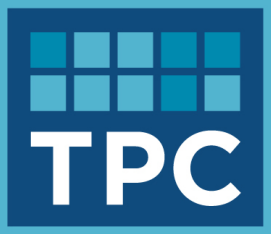


TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

A Synthetic Supplemental Public-Use File Of Low-Income Information Return Data: Methodology, Utility, And Privacy Implications

Claire Bowen, Victoria L. Bryant, Len Burman, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Philip Stallworth, Kyle Ueyama, Aaron R. Williams, and Noah Zwiefel

July 9, 2020; Revised November 11, 2020

ABSTRACT

The Statistics of Income division of the Internal Revenue Service releases an annual public-use file of individual income tax returns that is invaluable to tax analysts in government agencies, nonprofit research organizations, and the private sector. However, the Statistics of Income division has had to take increasingly aggressive measures to protect the data against growing disclosure risks, such as a data intruder matching the anonymized public data with other public information available in nontax databases. This project develops an alternative privacy protection method: a fully synthetic representation of the income tax data that is statistically representative of the original data. The method generates the synthetic data from a smoothed version of the empirical distribution of income tax returns. The resulting synthetic file includes no actual tax return records. In this report, we describe the methods used in the first part of this project, the creation of a synthetic public-use file of nonfilers. We show how the methodology protects the underlying data from disclosure and evaluates the quality of the data.

ABOUT THE TAX POLICY CENTER

The Urban-Brookings Tax Policy Center aims to provide independent analyses of current and longer-term tax issues and to communicate its analyses to the public and to policymakers in a timely and accessible manner. The Center combines top national experts in tax, expenditure, budget policy, and microsimulation modeling to concentrate on areas of tax policy that are critical to future debate.

CONTENTS

CONTENTS

ACKNOWLEDGMENTS

This paper relies on the analytical capability that was made possible in part by a grant from Arnold Ventures. We are grateful to the foundation and to all our funders, who make it possible for the Urban-Brookings Tax Policy Center to advance its mission.

The views expressed are those of the authors and should not be attributed the Urban-Brookings Tax Policy Center, the Urban Institute, the Brookings Institution, their trustees, or their funders. Funders do not determine research findings or the insights and recommendations of our experts. Further information on Urban's funding principles is available at http://www.urban.org/aboutus/our-funding/funding-principles; further information on Brookings' donor guidelines is available at http://www.brookings.edu/support-brookings/donor-guidelines.

We are grateful to Joe Ansaldi, Don Boyd, Vickie Bryant, Jim Cilke, John Czajka, Rick Evans, Dan Feenberg, Barry Johnson, Julia Lane, Graham MacDonald, Shannon Mok, Jim Nunns, James Pearce, Kevin Pierce, Alan Plumley, Daniel Silva-Inclan, Michael Strudler, Lars Vilhuber, Mike Weber, and Doug Wissoker for helpful comments and discussions.

INTRODUCTION

Tax data are a potentially invaluable resource for public policy analysis on many issues. The Internal Revenue Service (IRS) has for decades released a public-use file (PUF) with selected information from individual income tax return records that has been anonymized and altered to protect against the risk of disclosure. Analysts in academia, nonprofit research organizations, and the private sector use the PUF to study the effects of tax policy changes on revenues, the distribution of tax burdens, and economic incentives. But protecting taxpayers' privacy in the information age has required the IRS to limit the data released and distort those data in increasingly aggressive ways. Consequently, the released data are becoming less and less useful for analysis, and the PUF as currently conceived might soon no longer be produced.

This report presents an alternative approach to protecting tax data from disclosure: synthetic data generation, where values in a dataset are replaced with imputed values based on an underlying model. This methodology has been used to protect many administrative and other sensitive data sets against disclosure, but it has not been applied to US tax data. We develop a method for replacing administrative income tax data with fully synthetic records—that is, all of the values replaced with imputed amounts—discuss the method's implications for privacy, and apply the method to develop the 2012 Supplemental PUF, a database of individuals who were not dependents and did not file an individual income tax return in 2012. More information about the file is available in a report by the IRS (2019).[1] We demonstrate that our proposed method for generating fully synthetic data protects privacy because it would be impossible for a data intruder—even one possessing extensive information about most records in the administrative dataset—to determine with absolute certainty whether any individual is in the underlying administrative data used to create the synthetic file. This means that the synthetic dataset does not disclose whether someone had or had not filed a tax return. Because the synthetic data are imputed, the methodology also protects against disclosure of the confidential information of any individual in the underlying administrative data. Further, no PUF had been created from information returns before our work.[2] These data will allow researchers to gain a fuller picture of the distribution of income and tax burdens than one derived from income tax filing data alone.

Finding safe ways to provide access to administrative tax data is important because the data are valuable for public policy research. For instance, analysts use individual income tax return data to evaluate the distribution of tax burdens across income groups under current law; assess proposed changes to tax law; and model incentives to work, save, or invest under different policy scenarios. The data are also useful to analyze many nontax research questions because, in addition to detailed information about income, deductions, taxes, and credits, the datasets include key demographic information, such as taxpayers' marital status, their number of children, and the age and gender of both taxpayers and their dependents (available from the Social Security Administration).

Administrative tax data are comprehensive and high quality. Nearly all individuals in the United States are represented on an income tax return as a taxpayer or a dependent, and those who are not represented on an income tax return are generally represented on one or more information returns. And most taxpayers file quite complete and accurate tax returns either out of a sense of civic obligation or because filing a false return (or failing to file a return) can lead to severe

penalties. However, access to those data is limited to a few government agencies: the IRS, the Congressional Joint Committee on Taxation (JCT), Treasury's Office of Tax Analysis (OTA), and a few other agencies for specific purposes.

The microsimulation models of other organizations, such as the American Enterprise Institute, the Urban-Brookings Tax Policy Center, and the National Bureau of Economic Research, must rely on the PUF. Privacy protections are required by section 6103 of the Internal Revenue Code, which strictly limits access to tax return information and research based on tax return information. For example, statistical research to estimate taxpayer's behavioral responses to income tax parameters, such as the response of capital gains realizations to changes in tax rates, can only be performed directly by researchers in the Joint Committee on Taxation and the Office of Tax Analysis or in collaboration with them, or through a highly restrictive arrangement with the IRS.

This report is organized as follows. Section 2 reviews and defines terms and discusses approaches to protecting privacy and their implications for data utility. Section 3 details our proposed synthetic data generation method. Section 4 outlines how our synthesis process protects privacy, including protections against disclosure of outliers and attribute disclosure. Section 5 describes the implementation of our proposed synthetic data generation method on the Supplemental PUF data. Section 6 applies some metrics of privacy protection to the synthesized data. Section 7 evaluates the data utility measures and applies them to the synthesized Supplemental PUF data. We provide further discussion of our results and plans for future work in Section 8.

PRIVACY AND CONFIDENTIALITY

Any public tax data must protect the confidentiality of individual taxpayers.[3] Ensuring a dataset's confidentiality is challenging and complicated because more and more data that might appear on tax returns exists in other public and private databases, and the computational power required to match data from different sources continues to grow. The following section provides our definitions of some key concepts as well as a discussion of existing and proposed privacy protection methods and standards.

DEFINITIONS

Privacy may be defined as the ability "to determine what information about ourselves we will share with others" (Fellegi 1972). *Confidentiality* is "the agreement, explicit or implicit, between data subject and data collector regarding the extent to which access by others to personal information is allowed" (Fienberg and Jin 2009).

Disclosure is the act of making confidential information known. There are several types of disclosures. *Identity disclosure* occurs when a data intruder associates an individual with a specific record in the released data[4] and that discloses all the variables in the dataset with respect to that individual. This type of disclosure may be quite damaging. For example, an insurance company might increase insurance premiums for a participant based on information about health status gleaned from a medical survey, a credit card company could increase interest rates for an individual based on data obtained from a wealth survey, or a divorce lawyer might demand a larger settlement based on income data gathered from an income tax return.

Attribute disclosure occurs when a data intruder can determine certain characteristics of an individual based on information in the released data (Templ et al. 2019). This type of disclosure does not necessarily require identifying an individual in the data. For example, if all individuals in a census block are of one race and ethnicity, then a data intruder can know the race and ethnicity of someone who lives in the block without identifying the individual in the data. Although this may appear less harmful, attribute disclosure can lead to some of the same damage as identity disclosure.

Even without an identity or attribute disclosure, participants in a study may bear unintended costs. Wood et al. (2018) provide an example of a woman who decides to participate in a medical study and discovers she has a 50 percent chance of dying from a stroke in the next year. If a prospective insurer extracted her data from the survey, she might be denied life insurance coverage, or her premiums could skyrocket. But even if her identity isn't disclosed, including her data in the survey sample might increase the estimates of stroke risk for people like her. As a result, her life insurance premiums could increase even if her identity and individual data remain confidential.

However, improving the measurement of relationships among variables is not considered a disclosure. Otherwise, no statistical research using individual or household-level data would be permissible.

Utility, another important aspect of data privacy and confidentiality, is the usefulness of the data for analysis and research. *General utility* is the similarity of statistical properties, such as univariate and multivariate distributions, of the confidential data with the synthetic data (Snoke et al. 2018). *Specific utility* is the similarity of analytic results, such as regression estimates or summary tables, from the confidential data with the synthetic data (Snoke et al. 2018). In Section 7, we present specific measures of data quality and apply them to the synthetic nonfiler data.

LIMITATIONS OF TRADITIONAL METHODS FOR STATISTICAL DISCLOSURE

Data stewards (also referred to as data curators or data maintainers) have relied on many statistical disclosure control (SDC) or statistical disclosure limitation techniques to preserve data confidentiality while maintaining quality. However, some SDC techniques may fail to eliminate disclosure risk from data intruders armed with external data sources and high-powered computers (Dreschler and Reiter 2010; Winkler 2007). These techniques may also greatly reduce the usefulness of the released data for analysis and research.

Here, we list some key SDC methods that have been applied to tax data and other sensitive information and their limitations.

Adding random noise to continuous variables can maintain univariate distributions and prevent exact matches with external data sources. But the random noise added to sensitive variables creates measurement error in the perturbed variables, reducing the precision of statistical analyses and potentially introducing bias (Yancey, Winkler, and Creecy, 2002).

Data swapping is the exchange of sensitive values among sample units with similar characteristics other than the sensitive value. Mitra and Reiter (2006) found that a 5 percent random swapping of two identifying variables in the 1987 Survey of Youth in Custody invalidated statistical hypothesis tests in regression models that included those variables. Drechsler and Reiter (2010) also discovered that even 1 percent swapping of a subsample from the March 2000 US Current Population Survey can undermine statistical inference.

Top and bottom coding methods limit all values above or below a threshold to the threshold value. For example, the IRS currently "top codes" the number of children variable in the individual income tax return PUF at three for married-filing-jointly and head-of-household returns, two for single returns, and one for married-filing-separately returns (Bryant 2017). Top coding does not affect order statistics for values below the top coding threshold and, similarly, bottom coding does not affect order statistics for values above the bottom coding threshold. But the top and bottom coding approaches eliminate information at the tails of the distributions, degrading analyses that depend on the entire distribution (Fuller 1993; Reiter, Wang, and Zhang, 2014).

Aggregation combines several observations into one observation. The 2012 PUF aggregated 1,155 returns with extreme values into four observations (Bryant 2017). Aggregation does not alter simple statistics such as sums or means, but it may bias estimates from more complex statistical models and distort microsimulation model analyses that are

sensitive to outliers. Furthermore, aggregation of geographies may make small area estimation impossible and hides spatial variation (Reiter, Wang, and Zhang, 2014).

FULLY SYNTHETIC DATA AND DISCLOSURE RISKS

Replacing actual data with fully synthetic data has the potential to avoid the pitfalls of previous SDC techniques. The approach achieves this by attempting to simulate the data generation process of the confidential data based on a model of the underlying distribution. This method protects against identity disclosure because no real observations are released (Hu, Reiter, and Wang 2014; Kinney et al. 2011; Raab, Nowok, and Dibben 2017; Rubin 1993; Reiter 2002). Specifically, Hu, Reiter, and Wang (2014, 186), stated that "it is pointless to match fully synthetic records to records in other databases since each fully synthetic record does not correspond to any particular individual." Similarly, fully synthetic data prevent attribute disclosure because no actual values are released (Reiter 2002). Moreover, synthesized values limit a data intruder's confidence in any given value of a sensitive variable. For instance, if an intruder identifies a set of records with identical values for a sensitive variable (a simple attribute attack), they still cannot confirm that the value exists in the actual dataset.

If not carefully designed, fully synthetic data may still risk disclosing information (Raab, Nowok, and Dibben 2017). For example, overfitting the model used to generate the synthetic data might produce a synthetic file that is too close to the underlying data. In the extreme case, a data synthesizer could theoretically perfectly replicate the underlying confidential data (Elliot 2014).

The database reconstruction theorem (Dinur and Nissim 2003) proves that even noisy subset sums can be used to approximate individual records by solving a system of equations. If too many independent statistics are published based on confidential data, then the underlying confidential data can be reconstructed with little or no error.

The US Census Bureau produced its own application of the database reconstruction theorem using the 2010 Decennial Census. Based on published tables, researchers at the Census Bureau recreated the unreleased swapped and unswapped microdata with about 50 percent accuracy. They then correctly matched a small fraction of the records in the recreated microdata to credit bureau data (Ruggles 2018). These results are troubling, but a data intruder could not confirm that a match was correct—or even whether the reconstructed data were correct before the match—without access to the actual data.

Under certain conditions, many of the same techniques used to reconstruct nonsynthetic data might be used to reconstruct administrative data from fully synthetic data. For instance, Hu, Reiter, and Wang (2014) identified nontrivial disclosure risks in fully synthetic data processes.

To date, disclosure risks have only been identified for discrete variables and counts. Disclosure may be possible for categorical variables that have a limited number of possible values because they may be solved for with a finite set of simultaneous equations and with limited information. Hu, Reiter, and Wang (2014) determined reidentification risks on

synthetic data in the American Community Survey. The authors calculated posterior probability distributions for categorical variables based on the method used to synthesize the data.

Disclosure risks are difficult to estimate on complex synthetic datasets such as a synthetic individual income tax return database. Raab, Nowok, and Dibben (2017, 82) concluded that measuring disclosure risk in the synthesized data from the UK Longitudinal Series was impractical: "Hu et al. (2014); Reiter et al. (2014); McClure and Reiter (2012) proposed other methods that can be used to identify individual records with high disclosure potential, but these methods cannot at present provide measures that can be used with (the) sort of complex data that we are synthesizing."

