

TABLE 6
Product-Moment Correlations in True and Synthetic Data

	Age on Jan. 1, 1990			Quarterly Employment Earnings		
	True Value	Synthetic Value	Between-Implicate Std. Dev	True Value	Synthetic Value	Between-Implicate Std. Dev
Individual Characteristics						
True Age (years)	**1**	**0.818**	**(0.000)**	**0.143**	**0.146**	**(0.000)**
Synthetic Age	**0.818**	**1**	**(0.000)**	**0.129**	**0.132**	**(0.003)**
Education (years)	0.155	0.151	(0.004)	0.143	0.144	(0.001)
Male (0 or 1)	0.007	0.002	(0.016)	0.129	0.135	(0.001)
Foreign Born (0 or 1)	0.040	0.044	(0.002)	-0.003	-0.004	(0.001)
Race = Black (0 or 1)	-0.027	-0.028	(0.031)	-0.053	-0.056	(0.001)
Race = Hispanic (0 or 1)	-0.104	-0.108	(0.008)	-0.046	-0.048	(0.000)
Employment Characteristics						
Earnings (Dollars)	**0.143**	**0.129**	**(0.003)**	**1**	**0.960**	**(0.002)**
Synthetic Earnings	**0.146**	**0.132**	**(0.003)**	**0.960**	**1**	**(0.000)**
Not Employed in 1990 (0 or 1)	-0.356	-0.382	(0.004)	-0.118	-0.124	(0.000)
Not Employed in 1991 (0 or 1)	-0.326	-0.350	(0.004)	-0.111	-0.118	(0.000)
Not Employed in 1992 (0 or 1)	-0.302	-0.320	(0.003)	-0.115	-0.122	(0.000)
Not Employed in 1993 (0 or 1)	-0.265	-0.278	(0.004)	-0.108	-0.115	(0.000)
Not Employed in 1994 (0 or 1)	-0.229	-0.239	(0.004)	-0.105	-0.112	(0.001)
Not Employed in 1995 (0 or 1)	-0.184	-0.190	(0.004)	-0.099	-0.106	(0.001)
Not Employed in 1996 (0 or 1)	-0.140	-0.143	(0.004)	-0.095	-0.102	(0.001)
Not Employed in 1997 (0 or 1)	-0.092	-0.093	(0.004)	-0.090	-0.096	(0.001)
Not Employed in 1998 (0 or 1)	-0.040	-0.039	(0.004)	-0.081	-0.087	(0.001)
Not Employed in 1999 (0 or 1)	0.005	0.007	(0.004)	-0.073	-0.079	(0.001)
Not Employed in 2000 (0 or 1)	0.044	0.048	(0.003)	-0.063	-0.069	(0.001)
Not Employed in 2001 (0 or 1)	0.066	0.072	(0.003)	-0.057	-0.062	(0.000)
Employed Full Time (0 or 1)	0.037	0.043	(0.002)	0.027	0.029	(0.001)
Full Time Missing (0 or 1)	0.029	0.023	(0.002)	0.051	0.054	(0.002)
Employer Characteristics						
ln(Number of Employees)	0.058	0.059	(0.006)	0.065	0.069	(0.000)
ln(Total Payroll)	0.095	0.097	(0.005)	0.143	0.148	(0.000)
SIC Division = A (0 or 1)	-0.031	-0.033	(0.001)	-0.019	-0.019	(0.000)
SIC Division = B (0 or 1)	0.018	0.020	(0.003)	0.010	0.010	(0.001)
SIC Division = C (0 or 1)	0.014	0.012	(0.003)	0.011	0.012	(0.002)
SIC Division = E (0 or 1)	0.025	0.028	(0.001)	0.035	0.036	(0.002)
SIC Division = F (0 or 1)	0.020	0.024	(0.002)	0.046	0.049	(0.001)
SIC Division = G (0 or 1)	-0.184	-0.190	(0.003)	-0.117	-0.123	(0.000)
SIC Division = H (0 or 1)	0.009	0.012	(0.003)	0.073	0.076	(0.002)
SIC Division = I (0 or 1)	0.031	0.027	(0.006)	-0.038	-0.040	(0.001)
SIC Division = J (0 or 1)	0.073	0.072	(0.001)	0.011	0.011	(0.001)

Source: Authors' calculations based on the LEHD database.

Notes: To facilitate computation, all correlations are computed on a 2 percent random sample of observations. Statistics in the columns labeled “Synthetic Value” are averaged over three synthetic data implicates. The distribution of true and synthetic earnings is truncated at one and one million dollars.

