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

Len Burman, Surachai Khitatrakun, Philip Stallworth, Kyle Ueyama, and Aaron Williams

Tax Policy Center, Urban Institute

Draft of August 28, 2019
Comments Welcome

Len Burman, Urban Institute/Tax Policy Center, and Syracuse University/Center for Policy Research; Surachai Khitatrakun, Urban Institute/Tax Policy Center; Philip Stallworth, University of Michigan; Kyle Ueyama, Urban Institute; Aaron Williams, Urban Institute. We are grateful to Joe Ansaldi, Vickie Bryant, Jim Cilke, John Czajka, Rick Evans, Barry Johnson, Graham MacDonald, Rob McClelland, 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.

This draft is preliminary. Please do not quote without permission. Please send comments to lburman@urban.org. This paper relies on the analytical capability that was made possible in part by a grant from the Laura and John Arnold Foundation. The findings and conclusions are those of the authors and do not necessarily reflect positions or policies of the Tax Policy Center or its funders.

1. Introduction

The Internal Revenue Service (IRS) requires individuals and businesses to file annual income tax returns and related information returns for purposes of administering the income tax. These returns contain detailed information on the income, deductions, taxes, and credits of individuals and businesses. For individuals, these returns also provide key demographic information, such as taxpayers' marital status, their number of children, and the age and gender of both taxpayers and dependents (available from the Social Security Administration). Administrative tax data are comprehensive because 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 are generally represented on one or more information returns. Administrative tax data are of high quality because most taxpayers file returns that are quite complete and accurate, and because filing a false return (or failure to file a return) can result in severe penalties.

The detailed, comprehensive, and high-quality administrative tax data are enormously valuable for analysis and research that can help inform the public about a wide range of issues. For example, analysts use information from actual tax returns to model the effect of current or proposed law on the distribution of tax burdens across income groups, on incentives to work or to make charitable contributions, and on other effects of the income tax. All such microsimulation models are based on a sample of tax returns that is weighted to represent the entire filing population, but only the Congressional Joint Committee on Taxation (JCT) and the Treasury's Office of Tax Analysis (OTA) have statutory authority to use the confidential IRS administrative data in their models. The microsimulation models of other groups, such as the American Enterprise Institute and Urban-Brookings Tax Policy Center, must rely on a sample that has been heavily edited to avoid disclosure of any individual taxpayer's confidential information. Privacy protections are required by section 6103 of the Internal Revenue Code, which strictly limits access to tax return information. Research based on tax return information is also circumscribed by the requirements of section 6103. 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 JCT and OTA or in collaboration with them, or through a highly restrictive arrangement with the IRS.

We propose a two-part approach to expanding analysts' and researchers' access to tax data: (1) creating fully synthetic public use files–publicly releasable tax return datasets that replace all confidential administrative tax data with imputed data that protects the privacy of all taxpayers and preserves the relationships among variables necessary for valid analysis and research;[1] and (2) developing a secure process by which researchers can submit statistical programs to be executed on

[1] Burman, et al (2018) describe the current procedures the IRS uses to produce a public use file, outlines various synthesis methodologies and discusses the unique challenges of synthesizing tax return data.

the confidential administrative tax data that have been tested on the synthetic data, with all statistical results altered as necessary to protect taxpayer privacy.

Many public datasets are now partially synthetic, with some sensitive variables replaced by imputed values. Variables are deemed sensitive if an intruder could match them with unique data available in another database. By matching the externally available data to the confidential database, a particular record could be identified, disclosing all the other information contained in that observation. Because partially synthetic data contain some actual data, they always carry a risk of disclosure, which grows as more data become publicly available to potential intruders. Our approach addresses this concern by proposing fully synthetic tax return databases.

A number of synthesis techniques have been used in previous applications, including parametric (e.g., regression) and nonparametric models. One nonparametric method, classification and regression trees (CART), sorts observations into relatively homogeneous groups and draws from the empirical distribution of each group. The method is computationally simple and relatively flexible. We have used this method to date because CART out-performed regression-based parametric methods.