DIFFERENTIAL PRIVACY

Motivated by these data privacy concerns and lack of quantifiable privacy loss, the computer science community originally developed a formal privacy guarantee called ϵ-differential privacy (ϵ-DP) for query-based methods. Later research applied ϵ-DP to data synthesis. ϵ-DP creates a formal disclosure guarantee with a provable and quantifiable "privacy-loss budget," ϵ, for a given statistic (or query) such as a count or sum (Dwork 2008). Unlike other SDC methods, ϵ-DP does not make any assumptions about the prior knowledge of data intruders or how they would attempt to draw inferences about a dataset. ϵ-DP algorithms are fully transparent: data stewards may safely release methodology and the privacy budget, ϵ, without affecting the risk of disclosure (Abowd and Vilhuber 2008). Note that ϵ-DP is a mathematical condition that a mechanism or algorithm must satisfy to be considered private and not a statement about the characteristics of the confidential data. In layman's terms, a differentially private method associates the potential for privacy loss with how much a statistic changes given the absence or presence of any globally possible individual record in the target data.

More formally, ϵ-DP requires establishing that, for a chosen $\epsilon > 0$, the log of the ratio of the probability of any vector of statistics arising from the ϵ-DP algorithm with any single observation included to the probability with it excluded is less than ϵ. In other words, given a small value of ϵ, a data intruder possessing all the records in the underlying data set but one could not infer that the target record was in the data set used to generate the statistics (or synthetic data set).

Machanavajjhala and colleagues (2008) describe the intuition as follows:

> Differential privacy is a privacy definition that can be motivated in several ways. If an adversary knows complete information about all individuals in the data except one, the output of the anonymization algorithm [the synthetic data set] should not give the adversary too much additional information about the remaining individual. Alternatively, if one individual is considering lying about their data to a data collector (such as the US Census Bureau), the result of the anonymization algorithm will not be very different if the individual lied or not. (p. 277).

This definition assumes that the data intruder has detailed information about all but one individual in the dataset, which would prohibit release of even very aggregate statistics, such as unaltered population means. A synthesis process that precisely reflected the distribution of the underlying tax data would also violate this standard because a data intruder

could replicate the synthesis process with all but one record of data and infer information about the missing record based on the difference between the two distributions. Perturbing the distribution by adding a small amount of noise or reducing the size of the synthetic data set could protect data from most of the sample, but that might not be effective for outlier observations.

Given the high privacy guarantee, ϵ-DP is often criticized for adding too much noise to the data and subsequently producing low-quality data. Several relaxations of ϵ-DP exist, such as (ϵ-δ)-probabilistic differential privacy, which guarantees that ϵ-DP is met with probability 1- δ (Machanavajjhala et al. 2008). In other words, the probability that a data intruder with (1) full information about the data protection process, (2) knowledge of ϵ, and (3) knowledge of all but one row of the confidential data could gain significant information about any individual's data is at most δ.

Several researchers have attempted to develop fully synthetic data sets that satisfy differential privacy (DP), but either the method did not actually satisfy DP or the data were not high quality. The Census Bureau's "OnTheMap" application was designed to achieve (ϵ-δ)- probabilistic differential privacy, but with limited data quality (Machanavajjhala et al. 2008). Elliot (2014) created a measure of "empirical differential privacy." The measure makes assumptions about the data intruder's knowledge and methods, so it does not satisfy DP. Kinney and colleagues (2011) calculated *ex post* measures of privacy for individual variables in subgroups for the Synthetic Longitudinal Business Database, and they confirmed the database does not guarantee DP. McClure and Reiter (2012) found that the privacy-loss budget, ϵ, may not be closely related to the probability of disclosure given the specific differentially private mechanism. Bowen and Liu (2020) compared and evaluated several differentially private algorithms that generated synthetic data and showed that many produced low-quality data based on statistical measures such as bias and root-mean-square error.

In a critique of the Census Bureau's use of DP, Ruggles (2018) concluded "differential privacy requires protections that go well beyond [the Census Bureau's] standard; under the new approach, responses of individuals cannot be divulged even if the identity of those individuals is unknown and cannot be determined. In its pure form, differential privacy techniques could make the release of scientifically useful microdata impossible and severely limit the utility of tabular small-area data."

PROPOSED DATA SYNTHESIS METHODOLOGY

We propose a data synthesis methodology that protects against meaningful disclosure. We define a meaningful disclosure as information that would allow an intruder to (1) infer whether any individual is in or out of the underlying data set (i.e., whether the individual has filed a tax return) or (2) update that intruder's estimate of the range of any variable compared with the estimate ascertainable without access to the synthetic data file. We demonstrate that the synthetic data produced by the method protects taxpayer information from disclosure or statistically meaningful inference about taxpayer attributes. The synthetic data also do not disclose any useful information about any individuals, even if an intruder has extensive information about the underlying data.

BACKGROUND

Four aspects of the data and our proposed methodology protect against disclosure:

- The administrative databases are very large, so a substantial amount of information may be released without allowing a data intruder to infer anything useful about individuals unless they already possessed almost all the original data. Consider an attack on a dataset as an attempt to solve a system of simultaneous equations to infer a missing variable (for a particular record). More independent observations in the underlying dataset mean that the data intruder needs an enormous amount of data to solve the system. Moreover, the dimensionality of the solution problem becomes quite large and computationally demanding. This feature alone does not meet the standard of DP, where an intruder is assumed to possess all the records except one, but it is a useful protection in more realistic scenarios where the intruder has incomplete information.
- The synthetic dataset will be only a fraction (no more than one-tenth) of the size of the underlying administrative data. We show later in this section that this protects against meaningful disclosure about the idiosyncrasies of the underlying empirical distribution. Essentially, sampling reduces disclosure risk because there is no guarantee that a targeted individual is in the sample before it is synthesized (Duncan and Lambert 1989; Fienberg, Makov, and Sanil 1997; Matthews and Harel 2011; Reiter 2005b). Thus, an advantage of working with federal administrative tax data is that sampling rates can be quite low while still producing a large, representative dataset. This means that the vast majority of records in the underlying administrative database are not in the sample. However, this feature alone does not meet the standard of DP, where an intruder is assumed to possess all the records except one, but it is still a useful protection in more realistic scenarios where the data intruder has incomplete information.
- Previous research has focused on the special challenges created by outliers. Data intruders often have more information about outliers and may have more to gain from identifying them. We propose a method that smooths the distribution of underlying data, preserving the empirical distribution for nonsensitive observations. Specifically, the empirical distribution has a high population density and then flattens in the tails

to only reflect the general characteristics of the outlier observations. This protects against inference of even very sensitive observations.[5]

DATA SYNTHESIS METHODOLOGY

Overview of method

In general, synthetic-data generation techniques relying on a model can be roughly grouped into two categories: parametric (e.g., regression) and nonparametric models (e.g., classification and regression trees, or CART). CART uses predictor variables to sort observations of an outcome variable into relatively homogeneous groups and then draws from the empirical distribution of each group. We focus on implementing CART because the method is computationally simple and flexible. Further, CART outperformed regression-based parametric methods in preliminary tests.

CART is a collection of nonparametric models developed by Breiman and colleagues (1984) and brought to synthetic data generation by Reiter (2005a). In essence, CART creates a sequence of binary splits of the data that end in nodes that are intended to be homogenous and have predictive power. The method splits the data using classification trees for categorical variables and regression trees for continuous variables. According to Therneau and Atkinson (2019), a tree is built as follows:

1. Find the variable that best splits the data into two groups. Split the data.
2. For each subgroup, find the variable that best splits the data into two groups. Split the data.
3. Continue this process until the subgroups reach a user-specified minimum size or until no improvement can be made.
4. Optionally, use cross-validation to reduce the full tree to avoid overfitting.

We estimate CART models for each variable with all previously synthesized outcome variables as predictors. More specifically, our synthetic-data generation method is based on the insight that a joint multivariate probability distribution can be represented as the product of a sequence of conditional probability distributions. That is,

$$f(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_k \mid \theta_1, \theta_2, \ldots, \theta_k) = f_1(X_1 \mid \theta_1) \cdot f_2(X_2 \mid X_1, \theta_2) \cdots f_{\mathrm{k}}(X_k \mid X_1, X_2, \ldots, X_{k-1}, \theta_k) \quad (1)$$

where X_i , i = 1 to k, are the variables to be synthesized, θ_i are vectors of model parameters such as regression coefficients, and k is the total number of variables.

To create a synthetic record, first *gender* is assigned randomly based on the percentage distribution of records in the first-level groups (females and males). Then *age* is assigned randomly, taking into account the already randomly assigned *gender* values and the distribution of *ages* for each *gender*.

After assigning *gender* and *age*, CART predicts *Social Security benefits* based on the assigned values for *gender* and *age*. For continuous variables, such as *Social Security benefits*, random draws are made from the smoothed empirical distribution of each of the groups created by CART. Subsequent variables are then synthesized as a function of the previously synthesized variables.

Synthesizing discrete variables (X_1 and X_2)

The first variable (X_1) synthesized in the Supplemental PUF data is *gender*, which is simply split based on the distribution of gender in the administrative data and randomly assigned based on this distribution. *Age* (X_2) is the only other discrete variable and is split into groups to minimize the heterogeneity of values within groups. To measure heterogeneity, the algorithm in our synthesis uses a Gini index

$$I(A) = \sum_{i=1}^{C} p_i(1 - p_i) \quad (2)$$

A is a node, *C* is the number of classes in the node (e.g., 2 for binary *gender*), and p_i is the class probability for the *i*th class (e.g., 0.51 are Female). The best split minimizes

$$\frac{N_L}{N} I(A_L) + \frac{N_R}{N} I(A_R) \quad (3)$$

where N_L and N_R are the number of observations in the left and right nodes created by the split respectively, $N = N_L + N_R$ is the total number of observations in both nodes, and $I(A_L)$ and $I(A_R)$ are the Gini index in the left and right nodes respectively. Splits continue until there is no reduction in the heterogeneity or until the minimum size (50) for a final node is reached.

Synthesizing continuous variables (X_3, X_4, …, X_k)

Continuous variables are split and synthesized using regression trees, based on the values of previously synthesized variables. At each partition, the "best split" is the one that minimizes the error sum of squares given that the data are partitioned into two nodes. Thus,

$$SSE = \sum_{i \in A_L} (y_i - \overline{y_L})^2 + \sum_{i \in A_R} (y_i - \overline{y_R})^2 \quad (4)$$

where A_L and A_R are the left and right nodes created by the split. The variables $\overline{y_L}$ and $\overline{y_R}$ are the means of the left and right nodes, respectively (Kuhn and Johnson 2016). Splits continue until there is no improvement in the splitting criteria or until the minimum size for a final node is reached (50). Our data synthesis approach samples values from the appropriate final node and then applies our smoothing method.

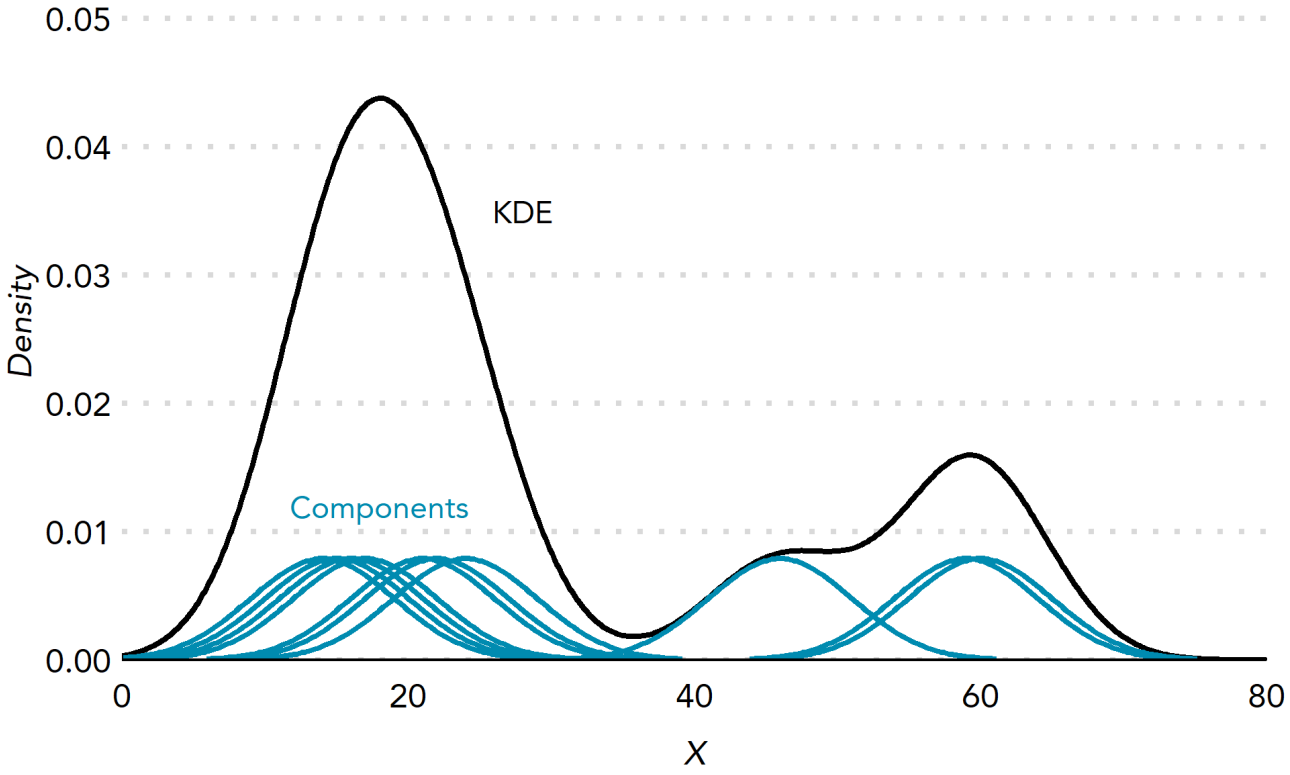
To synthesize the first continuous variable X_3, **(***Social Security benefits* in the Supplemental PUF data), we create a smoothed kernel density function for each percentile of values predicted by CART for this variable. As shown in figure 1, the kernel density estimator is the aggregation of individual normal densities centered around each observation.[6] In the example, each of the individual Gaussian kernels has the same variance. The kernel density distribution is smooth and unbounded.

However, we must tackle some complications with executing this approach. First, the variance of the Gaussian kernel must be larger when sampling outliers. If the variance was not larger, a data intruder who knows how the database is constructed might draw some fairly precise inferences because outlier observations in the synthetic dataset would likely be relatively close to an actual observation. We use percentile smoothing, which selects the variance based on the optimal variance for a kernel density estimator estimated on observations in the percentile for each observation. As we will discuss later, this causes the variance to grow with the value of the synthesized variable. Second, variables that are a deterministic

function of others, such as adjusted gross income or taxable income, will be calculated as a function of the synthesized variables. We do not calculate such variables for the Supplemental PUF data.