TABLE 7
Rank-Order Correlations in True and Synthetic Data

	Age on Jan. 1, 1990			Quarterly Employment Earnings		
	True Value	Synthetic Value	Between-Implicate Std. Dev	True Value	Synthetic Value	Between-Implicate Std. Dev
Individual Characteristics						
True Age (years)	**1**	**0.810**	**(0.000)**	**0.326**	**0.319**	**(0.002)**
Synthetic Age	**0.810**	**1**	**(0.000)**	**0.265**	**0.259**	**(0.004)**
Education (years)	0.202	0.179	(0.005)	0.272	0.262	(0.002)
Male (0 or 1)	0.008	0.003	(0.016)	0.199	0.199	(0.001)
Foreign Born (0 or 1)	0.047	0.046	(0.002)	0.011	0.011	(0.002)
Race = Black (0 or 1)	-0.024	-0.026	(0.031)	-0.084	-0.086	(0.001)
Race = Hispanic (0 or 1)	-0.105	-0.110	(0.009)	-0.070	-0.069	(0.002)
Employment Characteristics						
Earnings (Dollars)	**0.326**	**0.265**	**(0.002)**	**1**	**0.879**	**(0.000)**
Synthetic Earnings	**0.319**	**0.259**	**(0.004)**	**0.879**	**1**	**(0.000)**
Not Employed in 1990 (0 or 1)	-0.380	-0.398	(0.003)	-0.311	-0.308	(0.000)
Not Employed in 1991 (0 or 1)	-0.339	-0.354	(0.003)	-0.287	-0.286	(0.000)
Not Employed in 1992 (0 or 1)	-0.322	-0.329	(0.002)	-0.300	-0.300	(0.000)
Not Employed in 1993 (0 or 1)	-0.282	-0.283	(0.002)	-0.280	-0.280	(0.000)
Not Employed in 1994 (0 or 1)	-0.268	-0.264	(0.003)	-0.307	-0.307	(0.001)
Not Employed in 1995 (0 or 1)	-0.227	-0.218	(0.003)	-0.294	-0.294	(0.001)
Not Employed in 1996 (0 or 1)	-0.186	-0.174	(0.003)	-0.283	-0.284	(0.000)
Not Employed in 1997 (0 or 1)	-0.143	-0.128	(0.003)	-0.268	-0.270	(0.000)
Not Employed in 1998 (0 or 1)	-0.095	-0.078	(0.003)	-0.246	-0.248	(0.000)
Not Employed in 1999 (0 or 1)	-0.049	-0.033	(0.003)	-0.224	-0.226	(0.000)
Not Employed in 2000 (0 or 1)	-0.008	0.010	(0.003)	-0.199	-0.200	(0.000)
Not Employed in 2001 (0 or 1)	0.019	0.037	(0.002)	-0.181	-0.182	(0.001)
Employed Full Time (0 or 1)	0.058	0.057	(0.001)	0.142	0.142	(0.003)
Full Time Missing (0 or 1)	0.029	0.022	(0.002)	0.068	0.069	(0.003)
Employer Characteristics						
ln(Number of Employees)	0.071	0.063	(0.005)	0.199	0.195	(0.000)
ln(Total Payroll)	0.124	0.114	(0.005)	0.347	0.339	(0.000)
SIC Division = A (0 or 1)	-0.034	-0.034	(0.001)	-0.044	-0.042	(0.001)
SIC Division = B (0 or 1)	0.020	0.022	(0.002)	0.031	0.029	(0.003)
SIC Division = C (0 or 1)	0.018	0.015	(0.004)	0.043	0.043	(0.004)
SIC Division = E (0 or 1)	0.033	0.032	(0.001)	0.099	0.100	(0.002)
SIC Division = F (0 or 1)	0.023	0.026	(0.002)	0.096	0.097	(0.001)
SIC Division = G (0 or 1)	-0.205	-0.199	(0.002)	-0.286	-0.286	(0.001)
SIC Division = H (0 or 1)	0.011	0.013	(0.003)	0.101	0.101	(0.003)
SIC Division = I (0 or 1)	0.032	0.027	(0.006)	-0.108	-0.106	(0.002)
SIC Division = J (0 or 1)	0.074	0.071	(0.002)	0.073	0.070	(0.003)

Source: Authors' calculations based on the LEHD database.

Notes: To facilitate computation, all correlations are computed on a 2 percent random sample of observations. Statistics in the columns labeled “Synthetic Value” are averaged over three synthetic data implicates. The distribution of true and synthetic earnings is truncated at one and one million dollars.

TABLE 8
Time Series Correlation of Earnings in True and Synthetic Data

	Product-Moment Correlations					Rank-Order Correlations				
	t-2	*t-1*	*t*	*t+1*	*t+2*	*t-2*	*t-1*	*t*	*t+1*	*t+2*
True Data										
t-2	1					1				
t-1	0.542	1				0.921	1			
t	0.502	0.533	1			0.895	0.903	1		
t+1	0.512	0.509	0.554	1		0.885	0.896	0.903	1	
t+2	0.703	0.509	0.521	0.523	1	0.903	0.885	0.896	0.920	1
Synthetic Data										
t-2	1					1				
t-1	0.577	1				0.875	1			
t	0.530	0.549	1			0.863	0.852	1		
t+1	0.543	0.544	0.596	1		0.847	0.858	0.862	1	
t+2	0.698	0.544	0.560	0.602	1	0.839	0.841	0.864	0.883	1

Source: Authors' calculations based on the LEHD database.

Notes: To facilitate computation, all correlations are computed on a 2 percent random sample of observations. Statistics labeled “Synthetic Data” are averaged over three synthetic data implicates. Between-implicate standard deviations are available on request; all are less than 0.005. The distribution of true and synthetic earnings is truncated at one and one million dollars.

TABLE 9
Estimated Regression Coefficients in True and Synthetic Data

	True Data		Synthetic Data	
	Coefficient	Std. Error	Coefficient	Std. Error
Years of Experience	0.058	(0.000)	0.058	(0.001)
Experience^2/100	-0.258	(0.005)	-0.264	(0.006)
Experience^3/1000	0.048	(0.002)	0.049	(0.002)
Experience^4/10000	-0.003	(0.000)	-0.003	(0.000)
Initial Experience < 0	-0.181	(0.005)	-0.177	(0.002)
Years of Education	0.023	(0.010)	0.015	(0.005)
Education^2/100	-0.930	(0.111)	-0.854	(0.095)
Education^3/1000	1.09	(0.120)	1.05	(0.070)
Education^4/10000	-0.295	(0.041)	-0.288	(0.018)
Race = Black	-0.073	(0.002)	-0.083	(0.007)
Race = Hispanic	-0.132	(0.003)	-0.146	(0.017)
Foreign Born = 1	-0.031	(0.002)	-0.025	(0.007)
ln(Number of Employees)	-0.233	(0.001)	-0.247	(0.001)
ln(Total Payroll)	0.275	(0.001)	0.291	(0.001)
SIC Division = A	-0.261	(0.008)	-0.203	(0.048)
SIC Division = B	-0.068	(0.024)	0.023	(0.047)
SIC Division = C	-0.008	(0.008)	-0.041	(0.012)
SIC Division = E	0.044	(0.002)	0.076	(0.010)
SIC Division = F	0.040	(0.003)	0.066	(0.013)
SIC Division = G	-0.323	(0.002)	-0.278	(0.007)
SIC Division = H	0.050	(0.002)	0.073	(0.007)
SIC Division = I	-0.159	(0.002)	-0.131	(0.011)
SIC Division = J	-0.119	(0.002)	-0.101	(0.003)
Year = 1991	0.026	(0.002)	0.028	(0.002)
Year = 1992	0.072	(0.002)	0.070	(0.003)
Year = 1993	0.101	(0.001)	0.103	(0.002)
Year = 1994	0.132	(0.001)	0.134	(0.003)
Year = 1995	0.154	(0.001)	0.156	(0.002)
Year = 1996	0.179	(0.001)	0.183	(0.003)
Year = 1997	0.213	(0.001)	0.217	(0.003)
Year = 1998	0.245	(0.001)	0.250	(0.003)
Quarter = 2	0.048	(0.001)	0.048	(0.001)
Quarter = 3	0.011	(0.001)	0.015	(0.001)
Quarter = 4	0.084	(0.001)	0.092	(0.001)
Intercept	5.05	(0.036)	4.92	(0.021)
RMSE	0.741		0.787	
Number of Observations	13,140,425		13,140,425	

Notes: Dependent variable is the natural logarithm of quarterly employment earnings. Sample is restricted to full-quarter observations on males employed full time. An individual is defined as working a full-quarter in period *t* if she was employed at the same firm in periods *t-1*, *t*, and *t+1*. All calculations on synthetic data are based on three partially synthetic implicates.