We also propose to establish a secure method for researchers to submit statistical programs to run on a subset of the confidential administrative data. This model for research access to confidential data is referred to as a *validation server*. The synthetic dataset would have the same structure as the administrative data, so programs that are developed using the synthetic dataset would work on the confidential data with minimal alteration. Vilhuber and Abowd (2016) describe a system to provide access to the confidential version of the Survey of Income and Program Participation and receive statistical output after a privacy review by Census staff. We are exploring the creation of a similar system that would modify statistical outputs to guarantee privacy and preserve the statistical validity of estimates but without requiring human review. We will discuss the proposed validation server model in a subsequent paper.

This paper describes a synthesis methodology that we have implemented and tested on a database of individuals who did not file and were not dependent of any individual income tax return in 2012. This file is called the 2012 Supplemental Public Use File. More information about the use of the file is available in *2012 Supplemental Public Use File* (Internal Revenue Service, 2019).

The underlying administrative data are the information returns filed by employers, financial institutions, the Social Security Administration, and other entities that pay income to individuals or have certain other transactions with individuals. Prior to our work, no public use file had been created from information returns.[2] A synthetic version of these data would protect individuals'

[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.

privacy while allowing researchers to gain a fuller picture of the distribution of income and tax burdens than one derived from income tax filing data alone.

We show that the methodology protects privacy because it would be impossible for an intruder—even if possessing extensive information about most records in the administrative dataset—to determine with certainty if a particular 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. Since the synthetic data are imputed, the methodology also protects against disclosure of any individual's confidential information in the underlying administrative data.

The paper is organized as follows. Section 2 defines privacy and disclosure, and summarizes the characteristics of our synthesis process that protect privacy. Section 3 defines data utility and summarizes how it is maintained by our synthesis process. Section 4 provides an overview of the CART synthesis method, and Section 5 provides a detailed description of the method. Section 6 shows how our synthesis process protects privacy, including protections against disclosure of outliers and attribute disclosure. Section 7 describes the Supplemental Public Use File data and how we synthesized them. Section 8 describes measures of the privacy of synthetic data, and the results of applying these measures to the synthesized data. Section 9 describes data utility measures and applies them to the synthesized Supplemental Public Use File data. Section 10 offers conclusions and outlines our plans to extend this work to the production of a fully synthetic income tax return public use file.

2. Privacy and Confidentiality

A legal and moral imperative of this project is to protect the confidentiality of individual taxpayer information.[3] Assessing the confidentiality of a dataset is challenging because threats to privacy are continually evolving. The following section provides some key definitions. Subsequent sections discuss privacy protection methods and standards that have been implemented or proposed.

a. Some 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 is when an intruder associates an individual with a specific record in the released data (Templ et al. 2019) and it discloses all the variables in the data set with respect

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

to that individual. This can be quite damaging. For example, an insurance company might increase insurance premiums for a participant based on information about health status inferred from a medical survey, a credit card company could increase interest rates for an individual based on data gleaned from a wealth survey, or a divorce lawyer might demand a larger settlement based on income data inferred from an income tax return.

Attribute disclosure is when an intruder can determine certain characteristics of an individual based on information in the released data (Templ et al. 2019). This doesn't necessarily require identifying an individual in the data. For example, if all individuals in a Census block are of one race and ethnicity, then it is possible to know the race and ethnicity of someone who lives in the block without identifying the individual in the data. While appearing less harmful, attribute disclosure can create some of the same damages as identity disclosure.