FIGURE 1

Kernel Density Estimate as Weighted Sum of Component Densities



Although our use of a smoothed version of the empirical distribution function means there is a zero probability of drawing an actual sample value, information about particular observations still could be revealed if the empirical distribution too closely matches the population distribution. Our data synthesis method mitigates this risk by using only a fraction of the observations in the administrative dataset.

THE EFFECTS OF SAMPLING ON INFERENCE ABOUT THE UNDERLYING DISTRIBUTION

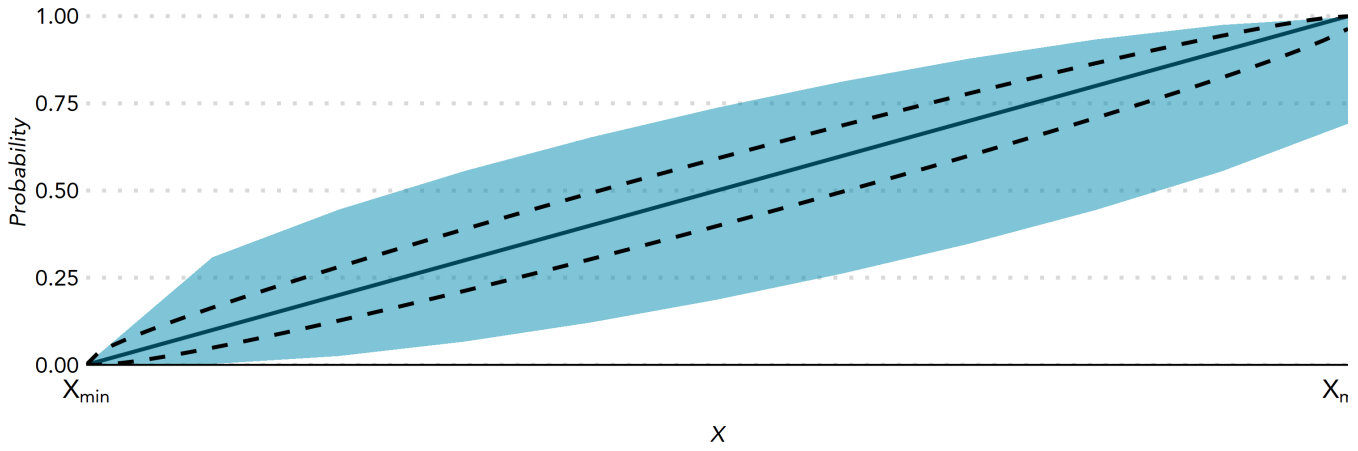
For the Supplemental PUF database, we start with a 10 in 9,999 (0.1 percent) sample. The odds are thus about 1,000 to 1 against any particular record from the population being in the sample. For the synthetic individual income tax return database, we plan to sample at different rates in different parts of the distribution. The synthetic file will be a stratified sample, with no portion of the dataset sampled at a rate higher than 1 in 10 (10 percent).

For the synthetic individual income tax return database, selecting a sample size that is at least an order of magnitude smaller than the underlying population obscures the nature of the underlying distribution. To illustrate, suppose the actual distribution of data in the administrative data set is uniform within an interval that includes 100 records. The actual distribution is the solid line in figure 2. A synthesis might draw *n* independent observations from the uniform distribution within this interval.[7] A data intruder might attempt to infer the underlying distribution by first ranking the observations from smallest to largest and then plotting the empirical distribution function. The data intruder would glean little information about the underlying distribution from this plot, especially if n is much smaller than 100, because many underlying distributions could be consistent with the sample distribution.[8]

FIGURE 2

95-Percent Confidence Interval Around Points Drawn from a Uniform Distribution Function

1-in-10 Draw (shaded area) Versus 1-in-1 draw (dashed lines) for n = 100



This simple example illustrates how drawing only a fraction of the observations in the underlying database will obscure many idiosyncrasies in the underlying empirical distribution.

OUTLIERS

Extreme values (outliers) are not close to uniformly distributed. Consider the most extreme case, where all but one of the observations are at the minimum value and one is at the maximum, x_m. How much could a data intruder infer about x_m? To simplify the algebra, assume that the minimum value is zero. Alternatively, think of x_m as the difference between minimum and maximum values. Suppose that there are 100 observations, 99 of which are zero. Then, the mean is

$$\mu = \frac{x_m}{100} \tag{5}$$

and the variance is

$$\sigma^2 = \sum_1^{100} \frac{(x_i - \mu)^2}{100} \tag{6}$$
$$= \frac{99(-0.01x_m)^2 + (0.99x_m)^2}{100} = \frac{\big(0.99(0.01 + 0.99)\big)x_m^2}{100}$$
$$= \frac{0.99x_m^2}{100}$$

Simply publishing the mean or variance for this subsample would disclose x_m if an intruder knew that the other values were all zero because x_m can be calculated as either 100 μ or $\sqrt{100\sigma^2/0.99}$.

Although this is a concern, our approach to simulating data by drawing from a kernel density estimator with variance σ^2 addresses it. Suppose we draw a 1-in-10 sample from the population of simulated values. The mean, $\bar{x}$, has the following properties:

$$E(\bar{x}) = \mu = \frac{x_m}{100} \quad (7)$$

$$Var(\bar{x}) = \frac{\sigma^2}{10} \quad (8)$$

Publishing the mean does not disclose much about the outlier, and there is a 90 percent probability that the outlier is not even in the database used to synthesize the data. The standard error of the mean will be quite large: the square root of $Var(\bar{x})$ from equation (4). Substituting from equation (2) yields the following:

$$se(\bar{x}) = \frac{\sigma}{\sqrt{10}} = \frac{\sqrt{0.99}x_m}{10\sqrt{10}} = 0.0315\ x_m \quad (9)$$

From equation (7), the best guess for x_m is $100\bar{x}$. The standard error of this estimate is $100se(\bar{x}) = 3.15x_m$. That is, the standard error of an estimate of x_m in this case is more than three times the actual value. So, any synthetic data sample that preserved the very high variance of the skewed sample would not reveal anything useful about the one outlier value.

If instead all the values are approximately the same, the simulated values will be very close to the outlier values, but there is no disclosure because those values are not unique.

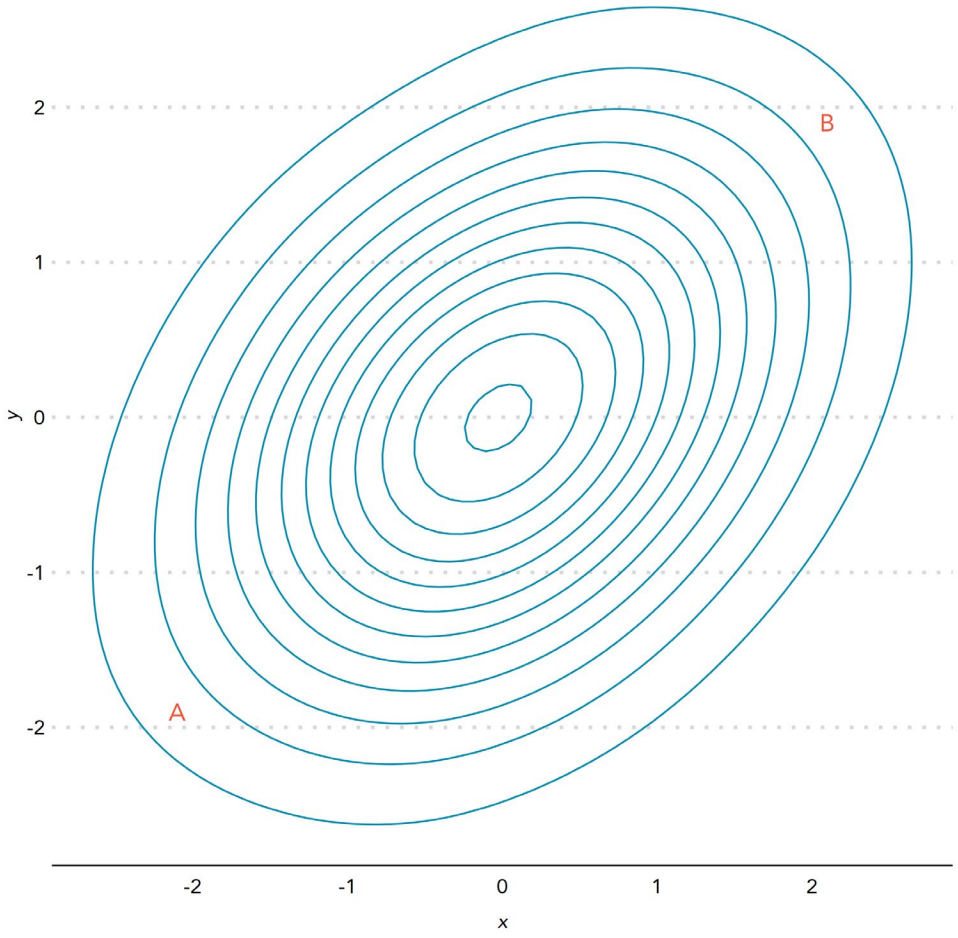
ATTRIBUTE DISCLOSURE

Tax return attribute disclosures raise two concerns. One is the revelation of information about particular taxpayers based on unique combinations of attributes. The other is the revelation that a person has filed a tax return, which the IRS treats as an impermissible disclosure. The data synthesis methodology described here protects against both types of disclosure. As shown, the data synthesis will prevent a data intruder from inferring any particular values of any individual's tax return, even if the intruder possesses extensive information about the taxpayer's other attributes. The probability of correctly inferring rare combinations of attributes will be even smaller. For example, in a bivariate normal distribution, rare pairings lie in the relatively flat part of the distribution along with other equally improbable combinations that do not occur in the original dataset. Thus, observing a point such as *A* on figure 3 (contour lines for a bivariate normal distribution with ρ = 0.5) tells us virtually nothing about whether a point like *A* exists in the original data. The true data point may be *B* or any of the infinite other disparate points in the tails of the distribution function.

More of a disclosure risk exists for discrete variables, such as *age*, but we address that by top coding.

FIGURE 3

Level Curves of Bivariate Normal Distribution with ρ = 0.5



The other type of disclosure is evidence of filing a tax return. The IRS has always viewed evidence that a return has been filed as disclosure of taxpayer information, which is prohibited under Internal Revenue Code Section 6103. The IRS revealing an individual had not filed an income tax return would reveal information about that person's income or evidence that the person might have violated the law. Both situations would be meaningful disclosures that could potentially harm that person. However, a data intruder could not infer from the synthetic income tax database that a person had not filed. Because the tax-filer database will be based on at most a 1-in-10 sample of actual tax returns, there is at least a 90 percent chance that any particular tax return filed will be excluded from the sample used to create the synthetic file. The highest risk of identification in a nonsynthetic sample would be for extreme outliers (i.e., those with extremely high incomes or with rare combinations of attributes). As noted, the data synthesis method effectively addresses this source of disclosure risk.

The synthetic Supplemental PUF database, described in the next section, is constructed from a 0.1 percent sample. Applying the same attribute disclosure reasoning, it would be virtually impossible to deduce from it that a particular individual was in the Supplemental PUF dataset.

SYNTHESIZING THE LOW-INCOME SUPPLEMENT SAMPLE DATA

Our main objective is to synthesize records from the IRS Master File to create a synthetic file that is similar to the current PUF released by the IRS Statistics of Income Division but that has stronger privacy protections. As a proof of concept and in an effort to release useful data that had never before been made public, we first create a fully synthetic file called the Supplemental PUF.

We begin with a definition of nonfilers from Cilke (2014): "Any US resident that does not appear on a Federal income tax return filed [for] a given year." We are not interested in people who are required to file but did not, so we exclude those with incomes above twice the filing threshold for married couples filing jointly. Our sample thus comprises people who do not file a federal income tax return for a given year and do not appear to have an income-tax filing requirement.[9]

Our data source is a random 0.1 percent sample of information returns for tax year 2012 maintained by the IRS Statistics of Income Division. Information returns are forms provided to the IRS by any business or other entity that pays income or has certain other transactions with an individual. Examples include the SSA-1099 filed by the Social Security Administration, W-2 filed by employers, and 1099-INT filed by banks and other financial institutions that pay interest. The sample comprises individuals whose Social Security number or individual taxpayer identification number (for those without Social Security numbers) ends in one of 10 four-digit combinations. The last four digits are randomly assigned at birth and range from 0001 to 9999. Thus, we used a random sample of 10 in 9,999 (approximately 1 in 1,000).[10]

TABLE 1

Percentage of Nonfilers with Specific Information Return Types

Type	Description	Percent of nonfilers with one or more information return	Percent of nonfilers with only one information return
SSA-1099	Social Security Benefits. Includes Form RRB-1099	55.9	34.6
W-2	Wage and Tax Statement	24.7	9.2
1099-INT	Interest Income	15.6	3.0
1099-R	Distributions from Pensions, Retirement Plans, etc.	14.3	0.7
1099-G	Certain Government Payments	11.1	3.4
1098	Mortgage Interest Statement	9.9	0.8
5498	Individual Retirement Arrangement Contributions	7.9	0.6
1099-MISC	Miscellaneous Income	7.8	2.8
1098-T	Tuition Statement	5.0	1.7
1099-DIV	Dividends and Distributions	4.3	0.8
1099-B	Proceeds From Broker and Barter Exchange Transactions	2.6	0.1
1098-E	Student Loan Interest Statement	2.1	0.4
W-2G	Certain Gambling Winnings	1.2	0.2
1099-C	Cancellation of Debt	1.1	0.3

Source: Table 2 in Cilke (2014).
Note: This table excludes specific Information Return types if held by less than 1 percent of nonfilers.

We delete records for those who should not be considered nonfilers by dropping late filers, deceased persons, foreign residents, and individuals with large dollar amounts for certain items. After dropping a few more observations because of missing or invalid ages or genders, the final administrative data set has about 26,000 observations.

We synthesize the data using a customized version of CART from the R package *synthpop* (Nowok et al. 2019). *Synthpop* contains several methods for creating partially synthetic and fully synthetic datasets and for evaluating the utility of synthetic data.

We use CART to partition the sample into relatively homogeneous groups with the constraint that none of the partitions be too small, to protect against overfitting (Benedetto, Stinson, and Abowd 2013). In testing on the Supplemental PUF database, we found that a minimum partition size of 50 produces a good fit with adequate diversity of values within each partition. Note that the optimal size may be different when synthesizing individual income tax return data.

To develop the synthetic Supplemental PUF dataset, we first split the data into two parts. One part includes the observations from the confidential data that have zeros for all 17 tax variables.[11] The other part includes the observations with at least one nonzero tax variable. For the part with zeros for all tax variables, we randomly assign the *gender* value based on the proportions in the zero subsample (see below), synthesize *age* based on *gender*, and assign zeros to all tax variables.