FIGURE 1
Estimated Density of True and Synthetic Age on Jan 1 1990

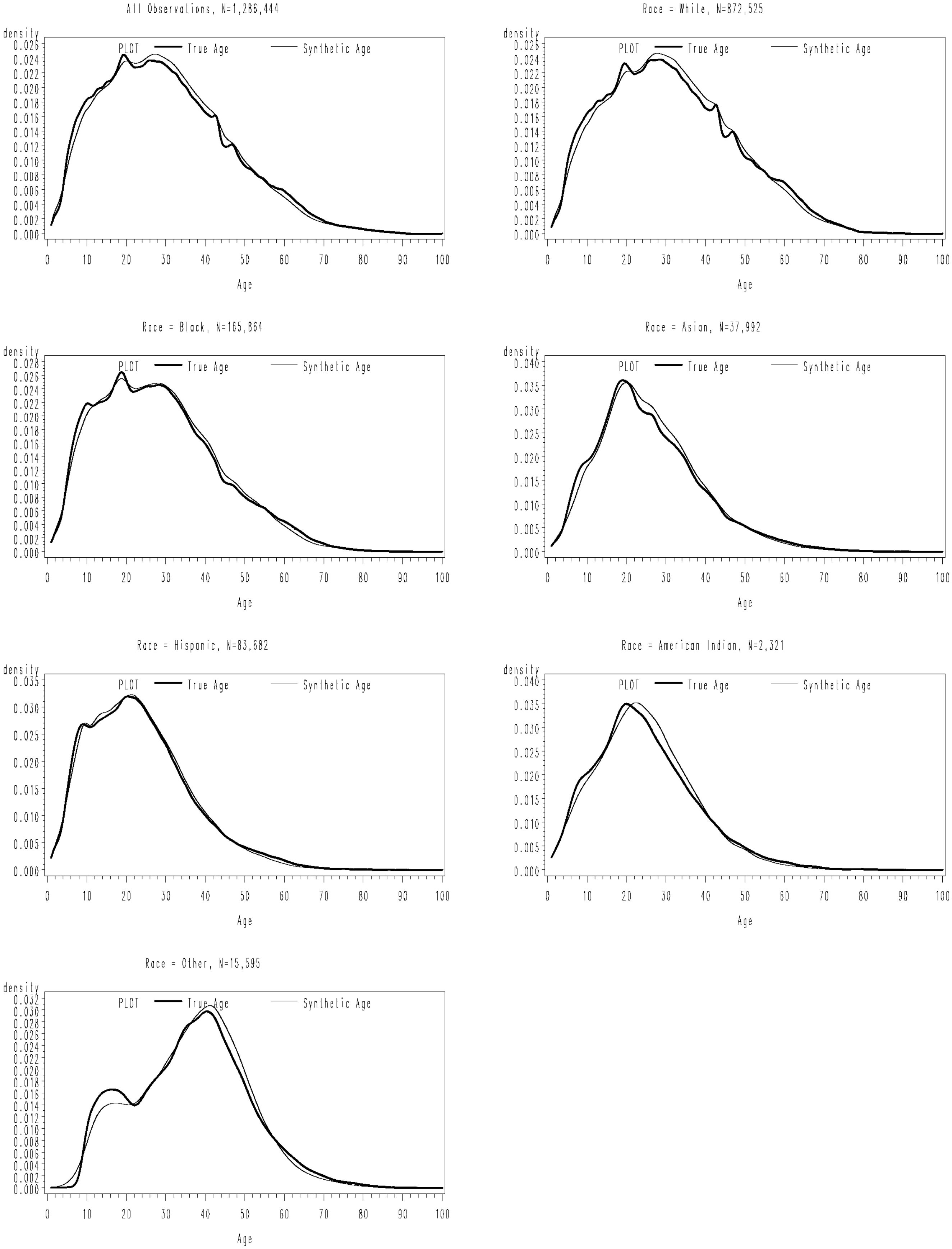
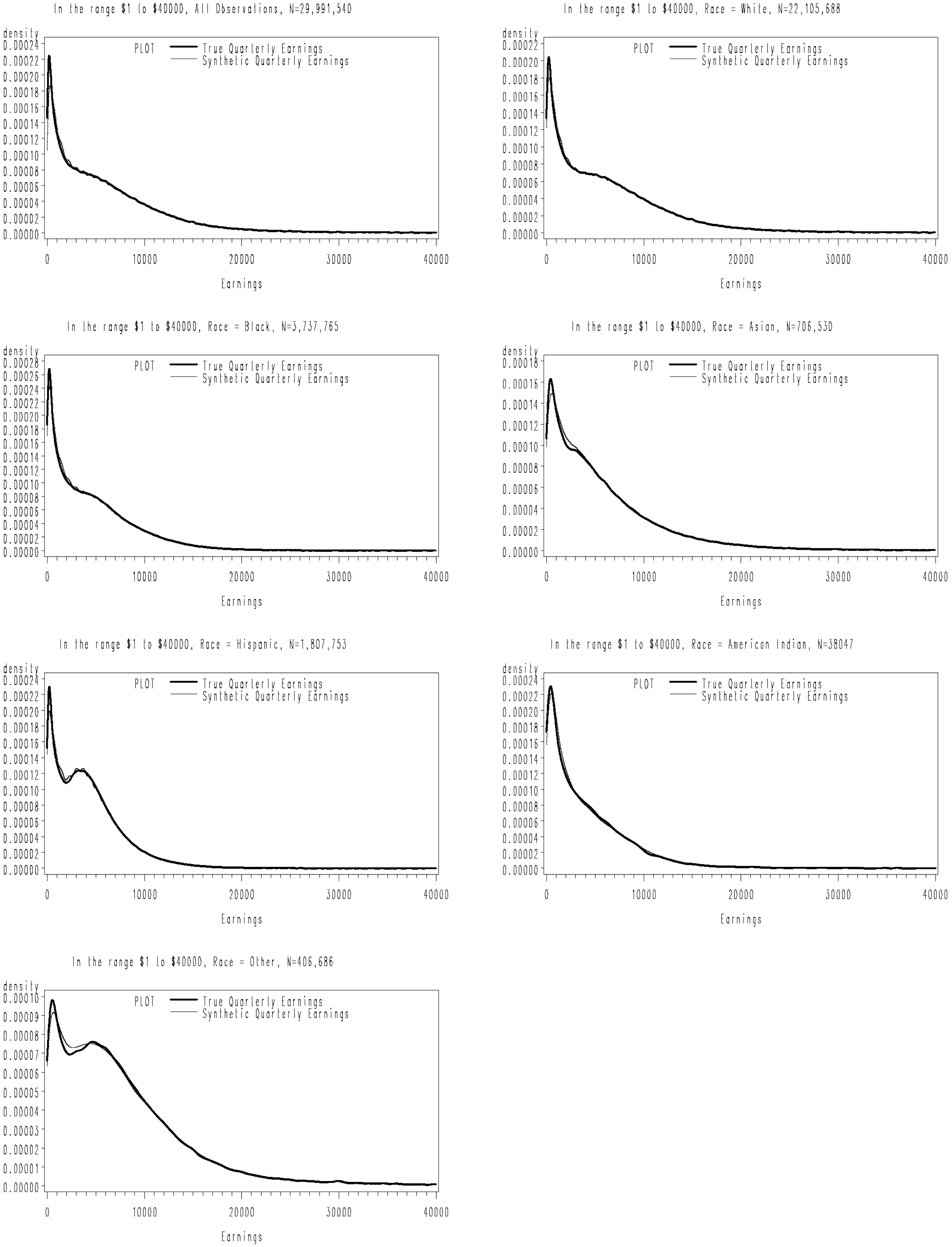