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

Notwithstanding the potential cost to a participant or others, 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.

b. **Limitations of traditional methods for statistical disclosure**

To avoid disclosure, data stewards have relied on a variety of statistical disclosure limitation techniques. However, many standard statistical disclosure limitation techniques fail to eliminate disclosure risk (Dreschler and Reiter, 2010; Winkler 2007). In addition, these techniques may greatly reduce the usefulness of the released data for analysis and research. In particular:

Adding random noise to continuous variables can maintain univariate distributions and prevent exact matches with external data sources. But adding random noise to sensitive variables creates measurement error in the perturbed variables that reduces the precision of statistical analyses and may introduce 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) found that even 1 percent swapping of a subsample from the March 2000 U.S. Current Population Survey can undermine statistical inference.

Top and bottom coding combine all values above or below a threshold into a single value. For example, for the individual income tax return public use file (PUF), the IRS currently top codes number of children 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 doesn't affect order statistics below where top coding begins and bottom coding doesn't affect order statistics above where bottom coding begins, but top and bottom coding eliminate information about the tails of distributions and thus degrade analyses that require the entire distribution (Reiter, Wang, and Zhang, 2014; Fuller 1993).

Aggregation combines multiple observations into one observation. The 2012 PUF aggregated 1,155 returns with extreme values into four observations in the microdata (Bryant 2017). Aggregation doesn't alter simple statistics such as sums or means, but it may bias estimates from more complex statistical models and distort microsimulation model analyses. Furthermore, aggregation of geographies may make small area estimation impossible and hides spatial variation (Reiter, Wang, and Zhang, 2014).

c. Fully synthetic data and identity disclosure

Fully synthetic data have the potential to avoid pitfalls of traditional statistical disclosure limitation techniques because the methods seek to replicate the data generation process of the confidential data while not disclosing the identity or attributes of any individual.

Fully synthetic data protect against identity disclosure because no real observations are released (Rubin 1993; Reiter, 2002; Kinney, Satkarta, Reiter, Reznek, Miranda, Jarmin, and Abowd, 2011; Hu, Reiter, and Wang 2014; Raab, Nowok, and Dibben 2017). To quote Hu, Reiter, and Wang (2014), "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 protect against attribute disclosure because no actual values are released (Reiter, 2002). In addition, synthesized values limit an intruder's confidence in any given value of a sensitive variable. For example, if an intruder identifies a set of records with identical values for a sensitive variable (a simple attribute attack), she still can't confirm if the value is the truth.

d. Potential disclosure risks in fully synthetic data

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, it is theoretically possible for a data synthesizer to 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 Census Bureau produced their own application of the database reconstruction theorem using the 2010 Census. Based on published tables, researchers at the Census Bureau recreated the unreleased swapped and unswapped microdata with about 50 percent accuracy. They then were able to correctly match a small fraction of the records in the recreated microdata to credit bureau data (Ruggles, 2018). This seems troubling, but there would be no way for an intruder to confirm if a match was correct or even if the reconstructed data were correct before the match.

Under certain conditions, many of the same techniques used to reconstruct non-synthetic data might be used to reconstruct administrative data from fully synthetic data. Indeed, researchers have identified non-trivial disclosure risks in fully synthetic data processes (Hu, Reiter, and Wang 2014).

To date, the only identified disclosure risks have been with respect to discrete variables and counts. Disclosure may be possible in the case of categorical variables that have a limited number of possible values, which means that they may be solved for with a finite set of simultaneous equations and a limited amount of information. Hu, Reiter, and Wang (2014) calculate re-identification risks on synthetic data in the American Community Survey. The authors were able to calculate 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) concluded that it was impractical to measure disclosure risk in the synthesized data from the UK Longitudinal Series: "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." [p. 82]