For the part with at least one nonzero value for a tax variable, we randomly assign *gender* (X_1) values based on the underlying proportions in the confidential dataset. With 51 percent female and 49 percent male in the administrative data set, the assigned *gender* value for each row in the synthetic data set will have a 51 percent probability of being female and a 49 percent probability of being male. Because *gender* is randomly assigned, the exact share of males and females in the synthetic data set may differ slightly from the distribution of *gender* in the administrative data, but the difference is likely to be small given the sample size.

We then use CART to assign *age* (X_2) to each record conditional on *gender*. Because the CART method selects values at random from the final nodes, the distribution may differ slightly from the distribution of *age* by *gender* in the administrative data, but again the differences are likely to be small given the sample size. Age is top coded at 85 after synthesis.[12]

For continuous variables, we start with the variable with the most nonzero values, *Social Security benefits* (X_3), and then order the remaining variables, ($X_4, X_5, \ldots, X_{19}$), in terms of their linear correlations with *Social Security benefits*, from most highly to least correlated.[13] CART partitions the data into relatively homogeneous *Social Security benefit* groups within each *gender*/*age* group and randomly selects a value for Social Security benefits from each of the values in that *Social Security* group. All nonzero values are replaced with random draws from a normal distribution in which the mean is the value being replaced and variance is the optimal variance from a kernel density estimator estimated on the corresponding percentile of the distribution of the variable being synthesized. This approach is a computationally efficient way to approximate a kernel density estimator and has the desirable feature that the variance is much larger for the sparse parts of the distribution than for the dense parts of the distribution. Further, variables in any given row in the synthetic dataset

come from many different rows in the confidential data. This means that even without smoothing empirical distributions, uncommon combinations of zeros and nonzeros within a synthesized record may be an artifact of the synthesizer and not an attribute of the underlying confidential data.

No smoothing is applied to values of 0, which is the most common value for all continuous variables in the Supplemental PUF data. We do not consider zeros to be a disclosure risk because the variable with the most nonzero values still contains about 40 percent zeros and all others contain 70 or more percent zeros. Many of the variables are zero for almost every record.

Subsequent variables (X_4, X_5, …, X_k) are synthesized in a similar way to X_3 by using CART to predict values based on random draws from the kernel density estimator of observations with similar characteristics. Classification trees and regression trees for prediction tend to overfit data, so most trees are reduced based on a penalty for the number of final nodes in the tree (Kuhn and Johnson 2016). For the Supplemental PUF, we do not reduce trees, because our minimum partition size is large (50).

MEASUREMENT OF PRIVACY PROTECTION IN THE SYNTHETIC SUPPLEMENTAL PUF

Our data synthesis method is designed to protect confidentiality *ex ante*. However, we also use privacy metrics to test whether the CART method might produce values that are too close to actual values or reveal too much about relationships between variables. We used these metrics to adjust the precision of the data synthesis by adjusting smoothing methods and parameters such as the minimum size of the final nodes in the CART synthesizer. We focus on three different types of metrics: (1) counts of the number of unique donors to each row in the synthetic data, (2) the frequency and uniqueness of synthesized rows in the confidential data, and (3) a formal privacy framework called ℓ-diversity to the CART synthesizer. With all of these measures, we should remember that the confidential data being synthesized come from a 10-in-9,999 sample of tax records. This means the low probability rows in the data are unlikely to be unique in the population. The results for duplicates, unique-uniques, and row-wise squared inverse frequency were all very small and thus are not reported in our results.

DUPLICATES

We examined several metrics for the frequency and uniqueness of synthesized rows in the confidential data. Even though all nonzero values are individually synthesized, a row in the synthetic data could match a row in the confidential data by chance.

The simplest metric of duplication is a count of rows in the unsmoothed synthetic data that match rows from the confidential data, but this is not particularly informative for two reasons. First, many rows have values for age, sex, and then all zeros for the tax variables. The probability of duplicating these rows is high but does not carry any disclosure risk. Second, there are many rows that occur in the confidential data that would be expected to appear as replicated in the confidential data by chance.

NUMBER OF UNIQUE-UNIQUES

The count of unique-uniques is the number of unique rows from the confidential data that are unique in the unsmoothed synthetic data. This narrows the focus to rows that are uncommon and could carry some inferential disclosure risk.

ROW-WISE SQUARED INVERSE FREQUENCY

Finally, we used a measure based on frequency. For any given row in the unsmoothed synthetic data, this metric counts the number of identical rows in the confidential data. We then take the inverse square of this metric such that rows that appear once are assigned a value of 1, rows that appear twice are assigned a value of $\frac{1}{4}$, rows that appear thrice are assigned a value of $\frac{1}{9}$, and so on.

ℓ-DIVERSITY OF FINAL NODES IN THE CART ALGORITHM

We were concerned that the CART algorithm could generate final nodes that lack adequate heterogeneity to protect confidentiality. Too little heterogeneity in the final nodes could lead to too much precision for the synthesizer. To ensure adequate heterogeneity, we applied ℓ-diversity (Machanavajjhala, Kifer, and Gehrke 2006) to the decision trees created by the CART algorithm.

ℓ-diversity is an extension of k-anonymity (Sweeney 2002). Let a quasi-identifier be a collection of nonsensitive variables in a dataset that could be linked to an external data source. Let a q*-block be a unique combination of the levels of quasi-identifiers. A q*-block is ℓ-diverse if it contains at least ℓ unique combinations of sensitive variables.

We apply this formal measure to the CART algorithm, where at each partition the split directions (left and right) are considered to be quasi-identifiers and the final nodes are considered to be q*-blocks. The trees create the discretized space formed by quasi-identifiers, the final nodes are q*-blocks, and the sensitive values are the values in the final nodes. We examine the minimum ℓ-diversity in a data synthesizer and the share of observations that came from final nodes with ℓ-diversity less than 3. In many cases, the minimum ℓ-diversity is 1 because some final nodes only contain zeros. We consider this to be acceptable because zeros carry negligible disclosure risk.

DESCRIPTION OF QUALITY MEASURES

General utility measures

Comparing *summary statistics* is a simple way to evaluate the quality of the synthesis. Ideally, for each variable, the distribution in the synthetic dataset is similar to the distribution in the underlying data. For discrete variables, the number of observations falling into each category should be similar in both the synthetic and underlying data. For continuous variables, the first four moments calculated on the synthetic data should be similar to the moments in the underlying data.

Correlation fit measures how well the synthesizer recreates the linear relationships between variables in the confidential dataset. The difference matrix is the lower triangle of a Pearson's linear correlation matrix from the synthetic data minus the same lower triangle from the confidential data. The difference matrix can be used to calculate two useful metrics. Values close to zero provide one measure of general utility in the synthetic data and are the result of similar correlation coefficients from the synthetic and confidential data sets.

First, we can rank the differences between each pair of variables from smallest to largest. Variable pairs with large differences indicate a poor job capturing the linear relationship (or lack thereof) between those two variables. Second, we can average the Euclidean distances between the pairs of variables in the confidential dataset and the synthetic dataset. This gives a general data synthesis-wide number that measures how well the data synthesis method is capturing linear relationships.

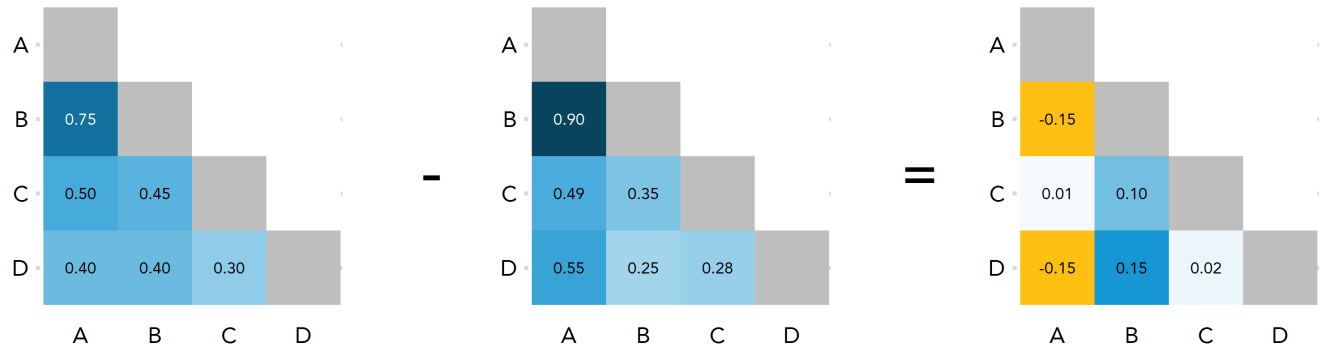
Let *S* and *O* be the correlation matrices corresponding to the synthetic and original data, respectively. The correlation fit is the average of distance between elements in the lower triangles of the two matrices.

$$correlation\ fit = \frac{\sqrt{\sum_{i=2}^{n}\sum_{j=1}^{i}(s_{ij}-o_{ij})^2}}{\binom{n}{2}} \quad (10)$$

The *Kolmogorov-Smirnov* (KS) *test* is a nonparametric test of the equivalence of univariate probability distributions. For synthetic data, the KS test statistic and its associated p value can be used to compare the distribution of an actual confidential variable and its synthesized counterpart. The null hypothesis is that the distributions are identical; a high p value indicates that the null hypothesis that the two distributions are identical cannot be rejected.

FIGURE 4

Example Calculation of Correlation Fit



The two-sample KS-test compares the empirical cumulative distribution functions for two samples. Let $I_{(-\infty,x_i]}(X_i)$ be an indicator function for the variable of interest. The empirical cumulative distribution function (ECDF) for the first sample, $F_{n,1}$, for *n* independent and identically distributed ordered observations is

$$F_{n,1} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x_i]}(X_i) \tag{11}$$

The ECDF for the first sample, $F_{m,2}$, for *m* independent and identically distributed ordered observations is

$$F_{m,2} = \frac{1}{m}\sum_{i=1}^{m} I_{(-\infty,x_i]}(X_i) \tag{12}$$

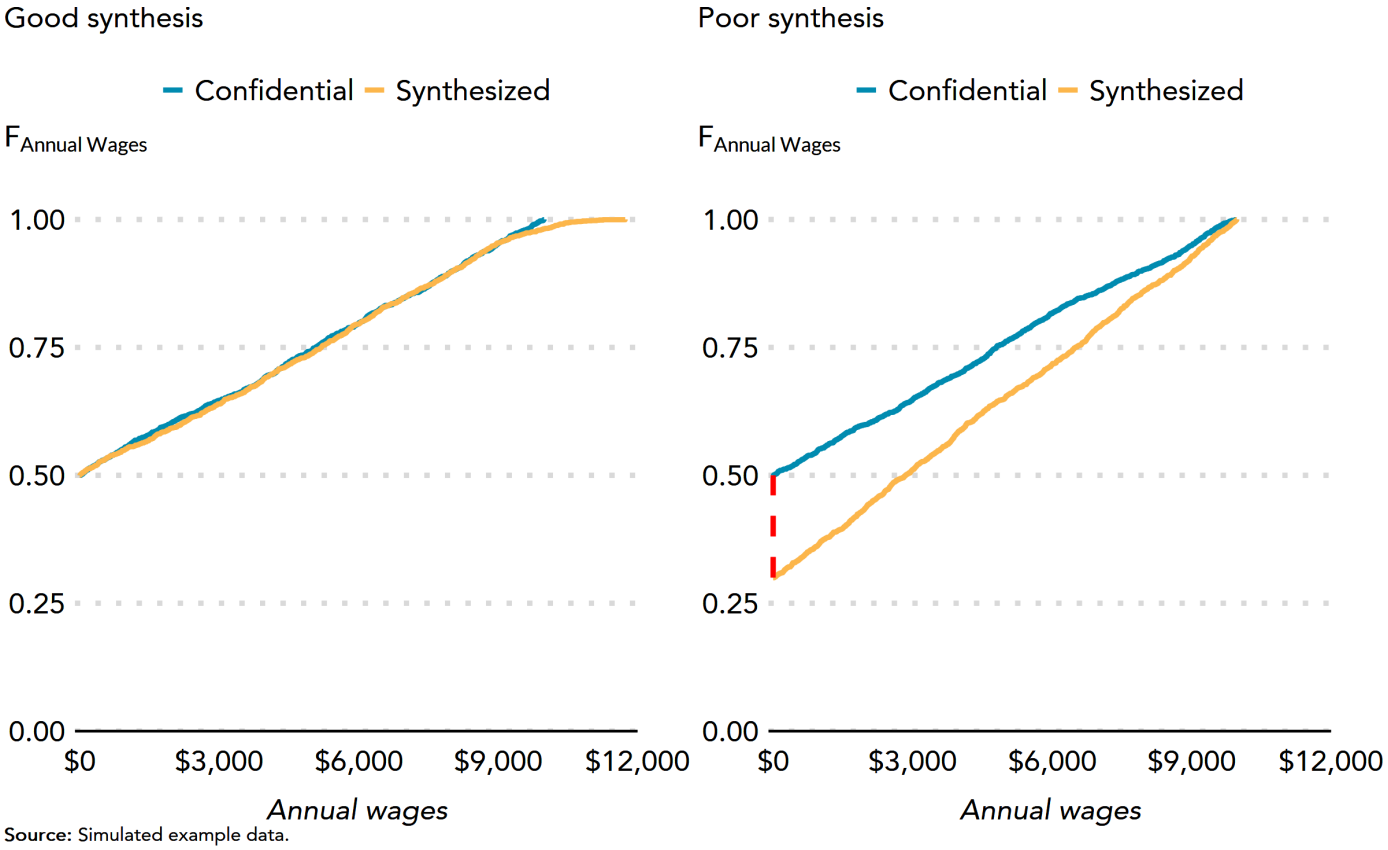
The KS statistic for the above samples and ECDFs is

$$D_{n,m} = \sup |F_{n,1}(x) - F_{m,2}(x)|. \tag{13}$$

This KS test essentially finds the largest absolute vertical distance between the two ECDFs and estimates the probability that it occurred by chance.

FIGURE 5

Example Kolmogorov-Smirnov Test



Source: Simulated example data.

The null hypothesis is rejected at level α if

$$D_{n,m} > \sqrt{-\frac{1}{2}\ln(\alpha)}\sqrt{\frac{n+m}{nm}} \quad (14)$$

If the test statistic is greater than the critical value, then we reject the null hypothesis that the samples come from the same underlying distributions. Figure 5 demonstrates the visual difference between a good synthesis with a modest test statistic and a poor synthesis with a large test statistic.

pMSE is a statistical test of whether a model can distinguish between the confidential and the synthetic data. Woo and colleagues (2009) introduced and Snoke and colleagues (2018) enhanced a propensity score measure for comparing distributions and evaluating the general utility of synthetic data. Propensity scores are probabilities of group membership introduced by Rosenbaum and Rubin (1983). The propensity score measure for general utility models group membership between the original and synthetic data as a measure of distinguishability. Low distinguishability corresponds with high general utility. The procedure is as follows:

1. Combine the rows of the confidential dataset and the rows of the synthetic dataset into one dataset. Add an indicator variable with 0 for the confidential data and 1 for the synthetic data.