FIGURE 2
Estimated Density of True and Synthetic Quarterly Earnings



In the range $1 to $40000, All Observations, N=29,991,540

In the range $1 to $40000, Race = White, N=22,105,688

In the range $1 to $40000, Race = Black, N=3,737,765

In the range $1 to $40000, Race = Asian, N=706,530

In the range $1 to $40000, Race = Hispanic, N=1,807,753

In the range $1 to $40000, Race = American Indian, N=38047

In the range $1 to $40000, Race = Other, N=406,686

APPENDIX TABLE 1

Marginal Distribution of Simulated True and Synthetic Values in Group g = 2

	y_1			y_2			y_3	
	True	Synthetic (Exact Transform)	Synthetic (Density-based Transform)	True	Synthetic (Exact Transform)	Synthetic (Density-based Transform)	True	Synthetic (Exact Transform)
Moments								
Mean	564	564	557	4772	4804	4722	3.20	3.20
	(7.53)	(8.98)	(8.57)	(65.4)	(307)	(73.0)	(0.05)	(0.06)
Standard Deviation	538	539	497	4632	4666	4344	2.59	2.59
	(18.0)	(16.9)	(18.7)	(145)	(327)	(168)	(0.03)	(0.03)
Skewness	3.21	3.22	2.69	3.12	3.13	2.70	-0.12	-0.10
	(0.47)	(0.30)	(0.37)	(0.37)	(0.23)	(0.34)	(0.03)	(0.04)
Kurtosis	19.5	19.7	13.2	17.6	17.6	12.7	-0.79	-(0.8)
	(10.03)	(7.23)	(6.2)	(7.29)	(4.34)	(5.2)	(0.05)	(0.1)
Percentiles								
1st	61.0	61.0	29.8	493	496	245	-2.44	-2.39
	(2.46)	(2.11)	(4.22)	(19.4)	(34.5)	(34.8)	0.15	0.15
5th	104	104	96.9	845	851	783	-1.02	-1.01
	(2.48)	(2.54)	(3.31)	(20.7)	(55.4)	(27.7)	(0.09)	(0.10)
50th	403	403	414	3366	3388	3435	3.15	3.15
	(5.94)	(6.09)	(7.34)	(51.6)	(214)	(61.1)	(0.08)	(0.09)
95th	1562	1561	1501	13429	13516	13037	7.13	7.16
	(37.2)	(37.0)	(44.5)	(328)	(902)	(405)	(0.06)	(0.06)
99th	2673	2673	2463	23039	23207	21538	8.03	8.08
	(110)	(91.8)	(126)	(940)	(1665)	(1152)	(0.08)	(0.09)

Notes: Main entry in each column is the sample mean of the statistic in 5000 simulations. Simulated standard errors are in parentheses. In each simulation, statistics based on synthetic data are computed in each synthetic implicate. Statistics are averaged over implicates before computing the mean and standard error over simulations.

APPENDIX TABLE 2

Correlations in Simulated True and Synthetic Data for Group g = 2

	Product-Moment Correlations				Rank-Order Correlations			
	x_1	x_2	y_1	y_2	x_1	x_2	y_1	y_2
True Data								
x_2	0.000				0.000			
y_1	0.490	0.490			0.567	0.567		
y_2	0.530	0.529	0.782		0.614	0.614	0.829	
y_3	0.576	-0.576	0.001	0.001	0.567	-0.567	0.000	0.000
Synthetic Data, Exact Transform								
y_1	0.490	0.490			0.567	0.566		
y_2	0.529	0.529	0.782		0.614	0.613	0.829	
Synthetic Data, Density-Based Transform								
y_1	0.504	0.504			0.566	0.566		
y_2	0.543	0.542	0.798		0.612	0.611	0.829	
y_3	0.576	-0.576	0.001	0.001	0.567	-0.567	0.000	0.000

Notes: Entry in each column is the sample mean of the statistic in 5000 simulations. Simulated standard errors are available on request. All standard errors are less than 0.016. In each simulation, statistics based on synthetic data are computed in each of three synthetic implicates. Statistics are averaged over implicates before computing the mean and standard error over simulations.

APPENDIX TABLE 3

Estimated Regression Coefficients in Simulated True and Synthetic Data, Group *g* = 2

	True	Synthetic (Exact Transform)	Synthetic (Density-Based Transform)
Intercept	6.00	6.00	6.03
	(0.065)	(0.096)	(0.089)
x_1	0.354	0.354	0.360
	(0.007)	(0.009)	(0.010)
x_2	0.353	0.353	0.360
	(0.007)	(0.009)	(0.010)
$ln(y_1)$	0.354	0.354	0.349
	(0.011)	(0.012)	(0.015)
RMSE	0.353	0.353	0.383
	(0.004)	(0.005)	(0.008)
Number of Observations	5000	5000	5000

Notes: Dependent variable is ln(y_2). Main entry in each column is the sample mean of the statistic in 5000 simulations. Simulated standard errors are in parentheses. In each simulation, statistics based on synthetic data are computed in each of three synthetic implicates. Statistics are averaged over implicates before computing the mean and standard error over simulations.