e. **Differential privacy**

The database reconstruction theorem motivated research into formal privacy guarantees for synthetic data, such as ϵ-differential privacy. Differential privacy is a definition that creates a formal disclosure guarantee for a given algorithm such as a count or sum (Dwork, 2008). Only ϵ-differential privacy guarantees protection against an intruder with full information about the data protection process, knowledge of ϵ, and knowledge of all but one row of the confidential data (Abowd and Vilhuber 2008). More formally, ϵ-differential privacy requires establishing that the log of the ratio of the probability that any single observation was in the data set that generated the output to the probability it was excluded is less than ϵ. Machanavajjhala, *et al.* (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 dataset] 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 U.S. 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 intruder has detailed data about all but one individual in the dataset. It would prohibit release of even very aggregate data; such as unaltered population means. A synthesis process that precisely reflected the distribution of the underlying tax data would also violate this standard since an intruder could replicate the synthesis process with all but one row of data and infer information about the missing row 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 dataset could protect data from most of the sample, but might not be effective for outlier observations.

ϵ-δ-probabilistic differential privacy, is an extension of ϵ-differential privacy. It guarantees that ϵ-differential privacy is met with probability $1-\delta$ (Machanavajjhala, Kifer, Abowd, Gehrke, and Vilhuber 2008). The probability that an intruder with full information about the data protection process, knowledge of ϵ, and knowledge of all but one row of the confidential data, could gain significant information about any individual's data is at most δ.

Several authors have attempted to develop fully synthetic datasets that satisfy differential privacy, but the data were of not of high quality. The Census Bureau's "OnTheMap" application was designed to achieve ϵ-δ-differential, but with limited data quality (Machanavajjhala, et al., 2008). Elliot (2014) created a measure of "empirical differential privacy." The measure makes assumptions about intruder knowledge and methods, so it does not satisfy ϵ-differential privacy. Kinney, et al., (2011) calculated *ex post* measures of privacy for individual variables in subgroups for the Synthetic Longitudinal Business Database (SynLBD). They confirmed that the SynLBD does not guarantee ϵ-differential privacy.

The usefulness and feasibility of differential privacy for microdata is currently under debate. McClure and Reiter (2012) found that the parameter ϵ is not closely related to the probability of disclosure. This is because differential privacy is based on algorithms and does not consider the specific values in a data set even though extreme values or uncommon combinations of values likely carry greater disclosure risk than common values or common combinations of values.

In a critique of the Census Bureau's use of differential privacy, 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."

f. The effects of sampling

Sampling limits 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; Reiter 2005b; Matthews and Harel, 2011). Skinner et al (1994) point out that: "Provided there is no measurement error, population uniqueness will be a sufficient condition for an exact match to be verified as correct." One of the great advantages 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.

g. **Our approach to protecting privacy**

We propose a synthesis methodology that protects against meaningful disclosure. We do not prove that our method satisfies differential privacy. Instead, we demonstrate that the synthetic data produced by the method protect 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 in the case where an intruder has extensive information about the underlying data.

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

- The administrative databases are very large. This means that a substantial amount of information may be released without allowing an intruder to infer anything useful about individuals unless they already possessed almost all the data. This alone does not meet the standard of differential privacy, where an intruder is assumed to possess all the records except one, but is a useful protection in more realistic scenarios where the intruder has incomplete information.

- We propose to generate synthetic observations by drawing from smoothed empirical distributions, rather than the actual data. This smoothing process significantly reduces the likelihood that a data point in the synthetic data will exactly match an actual observation unless there are many observations with similar characteristics in the underlying data.

- The synthetic dataset will have only a fraction (no more than 1 in 10) of the observations in the underlying administrative data. We show in Section 6 that this protects against meaningful disclosure about the idiosyncrasies of the underlying empirical distribution.

- Previous research has focused on the special problems created by outliers. 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 non-sensitive observations where the population density is high, and flattens the distribution in the tails to only reflect the general characteristics of the outlier observations. This protects against inference of even very sensitive observations.[4]

3. Data Utility

Utility 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 and 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 and the synthetic data (Snoke et al., 2018). In Section 9, we present specific measures of data quality and apply them to the synthetic nonfiler data.