2. Calculate propensity scores to estimate the probability that a row in the combined dataset belongs to the synthetic dataset. The propensity scores are modeled with a classifier such as logistic regression or CART. The predictors are all variables in the combined data without interactions. Interactions up to a specified maximum order of interactions are possible, estimation often struggles to converge.
3. Calculate the probability expected if the data did not distinguish the synthetic data from the original data. The probability expected is the share of synthetic data in the combined data. In most cases, this will be 0.5 because the confidential dataset and the synthetic data set usually have the same number of rows.
4. Finally, calculate the utility statistic. The utility statistic is the mean squared difference between the calculated propensity scores and the probability expected if the data did not distinguish the synthetic data from the original data.

Let *pMSE* be the utility statistic propensity score mean squared error. Let *N* be the number of rows in the combined data set, $\widehat{p_i}$ be the estimated propensities, and p_0 be the probability expected of the synthetic data in the combined data (typically 0.5).

$$pMSE = \frac{1}{N}\sum(\widehat{p_i} - p_0)^2 \tag{15}$$

We focused on the p values from a test with the null case of synthesizing data from the correct generative model of the original data. Failure to reject the null case suggests high general utility. The test statistic is a function of the *pMSE* and sample sizes. Let n_1 be the number of observations in the original data set. Let n_2 be the number of observations in the synthetic data set. Let $N = n_1 + n_2$.

$$test\ statistic = pMSE\ N^3 \frac{n_2}{n_1^2} \tag{16}$$

The null distribution of the test statistic is χ^2 with degrees of freedom equal to the number of parameters involving synthesized variables in the propensity score minus 1.

Specific utility metrics

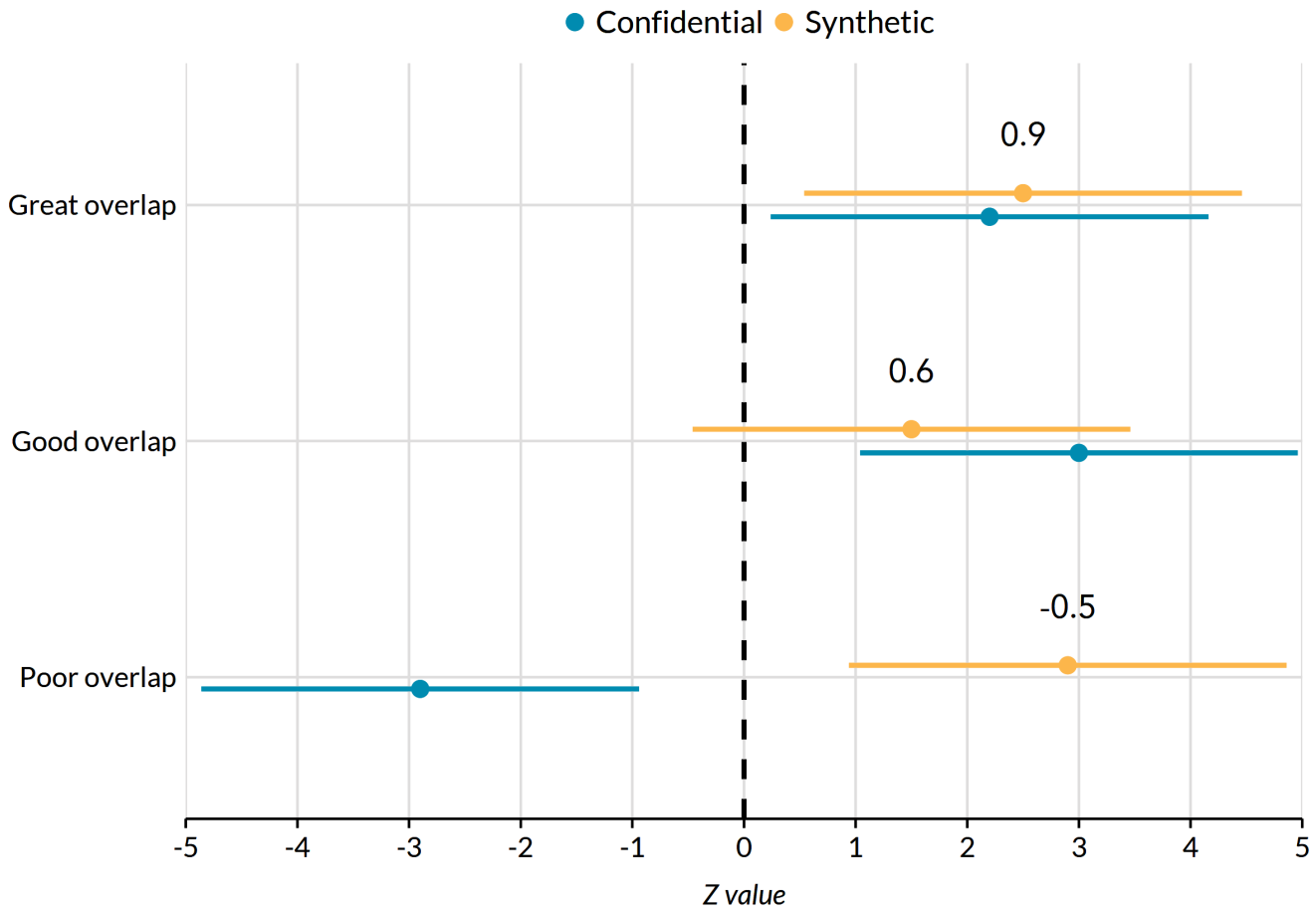
Regression confidence interval overlap (Karr et al. 2006) is a measure of the overlap between confidence intervals for each coefficient in a model estimated on the original data and a model estimated on the synthetic data. The overlap is calculated with the following where subscripts "o" and "s" denote the confidence interval bounds for the original and synthetic data:

$$IO = 0.5\left(\frac{\min(u_o,u_s) - \max(l_o,l_u)}{u_o - l_o} + \frac{\min(u_o,u_s) - \max(l_o,l_u)}{u_s - l_s}\right) \tag{17}$$

A value of 1 corresponds with perfect overlap between the intervals. A value of zero corresponds with no overlap but adjacent confidence intervals. Negative values correspond to the distance between intervals when the intervals do not overlap. Figure 6 demonstrates a great overlap, a good overlap, and a poor overlap.

FIGURE 6

Example Confidence Interval Overlap



Source: Simulated example data.

The synthetic Supplemental PUF dataset will be used for tax microsimulation. We built *a tax calculator* to compare calculations of adjusted gross income, personal exemptions, deductions, regular income tax, and tax on long-term capital gains and dividends based on the confidential data and the synthetic data. The tax calculator uses a simplified version of 2012 law (the year of the confidential and synthetic data). The calculator assumes that all individuals are single filers, it does not include any tax credits, it does not use standard or itemized deductions, and it lowers the personal exemption to $500. This unorthodox combination of rules is necessary to get useful calculations using the Supplemental PUF data, which come from a population that pays federal income tax only through withholding by payers of wages and other income (e.g., employers).

SUMMARY STATISTICS

Tables 2, 3, and 4 are based on all observations in the released synthetic dataset including rows with zeros for all 17 tax variables. All subsequent tables, figures, and metrics exclude rows that have zeros for every tax variable. This makes comparisons easier, and for tax microsimulation and analysis we are most interested in observations with nonzero values.

TABLE 2

Count of Genders by Data Source

Gender	Original	Synthetic
Female	13,082	13,086
Male	13,560	13,405

Source: Original IRS non-filer data; TPC synthetic non-filer data.

TABLE 3

Count of Age Groups by Data Source

Age Group	Original	Synthetic
1–17	855	812
18–24	2,298	2,362
25–34	2,807	2,727
35–54	6,081	6,085
55–64	3,886	3,842
65+	10,715	10,663

Source: Original IRS non-filer data; TPC synthetic non-filer data.

TABLE 4

Count of Age Groups by Data Source

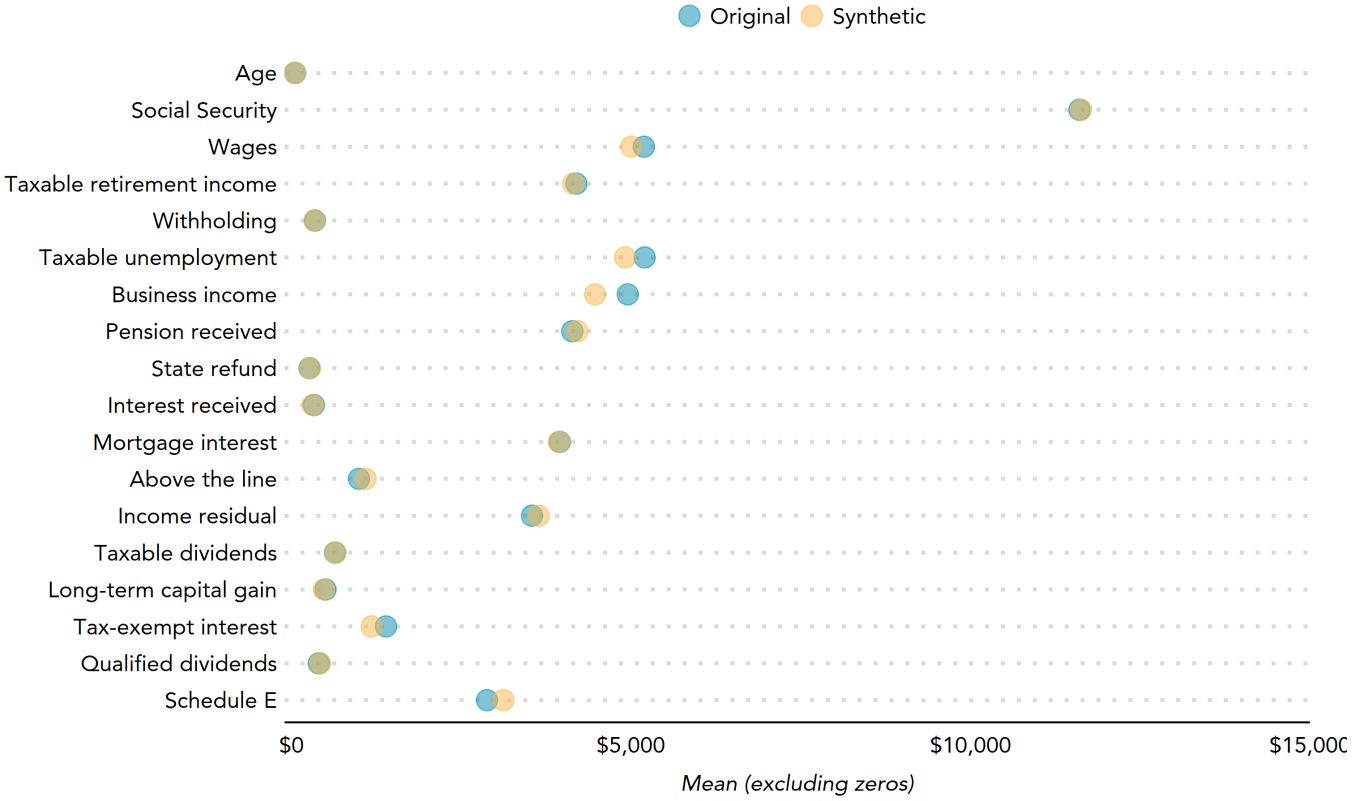
Age Group	Gender	Original	Synthetic
1–17	Female	412	415
1–17	Male	443	397
18–24	Female	1,030	1,047
18–24	Male	1,268	1,315
25–34	Female	1,090	1,069
25–34	Male	1,717	1,658
35–54	Female	2,494	2,442
35–54	Male	3,587	3,643
55–64	Female	1,892	1,903
55–64	Male	1,994	1,939
65+	Female	6,642	6,529
65+	Male	4,073	4,134

Source: Original IRS non-filer data; TPC synthetic non-filer data.

The data synthesis recreates the univariate distribution of the tax variables. Figures 7, 8, 9, and 10, respectively compare the mean, standard deviation, skewness, and kurtosis of the tax variables in the synthetic dataset with the tax variables in the confidential dataset. The four figures exclude any zeros.

FIGURE 7

Means from Original and Synthetic Data



Note: Calculations exclude all zeros.

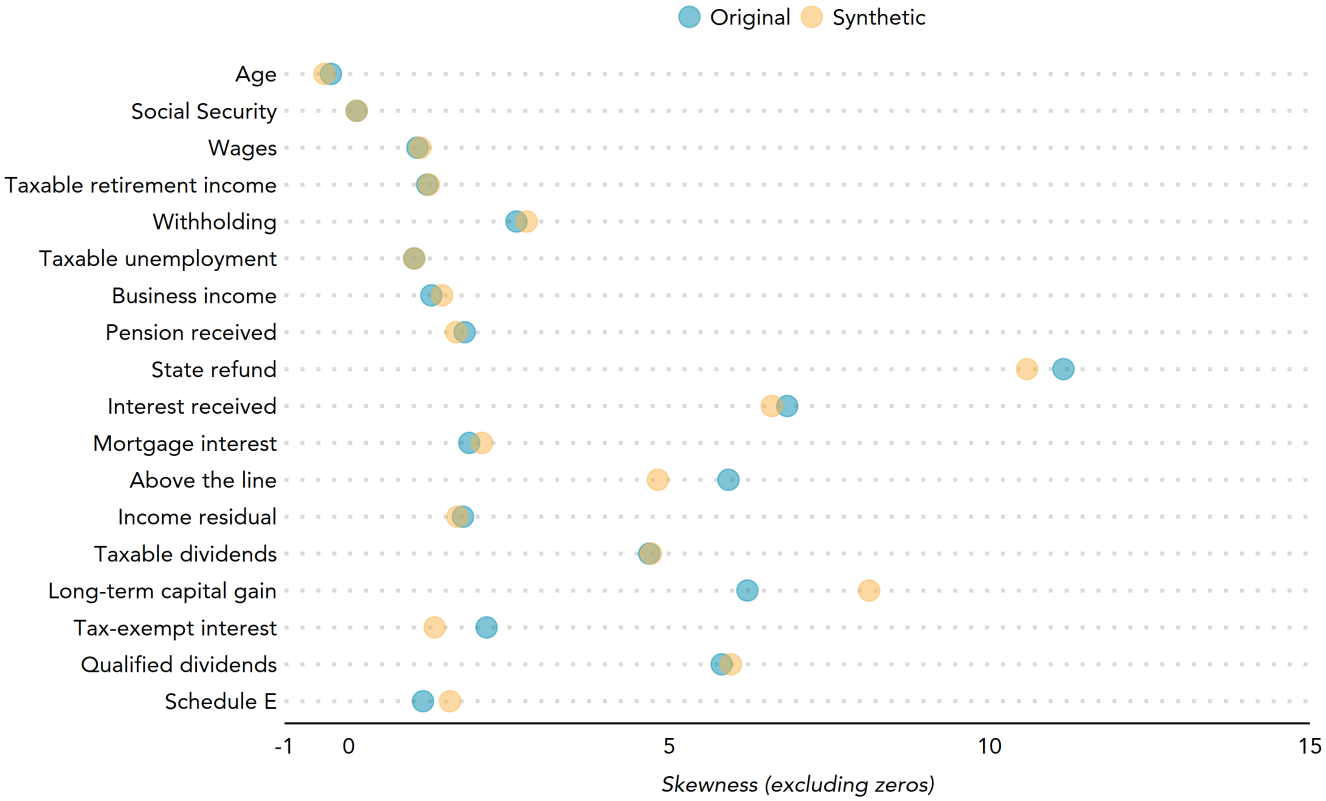
FIGURE 8

Standard Deviations from Original and Synthetic Data

Note: Calculations exclude all zeros.