APPENDIX TABLE 4

Simulated Re-identification Rates by Cell in Group *g* = 2

Value of x_1	Value of x_2				
	-2	-1	0	1	2
Synthetic Data, Density-Based Transform					
-2	0.046 (0.047) N = 22.3	0.013 (0.013) N = 80.8	0.008 (0.008) N = 128	0.013 (0.013) N = 80.5	0.047 (0.047) N = 22.3
-1	0.013 (0.013) N = 80.4	0.003 (0.003) N = 292	0.002 (0.002) N = 463	0.003 (0.003) N = 292	0.012 (0.012) N = 80.9
0	0.008 (0.008) N = 128	0.002 (0.002) N = 463	0.001 (0.001) N = 733	0.002 (0.002) N = 463	0.008 (0.008) N = 128
1	0.013 (0.013) N = 80.7	0.003 (0.003) N = 292	0.002 (0.002) N = 463	0.003 (0.003) N = 292	0.013 (0.013) N = 80.8
2	0.048 (0.047) N = 22.4	0.013 (0.013) N = 80.8	0.008 (0.008) N = 128	0.012 (0.012) N = 81.0	0.047 (0.046) N = 22.3
Synthetic Data, Exact Transform					
-2	0.047 (0.047)	0.013 (0.012)	0.008 (0.008)	0.013 (0.012)	0.047 (0.047)
-1	0.013 (0.013)	0.003 (0.003)	0.002 (0.002)	0.003 (0.003)	0.012 (0.012)
0	0.008 (0.008)	0.002 (0.002)	0.001 (0.001)	0.002 (0.002)	0.008 (0.008)
1	0.013 (0.012)	0.003 (0.003)	0.002 (0.002)	0.003 (0.003)	0.012 (0.013)
2	0.047 (0.046)	0.013 (0.013)	0.008 (0.008)	0.013 (0.012)	0.048 (0.048)

Notes: First entry in each cell is the average re-identification rate in that cell in 5000 simulations. The second entry, in parentheses, is the standard deviation of the re-identification rate in that cell in 5000 simulations. The third entry in the top panel is the average number of observations in that cell in 5000 simulations. Both panels are based on the same simulated data.

APPENDIX TABLE 5
Moments of the Distribution of Age on Jan. 1, 1990 by Sex and Race

	White		Black		Asian		Hispanic		American Indian		Other		Race Missing		TOTAL	
	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic
Men																
Mean	31.2	31.1	27.6	27.6	25.6	25.7	23.2	22.8	24.0	24.8	36.2	36.4	23.6	23.7	29.8	29.7
Std.Dev.	16.0	15.4	15.0	14.5	13.1	12.4	13.1	12.4	12.7	11.8	14.1	13.5	15.7	15.2	16.1	15.6
Skewness	0.40	0.34	0.56	0.47	0.78	0.65	0.65	0.56	0.61	0.44	0.20	0.04	0.04	-0.09	0.50	0.46
Kurtosis	-0.46	-0.40	-0.07	-0.17	0.71	0.57	0.54	0.32	0.53	0.27	-0.27	-0.30	-0.38	-0.58	-0.14	-0.09
N	458,356		81,000		20,157		47,547		1,203		9,319		60,239		677,821	
Women																
Mean	30.9	31.0	28.1	28.0	25.6	25.8	22.6	22.7	24.6	23.9	36.3	36.4	21.4	21.1	29.7	29.7
Std.Dev.	16.3	15.9	16.3	15.1	13.3	12.7	13.2	12.7	13.1	12.1	14.9	14.2	17.3	16.8	16.5	16.1
Skewness	0.45	0.39	0.56	0.48	0.75	0.61	0.68	0.57	0.62	0.42	0.13	-0.03	0.51	0.37	0.55	0.50
Kurtosis	-0.35	-0.31	-0.12	-0.13	0.65	0.43	0.48	0.24	0.60	0.29	-0.32	-0.35	0.07	-0.22	-0.06	-0.02
N	414,169		84,863		17,835		36,135		1,118		6,276		48,226		608,622	
TOTAL																
Mean	31.1	31.1	27.9	27.8	25.6	25.7	23.0	22.8	24.3	24.3	36.2	36.4	22.7	22.7	29.7	29.7
Std.Dev.	16.1	15.6	15.3	14.9	13.2	14.9	13.1	12.5	12.9	12.0	14.4	13.8	16.4	15.9	16.3	15.8
Skewness	0.42	0.36	0.56	0.48	0.76	0.63	0.66	0.56	0.62	0.42	0.17	0.01	0.25	0.11	0.53	0.48
Kurtosis	-0.40	-0.35	-0.09	-0.15	0.68	0.50	0.51	0.28	0.57	0.28	-0.28	-0.31	-0.20	-0.45	-0.10	-0.05
N	872,525		165,864		37,992		83,682		2,321		15,595		108,465		1,286,444	

Source: Authors' calculations based on the LEHD database.

Notes: Statistics in the columns labeled “Synthetic Value” are averaged over three synthetic data implicates.