4. Overview of Proposed Synthesis Methodology

CART is a collection of non-parametric models developed by Breiman, Friedman, Olshen, and Stone (1984) and brought to synthetic data by Reiter (2005a). CART creates a sequence of binary splits of the data that end in nodes that are intended to be homogenous and have predictive power. CART uses 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. Separately and 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. Optional: use cross-validation to reduce the full tree to avoid overfitting.

We estimate CART models for each variable with all previously estimated variables as predictors. To create a synthetic record, a gender is assigned randomly, based on the percentage distribution of records in the first level groups (female and males). For example, because 51 percent of Supplemental Public Use File records are female, the synthetic record had a 51% chance of being assigned a female gender, and a 49% chance of being assigned a male gender. Then an age is assigned randomly, taking into account the gender already randomly assigned and the distribution of ages for that gender. For example, if the gender already randomly assigned was female, and 10 percent of the females in the confidential data were between the ages of 25 and 30, there would be a 10 percent chance of assigning an age between 25 and 30 to the synthetic record. Other age groups would likewise have a random chance of being assigned, with all of the chances computed from the ages in the confidential data for females.

[4] This approach is consistent with the advice of Machanavajjhala, *et al.* (2008): “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.” (p. 285)

After assigning gender and age, wage income is assigned using the estimated CART model for wages, taking into account both the gender and age already assigned. The process for synthesizing tax variables like wages is somewhat different, because the tax variables are continuous (can take on any dollar value), unlike gender and age groups which are discrete (can take on only a limited number of specific values). For continuous variables, we draw from a smoothed version of the empirical distribution function with variances for each value determined by the sparseness of where they belong in the overall distribution.

5. Details of the Synthesis Procedure

The procedure involves drawing from a variable's smoothed empirical distribution and then deriving subsequent variables as a function of the previously synthesized values. The function could be estimated on the administrative tax data using parametric techniques such a multiple regression, but we have found that CART produces better results. For continuous variables—wages, interest, dividends, etc.—we use CART and draw from a Gaussian kernel density estimator fitted to the predicted values. This method smooths the distribution by filling in the gaps between observed values and is unbounded. Both smoothing and unbounding are desirable attributes as discussed below. For discrete values, such as age, we use CART without kernel density smoothing but top-code the synthesized value if necessary to suppress information about outliers. For example, age is capped at 85 in the synthetic Supplemental Public Use File data.

a. Overview of methodology

Our synthesis methodology is based on the insight that a joint multivariate probability distribution can be represented as the product of a sequence of conditional probability distributions.

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

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

b. Simulating discrete variables (X_1 and X_2)

The first variable (X_1) synthesized in the Supplemental Public Use File 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. It is split into groups to minimize the heterogeneity of values within groups. To measure heterogeneity, the algorithm in our synthesis uses a Gini index

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

where A is a node, C is the number of classes in the node (i.e. Male/Female), and p_i is the class probability for the i^{th} class (i.e. 0.65 are Female). So the best split minimizes

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

where N_L and N_R are the number of observations in the left and right nodes created by the split respectively, *N* is the 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 for a final node is reached (50).

c. Simulating continuous variables (*X*$_3$, *X*$_4$, …, *X*$_k$)

In CART, continuous variables are split and simulated using regression trees. The "best split" is defined by the split that minimizes the sum of squares in the two resulting nodes. This minimizes

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

where A_L and A_R are the left and right nodes created by the split respectively and $\overline{y_L}$ 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 synthesis samples values from the appropriate final node and then applies our smoothing method.