FIGURE 9

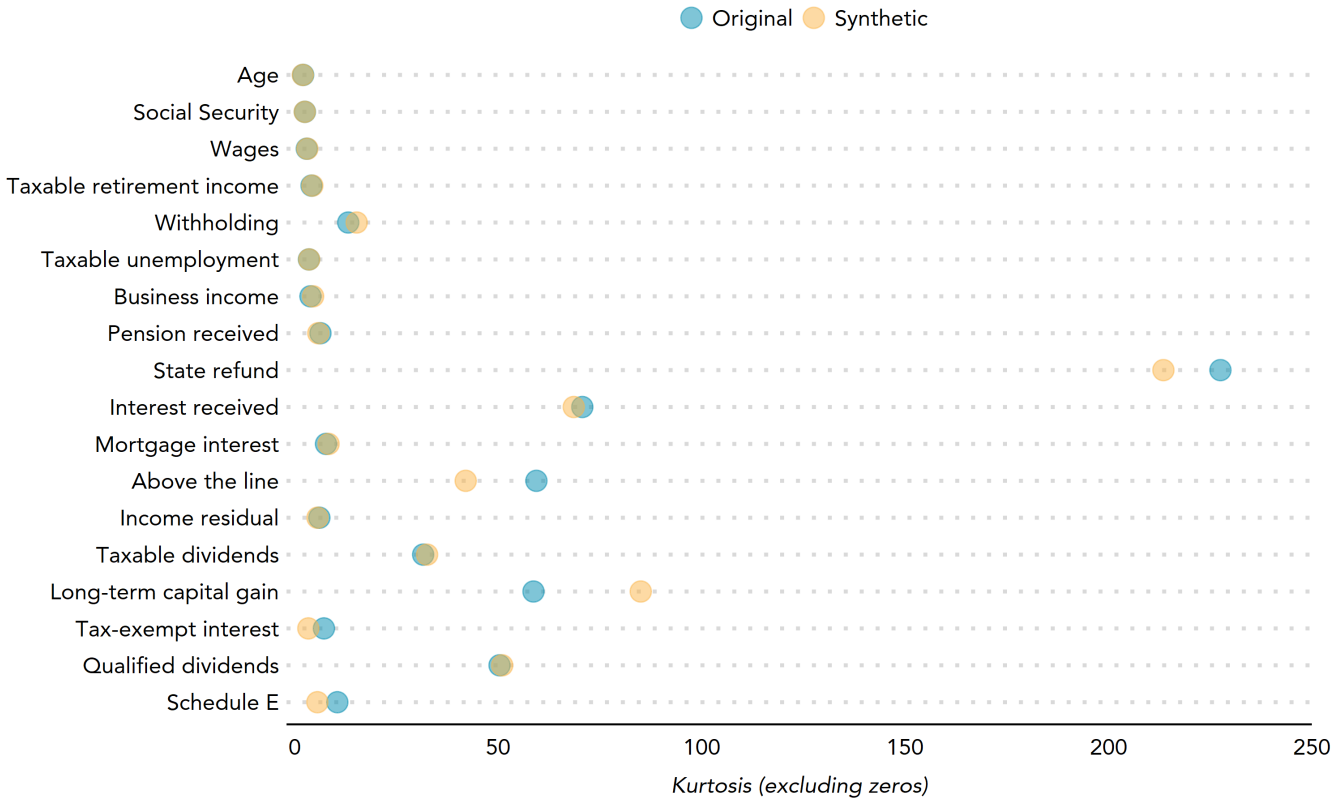
Skewness from Original and Synthetic Data



Note: Calculations exclude all zeros.

FIGURE 10

Kurtosis from Original and Synthetic Data



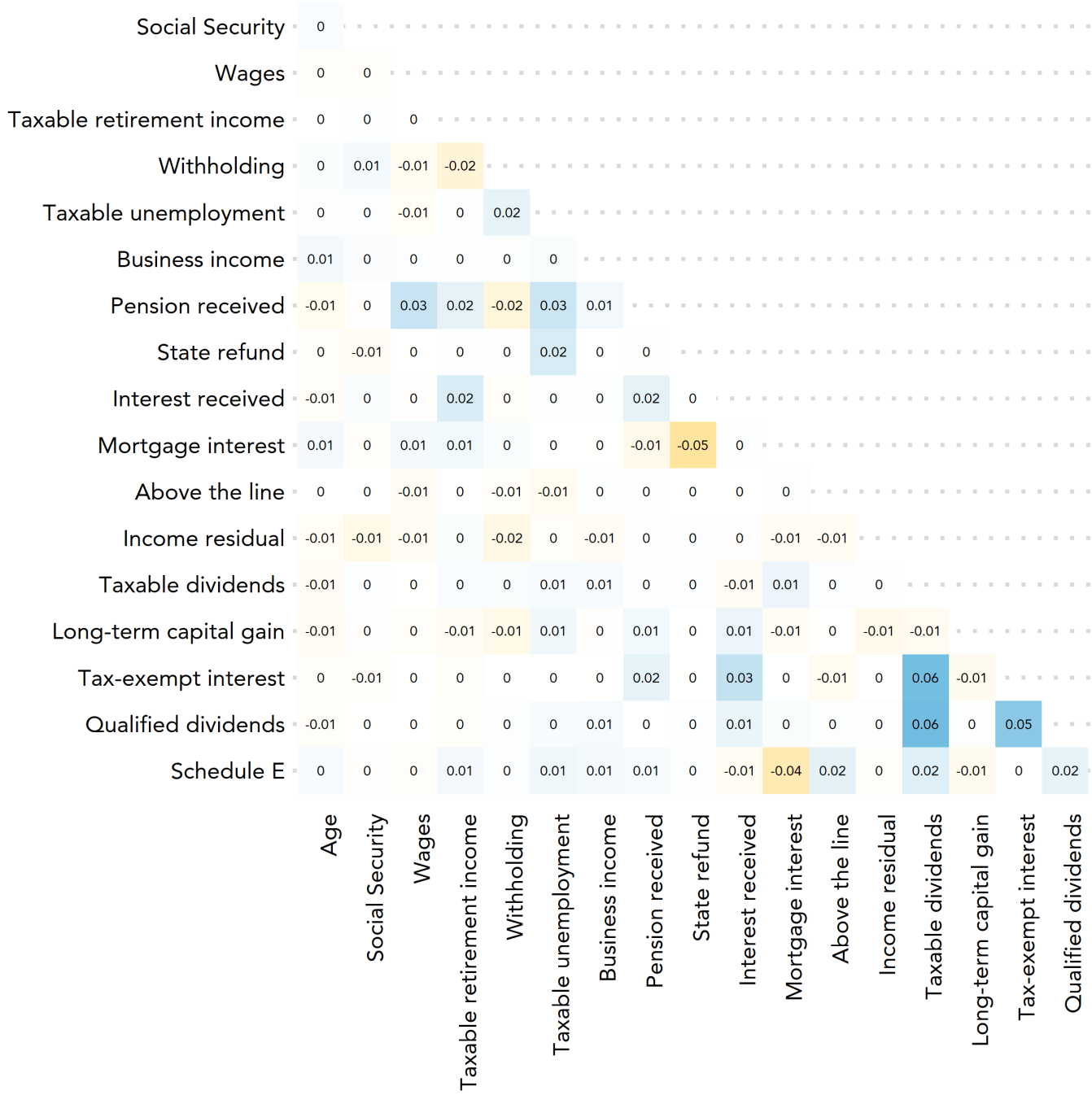
Note: Calculations exclude all zeros.

CORRELATION FIT

Our synthesizer also replicates the linear relationships between variables well. Overall, the correlation fit was 0.0013. Figure 11 illustrates the correlation difference between every combination of tax variables. Most differences are close to zero. Taxable dividends, qualified dividends, tax-exempt interest, and long-term capital gains all have correlation differences that are not close to zero. This is not surprising, because these variables have very few nonzero values and are uncommon sources of income for nonfilers. We do not consider this a cause for concern, but it is an area for future improvement.

FIGURE 11

Correlation Differences (Synthetic minus Original)



Note: Calculation excludes rows with zeros for all seventeen tax variables.

PMSE

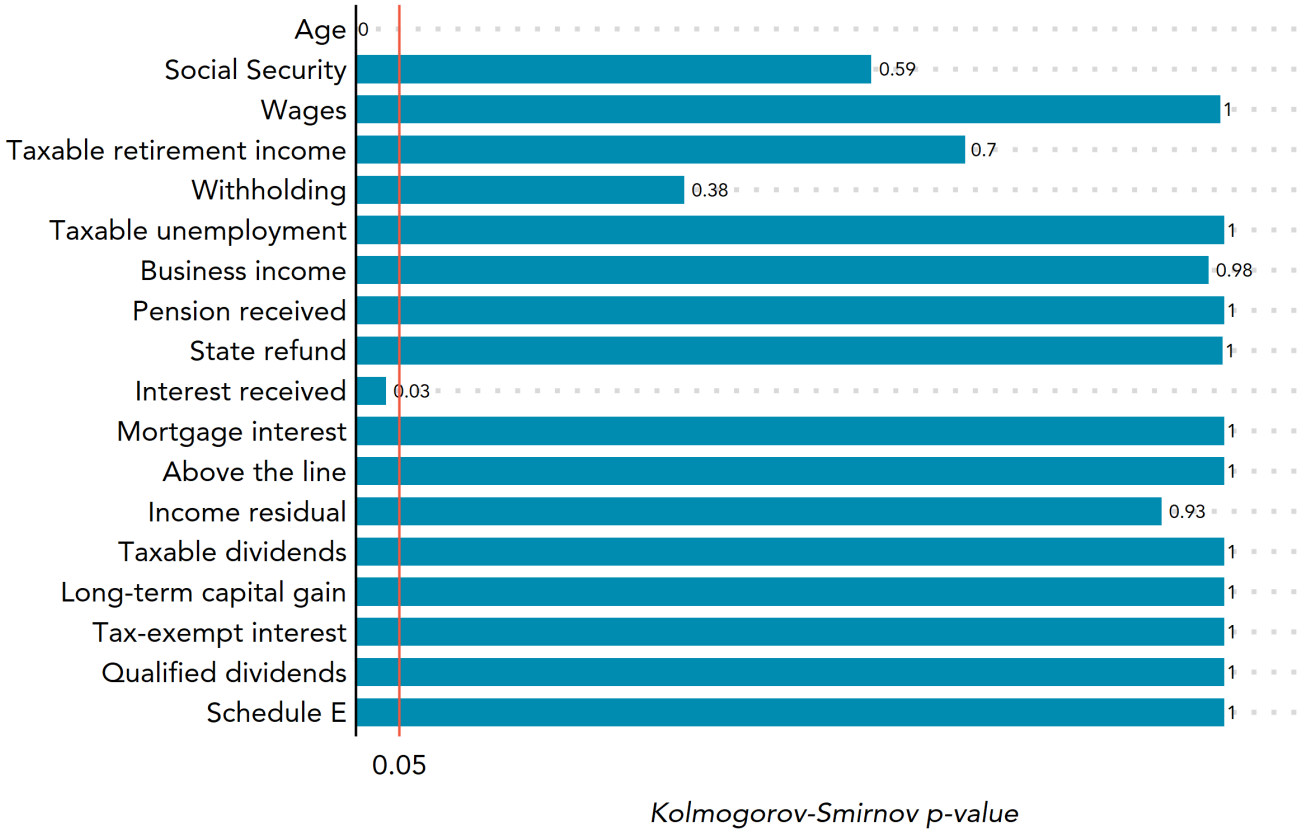
The p value of the pMSE with main effects and no interactions or higher-order terms is 0.26. This means we fail to reject the null hypothesis, which suggests distinguishing between the confidential and synthetic data is difficult.

KS TEST

The KS test also suggests that our data synthesis method performs well in recreating the univariate distributions of the tax variables. *Age* failed the KS test because of top coding but passes without top coding. *Interest received* fails the KS test because of rounding but passes without rounding. No other variables failed the KS test and, as figure 12 shows, none of the p values is near common cutoffs of 0.01, 0.05, or 0.1.

FIGURE 12

P-Values from Two-Sample Kolmogorov-Smirnov Tests on Original and Synthetic Data

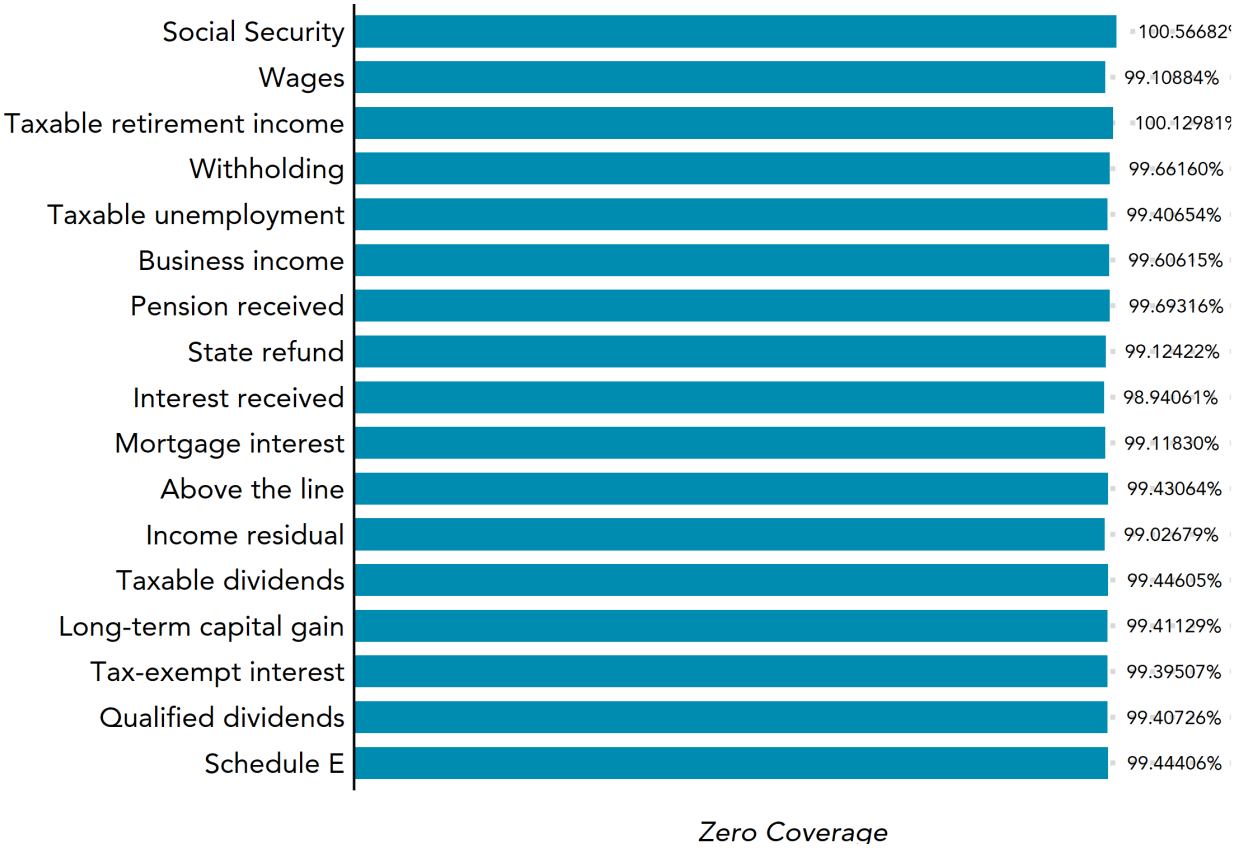


Note: Calculation excludes rows with zeros for all seventeen tax variables.