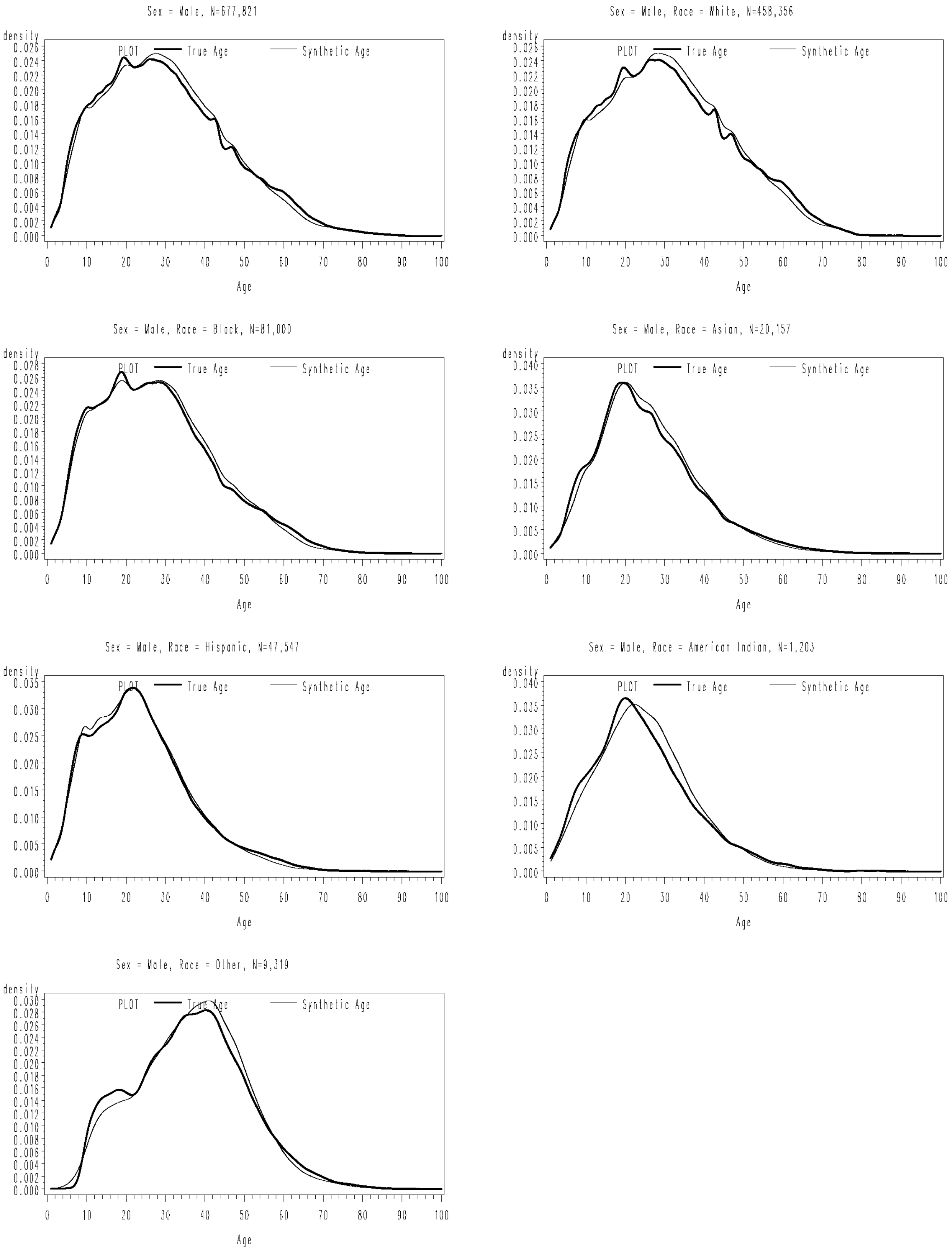
APPENDIX TABLE 6
Moments of the Distribution of Quarterly Employment Earnings by Sex and Race

	White		Black		Asian		Hispanic		American Indian		Other		Race Missing		TOTAL	
	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic	True	Synthetic
Men																
Mean	9,340	9,306	5,153	5,114	7,383	7,307	4,595	4,590	4,578	4,611	10,877	10,726	10,680	10,566	8,482	8,444
Std.Dev.	19,620	18,751	8,454	7,368	12,574	11,318	5,028	4,595	8,594	6,112	24,582	22,428	21,900	19,719	18,060	17,150
Skewness	24.4	24.0	48.6	37.3	23.1	18.1	55.7	40.6	41.9	7.74	18.8	17.6	18.1	15.3	25.9	25.2
Kurtosis	965	957	4,711	3,627	1,208	885	9,422	7,420	3,233	246	548	523	587	464	1,107	1,092
N	11,562,073		1,700,810		351,162		1,046,917		17,681		242,345		696,265		15,617,253	
Women																
Mean	4,990	4,995	4,397	4,399	5,398	5,393	3,625	3,660	3,830	3,867	6,654	6,662	5,109	5,121	4,829	4,836
Std.Dev.	7,498	6,953	7,498	4,492	6,893	6,423	3,301	3,310	3,995	3,945	8,361	7,898	8,899	7,573	7,002	6,499
Skewness	39.9	32.7	13.9	7.79	18.5	13.6	4.23	3.32	2.57	2.09	12.6	9.61	32.1	14.7	39.0	31.0
Kurtosis	3,969	3,177	1,284	531	1,097	762	79.7	44.5	18.0	8.87	453	267	2,566	641	4,085	3,145
N	10,543,615		2,036,955		355,368		760,836		20,366		164,341		492,806		14,374,287	
TOTAL																
Mean	7,265	7,250	4,741	4,724	6,385	6,344	4,187	4,199	4,178	4,212	9,169	9,082	8,568	8,501	6,731	6,715
Std.Dev.	15,260	14,547	6,674	5,985	10,172	5,985	4,411	4,130	6,558	5,082	19,814	18,132	18,306	16,435	14,024	13,291
Skewness	29.7	28.9	47.6	33.5	24.7	19.0	48.7	33.2	44.3	7.03	22.0	20.4	20.8	17.2	31.4	30.3
Kurtosis	1,504	1,477	5,698	3,889	1,526	1,089	9,240	6,596	4,444	261	788	741	810	617	1,722	1,678
N	22,105,688		3,737,765		706,530		1,807,753		38,047		406,686		1,189,071		29,991,540	

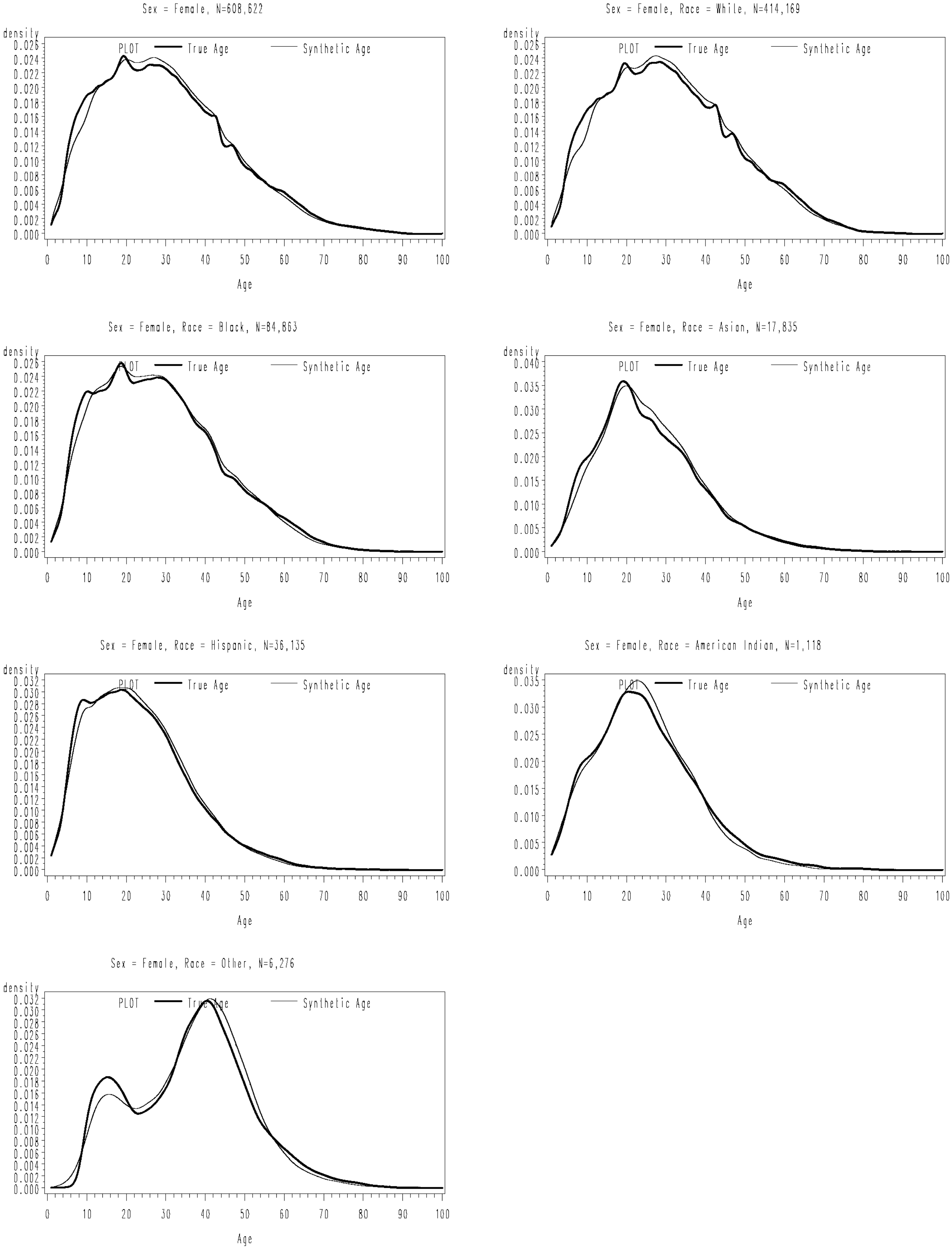
Source: Authors' calculations based on the LEHD database.

Notes: Statistics in the columns labeled “Synthetic Value” are averaged over three synthetic data implicates. The distribution of quarterly employment earnings is truncated at one and one million dollars.

APPENDIX FIGURE 1
Estimated Density of True and Synthetic Age on Jan 1 1990, Men



APPENDIX FIGURE 2
Estimated Density of True and Synthetic Age on Jan 1 1990, Women



Sex = Female, N=608,622

Sex = Female, Race = While, N=414,169

Sex = Female, Race = Black, N=84,863

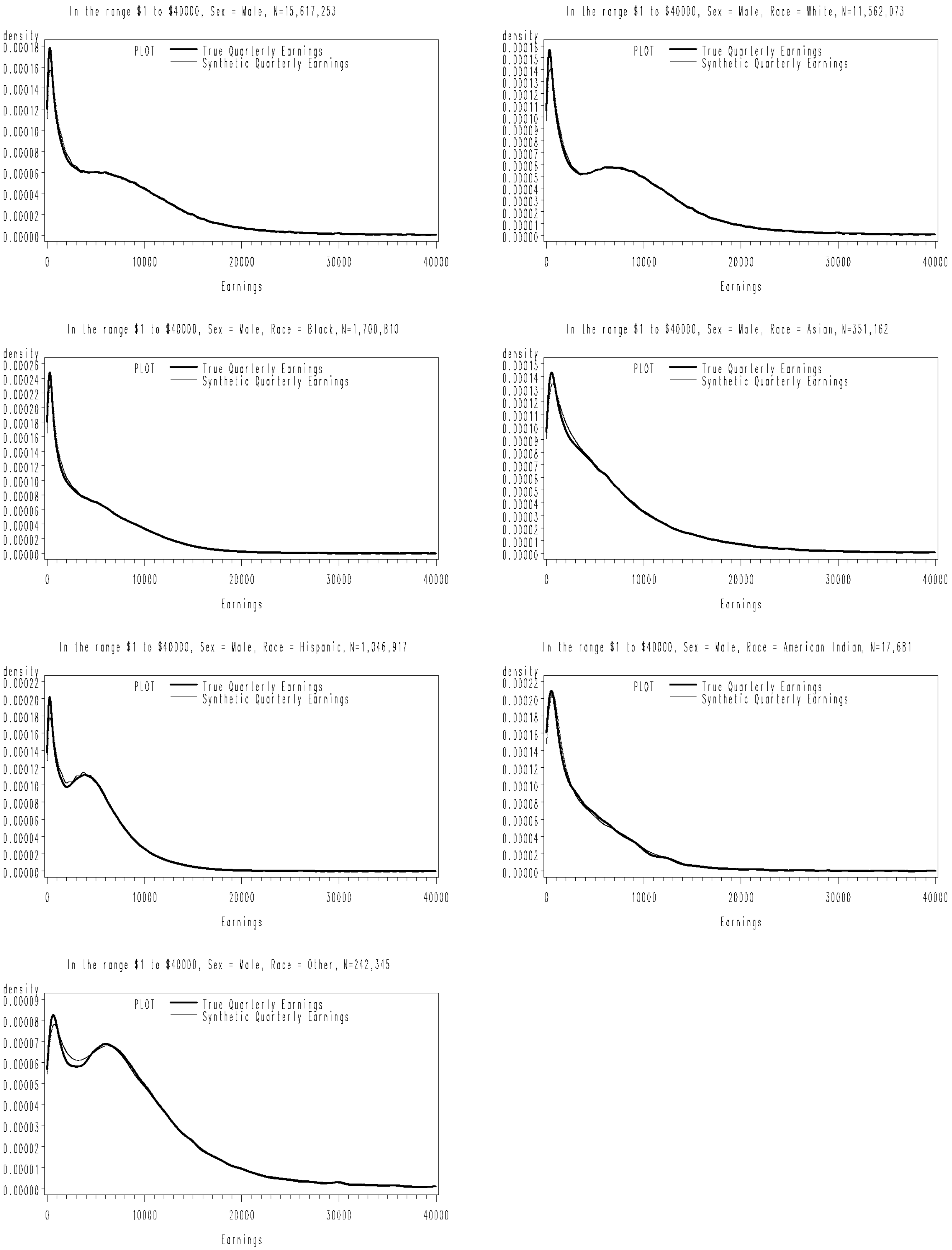
Sex = Female, Race = Asian, N=17,835

Sex = Female, Race = Hispanic, N=36,135

Sex = Female, Race = American Indian, N=1,118

Sex = Female, Race = Other, N=6,276

APPENDIX FIGURE 3
Estimated Density of True and Synthetic Quarterly Earnings, Men