The synthesizer matched closely the number of zero values in each record. The proportion of zero values for each variable is within 1 percent of the correct number of zeros (figure 13).

FIGURE 13

Percentage of Values that are Zeros in the Synthetic Data Relative to the Original Data



Note: Calculation excludes rows with zeros for all seventeen tax variables.

CONFIDENCE INTERVAL OVERLAP

Figure 14 compares the coefficient estimates and confidence intervals for a regression with *wages* as the dependent variable and all other variables as independent variables. The figure is broken into three sections to ease visual comparisons. Most of the estimates are very close. The two variables with negative confidence interval overlaps are pension and withholding. Measured by *z* score, the estimates appear to differ markedly, but the difference is exaggerated because the standard deviations are small (especially for withholding, which is very highly correlated with wages). In the case of withholding, the actual coefficient is 5.6 and the coefficient in the synthetic sample is 5.1. For pensions, the estimates are -0.2 and -0.1, respectively.

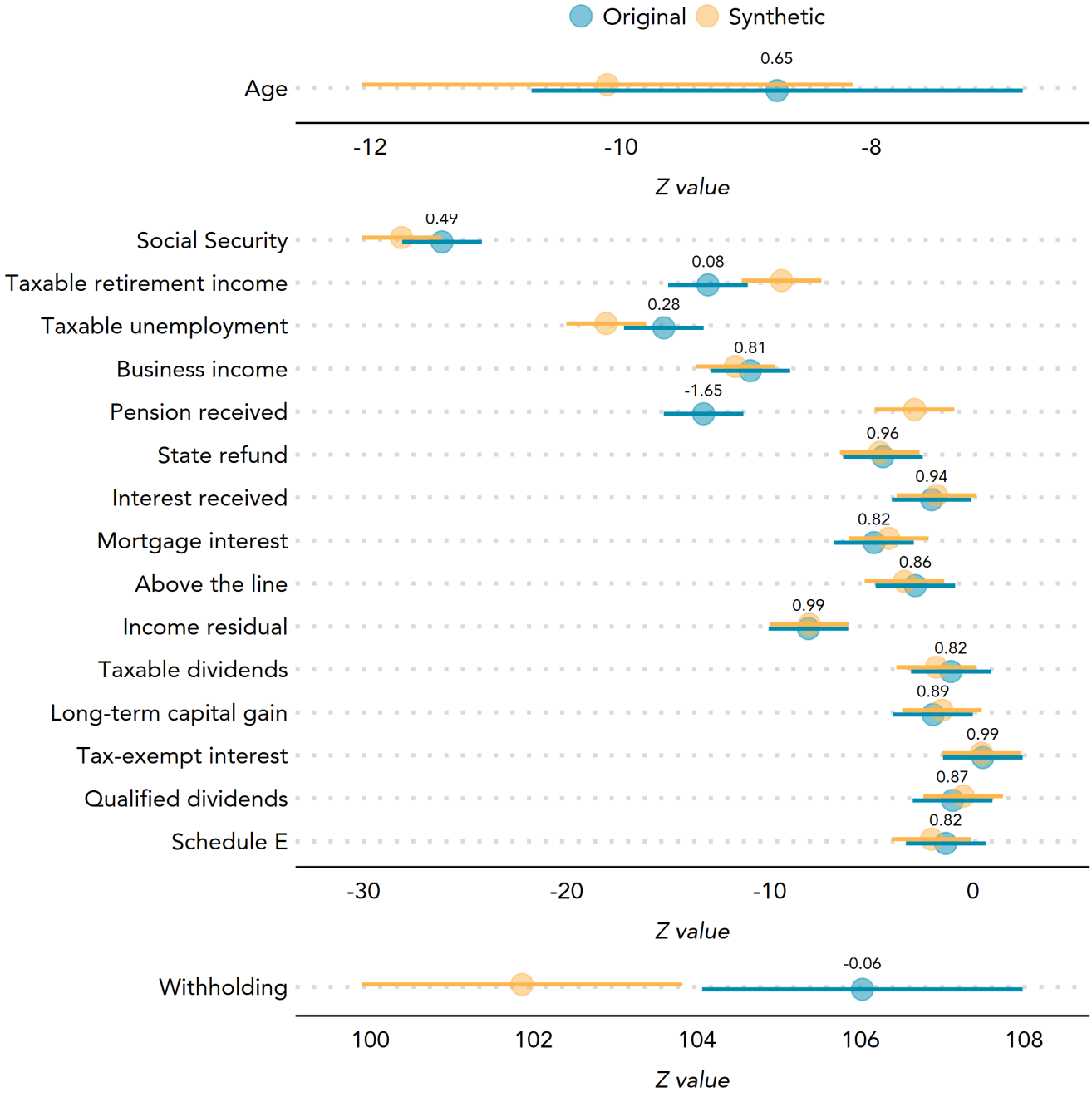
SIMPLIFIED TAX CALCULATOR

One use of the synthetic Supplemental PUF dataset is as an input to tax microsimulation. By definition, these records do not have federal income tax liability, but we created a simple alternative tax system that applies to low-income filers (as exists

in some states). The synthetic file performs well in our simple tax calculator and approximates the results from the confidential dataset. Figure 15 compares results for the original and synthetic datasets across different adjusted gross income groups for count, mean tax, and total tax.

FIGURE 14

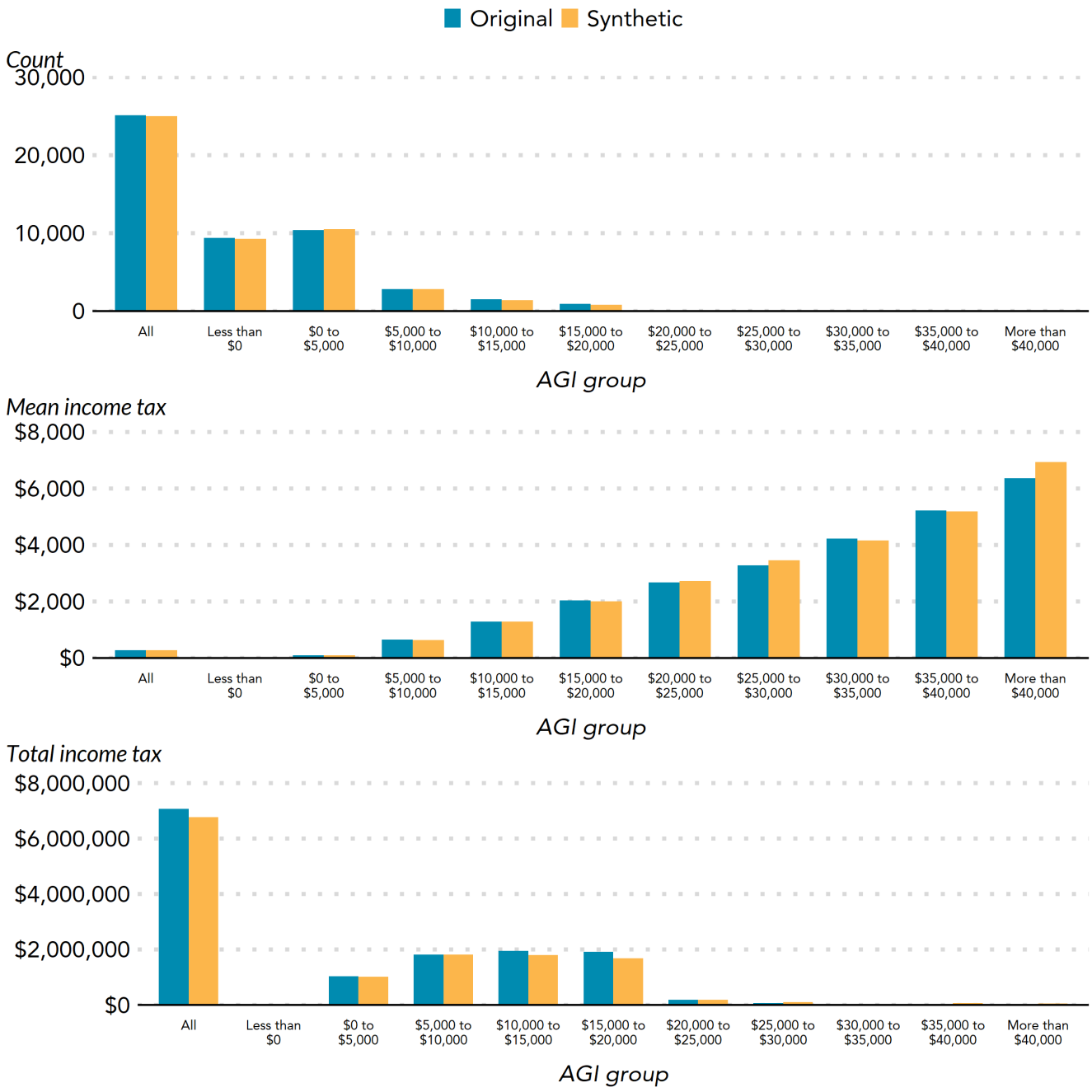
Coefficient Estimates and Confidence Intervals for a Regression with Wages as the Dependent Variable



Note: Calculation excludes rows with zeros for all seventeen tax variables. Confidence interval overlap is shown above each point estimate. The z value is the ratio of the estimated coefficient in the original data to the standard error of the regression estimate.

FIGURE 15

Simplified Tax Calculator Results for the Original and Synthetic Data



Note: Calculation excludes rows with zeros for all seventeen tax variables.

CONCLUSIONS AND PLANNED FUTURE WORK

In this report, we developed and evaluated a method to create a fully synthetic version of the IRS Supplemental PUF database. We demonstrated that the synthetic dataset would not allow a data intruder with extensive knowledge to meaningfully update his or her prior distribution about any variable on a tax return or even about whether someone had filed a tax return beyond statistical relationships between variables. Moreover, our method generated a synthetic data set that replicates the characteristics of the underlying administrative data while protecting individual information from disclosure.

In future work, we will develop a synthetic dataset based on the much more complex and diverse individual income tax return data. We do not know how well the data synthesis method used for the Supplemental PUF data will replicate the underlying distributions of individual income tax return data. For instance, we found that random forests performed worse than CART for the Supplemental PUF data, but random forests might outperform CART for the individual income tax return data. We plan to test a range of data synthesis methods. At a minimum, our goal is to create a synthetic file that protects individuals' privacy and reproduces the conditional means and variances of the administrative data. The synthetic data should also be useful for estimating the revenue and distributional effects of tax law changes and for other exploratory statistical analysis.

Experience suggests that the synthetic data will not provide accurate estimates for complex statistical models, so a key component of this project is to create a way for researchers to run their models using the actual administrative data with parameter estimates altered to protect privacy and standard errors adjusted accordingly. Other future work includes developing and establishing a *validation server*, a secure process to analyze the raw confidential data. This is a natural complement to the synthetic data because researchers could use the synthetic data, which have the same record layout as the confidential data, for exploratory analysis and to test and debug complex statistical programs. Vilhuber and Abowd (2016) describe a system that provides access to the confidential version of the Survey of Income and Program Participation and receives statistical output after a privacy review by a US Census Bureau staff person. Our goal is to create a similar system that would modify statistical outputs to guarantee privacy and preserve the statistical validity of estimates without requiring human review.

NOTES

1 Burman and colleagues (2018) describe the current procedures the IRS uses to produce a PUF, outlines various synthesis methods, and discusses the unique challenges of synthesizing tax return data.

2 Several papers have analyzed the confidential administrative data on nonfilers and compared them with information in survey datasets. See Cilke (2014); Mok (2017); and Langetieg, Payne, and Plumley (2017). All conclude that publicly available survey data provide biased estimates of the nonfiling population.

3 See National Research Council (1993) and Matthews and Harel (2011) for a discussion of data confidentiality and protecting privacy.

4 Matthias Templ, Bernhard Meindl, and Alexander Kowarik, "Introduction to Statistical Disclosure Control (SDC)," Comprehensive R Archive Network, February 11, 2020, https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf.

5 This approach is consistent with the advice of Machanavajjhala, Kifer, and Gehrke (2008, 285): "We believe that judicious suppression and separate modeling of outliers may be the key since we would not have to add noise to parts of the domain where outliers are expected."

6 Rick Wicklin, "How to Visualize a Kernel Density Estimate," *The DO Loop* (blog), July 27, 2016, https://blogs.sas.com/content/iml/2016/07/27/visualize-kernel-density-estimate.html.

7 In practice, our use of a kernel density estimator to approximate the distribution would add some additional noise to any synthetic data.

8 Here is how we calculate the confidence intervals shown in figure 2: The probability that the kth observation, x(k), is less than z is $\binom{n}{k}\left[F(z)^k\left(1-F(z)\right)^{n-k}\right]$. In the case of a uniform distribution on [0,1], F(z) = z, so the probability is simply $\binom{n}{k}[z^k(1-z)^{n-k}]$. The distribution of the kth order statistic of a random sample of size n, x(k), is approximately Beta(k, n – k + 1). Using the Beta distribution, we can derive the confidence interval around each order statistic. If we draw 100 observations, the distribution of each point is Beta(k, 100 – k + 1), k = 1,…, 100. If we use just 10 observations (a 1-in-10 sample), the distribution is Beta(k, 10 – k + 1), k = 1,…,10.