In the range $1 to $40000, Sex = Male, N=15,617,253

In the range $1 to $40000, Sex = Male, Race = White, N=11,562,073

In the range $1 to $40000, Sex = Male, Race = Black, N=1,700,810

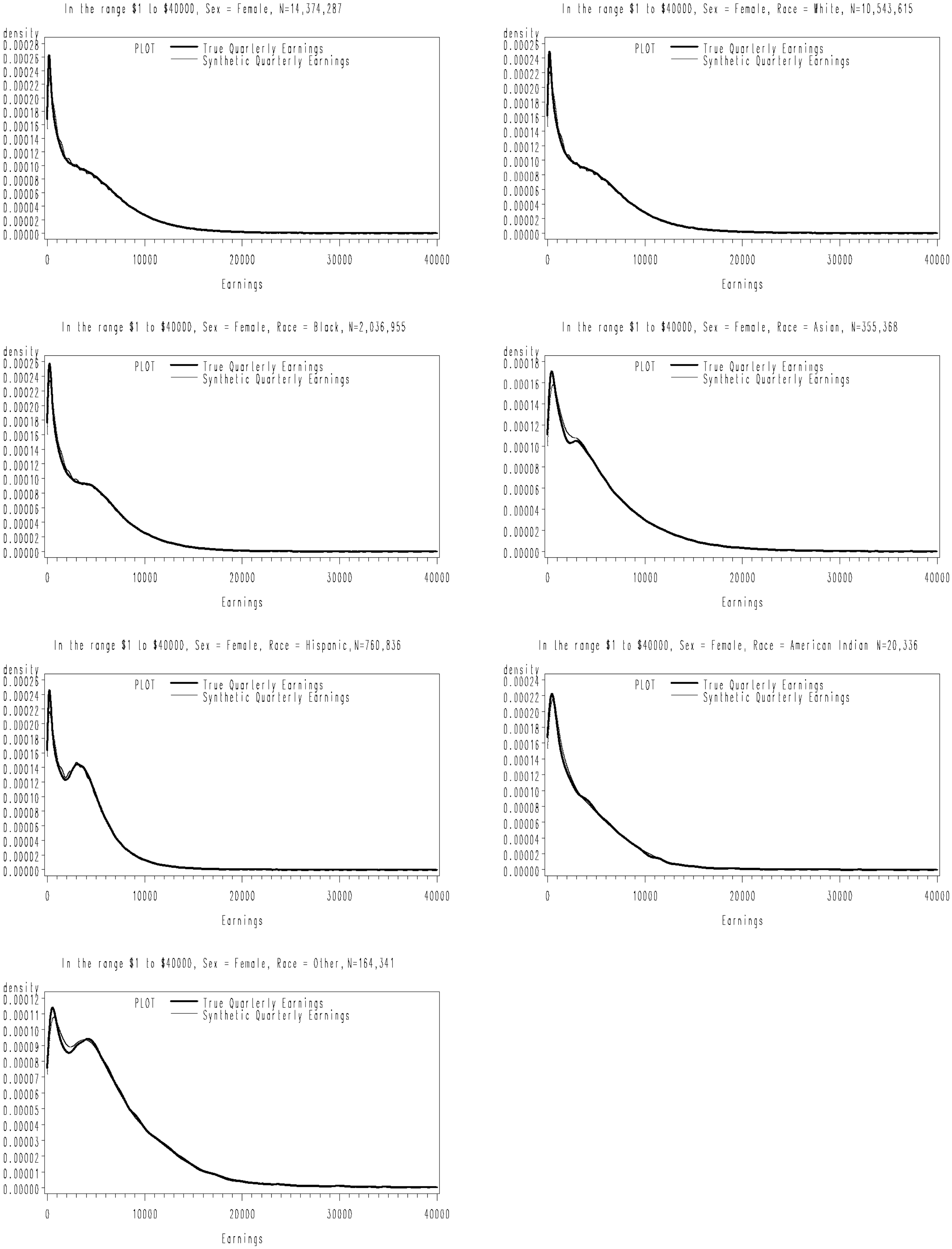
In the range $1 to $40000, Sex = Male, Race = Asian, N=351,162

In the range $1 to $40000, Sex = Male, Race = Hispanic, N=1,046,917

In the range $1 to $40000, Sex = Male, Race = American Indian, N=17,681

In the range $1 to $40000, Sex = Male, Race = Other, N=242,345

APPENDIX FIGURE 4
Estimated Density of True and Synthetic Quarterly Earnings, Women


In the range $1 to $40000, Sex = Female, N=14,374,287
density
PLOT
True Quarterly Earnings
Synthetic Quarterly Earnings
Earnings
In the range $1 to $40000, Sex = Female, Race = White, N=10,543,615
density
PLOT
True Quarterly Earnings
Synthetic Quarterly Earnings
Earnings
In the range $1 to $40000, Sex = Female, Race = Black, N=2,036,955
density
PLOT
True Quarterly Earnings
Synthetic Quarterly Earnings
Earnings
In the range $1 to $40000, Sex = Female, Race = Asian, N=355,368
density
PLOT
True Quarterly Earnings
Synthetic Quarterly Earnings
Earnings
In the range $1 to $40000, Sex = Female, Race = Hispanic,N=760,836
density
PLOT
True Quarterly Earnings
Synthetic Quarterly Earnings
Earnings
In the range $1 to $40000, Sex = Female, Race = American Indian N=20,336
density
PLOT
True Quarterly Earnings
Synthetic Quarterly Earnings
Earnings
In the range $1 to $40000, Sex = Female, Race = Other,N=164,341
density
PLOT
True Quarterly Earnings
Synthetic Quarterly Earnings
Earnings