9 Some self-employed people may not owe income tax but still be required to file a Form 1040 because they owe payroll taxes under the Self-Employed Contributions Act (Langetieg, Payne, and Plumley 2017). We retain those people in the sample.

10 The sample is called the Continuous Work History Sample and has been maintained by the IRS for many decades, although some of the 10 digits were not selected in earlier years of the panel. The last four digits of Social Security numbers and of Individual Taxpayer Identification Numbers are randomly assigned, but 0000 is never assigned. Thus, only 9,999 four-digit endings are possibler.

11 Note that this peculiarity is limited to the information return dataset of nonfilers, where a sizable share of records have zero values for all variables other than age and gender. The individual income tax return data should always include at least one nonzero value; otherwise, the individual has no reason to file a tax return.

12 Based on US Census Bureau data, the age 85 cut-off groups together total about 2 percent of the adult population (3 percent of females and 1 percent of males). The percentages are probably higher for nonfilers because people whose income comes mostly or entirely from Social Security generally do not have a filing requirement.

13 Ordering from the variable with the most nonzero observations to the variable with the fewest nonzero observations is the norm for creating synthetic data, but we found that the correlation-order with Social Security benefits worked better in preliminary tests

REFERENCES

Abowd, John M., and Lars Vilhuber. 2008. "How Protective Are Synthetic Data?" In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Yücel Saygın, 239–46. Heidelberg, Germany: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-87471-3_20.

Benedetto, Gary, Martha H. Stinson, and John M. Abowd. 2013. "The Creation and Use of the SIPP Synthetic Beta." Washington, DC: US Census Bureau.

Bowen, Claire McKay., and Fang Liu. 2020. "Comparative Study of Differentially Private Data Synthesis Methods." *Statistical Science*, 35(2): 280-307.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. New York: Chapman and Hall.

Bryant, Victoria L. 2017. "General Description Booklet For the 2012 Public Use Tax File." Washington, DC: Internal Revenue Service, Statistics of Income Division, Individual Statistics Branch. https://users.nber.org/~taxsim/gdb/gdb12.pdf.

Burman, Leonard E., Alex Engler, Surachai Khitatrakun, James R. Nunns, Sarah Armstrong, John Iselin, Graham MacDonald, and Philip Stallworth. 2018. "Safely Expanding Research Access to Administrative Tax Data: Creating a Synthetic Public Use File and a Validation Server." Washington, DC: Urban-Brookings Tax Policy Center.

Cilke, James. 2014. "The Case of the Missing Strangers: What We Know and Do Not Know about Non-Filers." Washington, DC: Joint Committee on Taxation. https://www.ntanet.org/wp-content/uploads/proceedings/2014/029-cilke-case-missing-strangers-know-don.pdf .

Dinur, Irit, and Kobbi Nissim. 2003. "Revealing Information while Preserving Privacy." In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 202–10. San Diego, CA: ACM Press. https://doi.org/10.1145/773153.773173.

Drechsler, Jörg, and Jerome P. Reiter. 2010. "Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata." *Journal of the American Statistical Association* 105 (492): 1347–57. https://doi.org/10.1198/jasa.2010.ap09480.

Duncan, George, and Diane Lambert. 1989. "The Risk of Disclosure for Microdata." *Journal of Business and Economic Statistics* 7 (2): 207–17.

Dwork, Cynthia. 2008. "Differential Privacy: A Survey of Results." In *Theory and Applications of Models of Computation*, edited by Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, 1–19. Heidelberg, Germany: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-79228-4_1.

Elliot, Mark. 2014. "Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team." http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf.

Fellegi, I. P. 1972. "On the Question of Statistical Confidentiality." *Journal of the American Statistical Association* 67 (337): 7–18.

Fienberg, Stephen E., and Jiashun Jin. 2009. "Statistical Disclosure Limitation for Data Access." In Encyclopedia of Database Systems, edited by Ling Liu and M. Tamer Özsu. New York: Springer.

Fienberg, Stephen E., Udi E. Makov, and Ashish P. Sanil. 1997. "A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data." *Journal of Official Statistics* 13, (1): 75–89.

Fuller, Wayne A. 1993. "Masking Procedures for Microdata Disclosure Limitation." *Journal of Official Statistics* 9 (2): 383–406.

Hu, Jingchen, Jerome P. Reiter, and Quanli Wang. 2014. "Disclosure Risk Evaluation for Fully Synthetic Categorical Data." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer, 185–99. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-11257-2_15.

IRS (Internal Revenue Service). 2019. "2012 Supplemental Public Use File." Washington, DC: IRS.

Karr, Alan F, C. N. Kohnen, Anna Oganian, J. P. Reiter, and A. P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." American Statistician 60 (3): 224–32. https://doi.org/10.1198/000313006X124640 .

REFERENCES

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." International Statistical Review 79 (3): 362–84.

Kuhn, Max, and Kjell Johnson. 2016. *Applied Predictive Modeling*. New York: Springer.

Langetieg, Patrick, Mark Payne, and Alan Plumley. 2017. "Counting Elusive Nonfilers Using IRS Rather Than Census Data." In *IRS Research Bulletin, Papers Given at the 7th Annual Joint Research Conference on Tax Administration*, edited by Alan Plumley, 197–222. Washington, DC: Internal Revenue Service.

Machanavajjhala, Ashwin, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In 2008 IEEE 24th International Conference on Data Engineering, 277–86. Cancun, Mexico: IEEE. https://doi.org/10.1109/ICDE.2008.4497436.

Machanavajjhala, Ashwin, Johannes Gehrke, and Daniel Kifer. 2006. "L-Diversity: Privacy beyond k-Anonymity." Paper presented at the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, April 3–7.

Matthews, Gregory J., and Ofer Harel. 2011. "Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy." *Statistics Surveys* 5 (0): 1–29. https://doi.org/10.1214/11-SS074.

McClure, David, and Jerome P Reiter. 2012. "Differential Privacy and Statistical Disclo- Sure Risk Measures: An Investigation with Binary Synthetic Data." Transactions on Data Privacy 5, no. 3: 535–52.

Mitra, Robin, and Jerome P. Reiter. 2006. "Adjusting Survey Weights When Altering Identifying Design Variables Via Synthetic Data." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Luisa Franconi, 177–88. Heidelberg, Germany: Springer Berlin Heidelberg, https://doi.org/10.1007/11930242_16.

Mok, Shannon. 2017. "An Evaluation of Using Linked Survey and Administrative Data to Impute Nonfilers to the Population of Tax Return Filers." Washington, DC: Congressional Budget Office. https://www.cbo.gov/publication/53125.

National Research Council. 1993. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, DC: National Academies Press.

Nowok, Beata, Gillian M Raab, Joshua Snoke, and Chris Dibben. 2019. Synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/synthpop/index.html.

Raab, Gillian M., Beata Nowok, and Chris Dibben. 2017. "Practical Data Synthesis for Large Samples." *Journal of Privacy and Confidentiality* 7 (3): 67–97.

Reiter, Jerome P. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics* 18 (4): 531–43.

———. 2005a. "Using CART to Generate Partially Synthetic Public Use Microdata." *Journal of Official Statistics* 21 (3): 441–62.

———. 2005b. "Estimating Risks of Identification Disclosure in Microdata." *Journal of the American Statistical Association* 100 (472): 1103–12. https://doi.org/10.1198/016214505000000619.

Reiter, Jerome P., Quanli Wang, and Biyuan Zhang. 2014. "Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data." *Journal of Privacy and Confidentiality* 6 (1): 17–33. https://doi.org/10.29012/jpc.v6i1.635.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 1993. "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68.

Ruggles, Steven. 2018. "Implications of Differential Privacy for Census Bureau Data and Scientific Research." Minneapolis: Minnesota Population Center. https://assets.ipums.org/_files/mpc/wp2018-06.pdf.

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Syntheticdata." *Journal of the Royal Statistical Society* 181 (3): 663–88.

Sweeney, Latanya. 2002. "K-Anonymity: A Model for Protecting Privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (5): 557–70.

REFERENCES

Therneau, Terry M., and Elizabeth J. Atkinson. 2019. "An Introduction to Recursive Partitioning Using the RPART Routines." Rochester, MN: Mayo Foundation.

Vilhuber, Lars, and John M. Abowd. 2016. "Usage and Outcomes of the Synthetic Data Server. Presentation at the Society of Labor Economics Meetings, Cornell University, May 9.

Winkler, William E. 2007. "Examples of Easy-to-Implement, Widely Used Methods of Masking for Which Analytic Properties Are Not Justified." Research report series, Statistics #2007-21. Washington, DC: US Census Bureau. https://www.census.gov/srd/papers/pdf/rrs2007-21.pdf.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1 (1): 111–24. https://doi.org/10.29012/jpc.v1i1.568 .

Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David O'Brien, Thomas Steinke, and Salil Vadhan. 2018. "Differential Privacy: A Primer for a Non-Technical Audience." Research publication 2019-2. Amsterdam: Elsevier SSRN. https://doi.org/10.2139/ssrn.3338027 .

Yancey, William E., William E. Winkler, and Robert H. Creecy. 2002. "Disclosure Risk Assessment in Perturbative Microdata Protection." In *Inference Control in Statistical Databases*, edited by Josep Domingo-Ferrer, 135–52. Heidelberg, Germany: Springer Berlin Heidelberg, https://doi.org/10.1007/3-540-47804-3_11.

ABOUT THE AUTHORS

Claire McKay Bowen is the lead data scientist for privacy and data security at the Urban Institute. She develops and researches algorithms that create synthetic versions of confidential data, such as income tax return data, and assesses the quality and privacy guarantee of the generated synthetic data. Bowen obtained her honors BS in mathematics and physics from Idaho State University before receiving her MS and PhD in statistics at the University of Notre Dame. Before her current position, Bowen worked at the Los Alamos National Laboratory, where she investigated cosmic ray effects on supercomputers and developed functional data analysis techniques for iterative design problems. Bowen is a past National Science Foundation Graduate Research Fellow.

Victoria L. Bryant is a Senior Economist at the Internal Revenue Service's Statistics of Income Division. She works primarily on tax research development on topics ranging from retirement savings, intergenerational mobility and wealth, information returns linkage, to data confidentiality. She earned her BA in Economics from Virginia Tech in 2003, her MA in Economics from George Mason in 2008, and her PhD in Public Policy also at George Mason in 2020.

Leonard E. Burman is an Institute Fellow at the Urban Institute, the Paul Volcker Professor of Public Administration and International Affairs at the Maxwell School of Syracuse University, and senior research associate at Syracuse University's Center for Policy Research. He co-founded the Tax Policy Center, a joint project of the Urban Institute and the Brookings Institution, in 2002. He was Deputy Assistant Secretary for Tax Analysis at the Treasury from 1998 to 2000 and Senior Analyst at the Congressional Budget Office from 1988 to 1997. He is past-president of the National Tax Association (NTA) and 2016 recipient of the NTA's Davie-Davis Award for Public Service. Burman is the coauthor with Joel Slemrod of Taxes in America: What Everyone Needs to Know and author of The Labyrinth of Capital Gains Tax Policy: A Guide for the Perplexed, and co-editor of several books. He is often invited to testify before Congress and has written for scholarly journals as well as media outlets such as the Washington Post, New York Times, and the Wall Street Journal. He holds a Ph.D. from the University of Minnesota and a B.A. from Wesleyan University**.**

Surachai Khitatrakun is a Senior Research Associate at the Urban-Brookings Tax Policy Center. Khitatrakun extends the TPC's microsimulation model of the federal tax system to incorporate aspects of education, health, and retirement saving issues and to analyze federal income tax distributions among the states.

Robert McClelland is a senior fellow in the Urban-Brookings Tax Policy Center. Previously, he worked in the tax analysis division of the Congressional Budget Office (CBO), where he examined the impact of federal tax policy on charitable giving and bequests, the realization of capital gains, labor supply, and small businesses. He worked for the CBO from 1999 to 2005 and from 2011 to 2016, and in between, he directed the division of price and index number research at the Bureau of Labor Statistics. He is a member of the Conference on Research in Income and Wealth. He taught econometrics at Johns Hopkins University for 20 years, where he won an Excellence in Teaching award in 2006. He received a PhD in economics from the University of California, Davis.

ABOUT THE AUTHORS

Philip Stallworth worked as a Research Analyst at the Urban-Brookings Tax Policy Center from 2016 - 2019. He primarily worked on the Tax Policy Center's federal tax model. He is currently a JD Candidate at the University of Michigan Law School. He holds a bachelor's degree in mathematics with a concentration in statistics from Reed College.

Kyle Ueyama (M.A. Quantitative Methods in the Social Sciences, Columbia University) is a Senior Research Programmer and Data Scientist in the Office of Technology and Data Science at the Urban Institute. Ueyama has worked on a wide range of Urban Institute projects including Urban's Modeling in the Cloud initiative. He has developed custom interfaces, tools, and websites that allow users to interact with data on multiple platforms, including the Education Data Portal. He is an experienced R, Python, and Stata programmer and an Amazon Web Services certified solutions architect associate.

Aaron R. Williams is a data scientist in the Income and Benefits Policy Center and Program on Retirement Policy at the Urban Institute, where he works on retirement policy, tax policy, microsimulation models, data imputation methods, and data visualization. He has worked on Urban's Dynamic Simulation of Income (DYNASIM) microsimulation model and the Social Security Administration's Modeling Income in the Near Term (MINT) microsimulation model. Williams leads the Urban Institute's R Users Group and assists researchers across Urban with projects that use R for statistical analysis, data visualization, mapping, and automation. Previously, Williams worked at the Commonwealth Institute for Fiscal Analysis. He holds a BS in economics and a BA in music from Virginia Commonwealth University.

Graham MacDonald is chief data scientist at the Urban Institute, where he works with researchers to improve access to data, analytics tools, and innovative research methods. MacDonald uses and advises on such tools as machine learning, natural language processing, web scraping, big data platforms, and data visualization techniques and their application to relevant public policy issues.

Before starting his current role, MacDonald worked for the Terner Center for Housing Innovation, the San Francisco Planning Department, and the California Housing Partnership Corporation, where he used innovative data collection, machine learning, data linking, and microsimulation techniques to solve problems. Before that, he was a research associate in the Urban Institute's Metropolitan Housing and Communities Policy Center and was the lead data visualization developer for interactive digital communications projects.

MacDonald holds a BA in economics from Vanderbilt University and earned an MPP from the University of California, Berkeley.

Noah Zwiefel is a research assistant in the Tax Policy Center at the Urban Institute. Zwiefel primarily works on TPC's federal tax model. He graduated summa cum laude and with honors from Macalester College, where he holds a BA in economics.

The Tax Policy Center is a joint venture of the
Urban Institute and Brookings Institution.

BROOKINGS

For more information, visit taxpolicycenter.org
or email info@taxpolicycenter.org