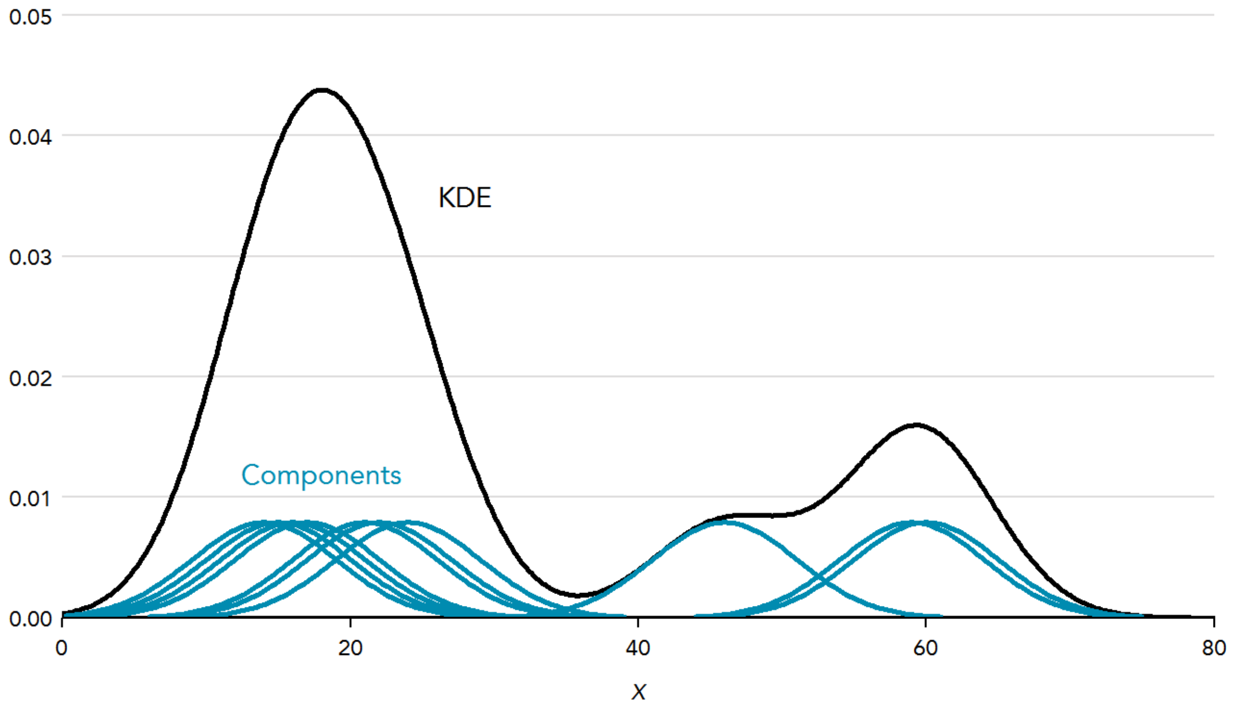
To simulate the first continuous variable (*X*$_3$**,** which is wages in the Supplemental Public Use File data), we create a smoothed kernel density function for each percentile of values predicted by CART for this variable.

As shown in Figure 1, the kernel density estimator is the aggregation of individual normal densities centered around each observation (Wicklin 2016). In the example, each of the individual Gaussian kernels has the same standard deviation. The kernel density distribution is smooth and unbounded.

We have to deal with some complications. First, the variance of the Gaussian kernel must grow with the tax variables. Otherwise, an intruder who knows how the database is constructed could draw some fairly precise inferences about outliers since any 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 discussed below this causes the variance to grow with the value of the synthesized variable.

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 Public Use File data.

FIGURE 1
Kernel Density Estimate as Weighted Sum of Component Densities



6. How the Synthesis Procedure Protects Privacy

The synthesis methodology draws values from a smoothed version of the empirical distribution function. As discussed above, the distribution is smoothed so that the probability of drawing any actual sample value is zero. However, there is a risk that the empirical distribution would be too close to the population distribution, revealing sensitive information about particular observations. In addition to smoothing, a key feature in our synthesis methodology is that we use only a fraction of the observations in the administrative dataset to generate the synthetic dataset.

a. The effects of sampling on inference about the underlying distribution

For the Supplemental Public Use File database, we start with a 10 in 9,999 (approximately 1 in 1,000) sample. That means that the odds are 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).

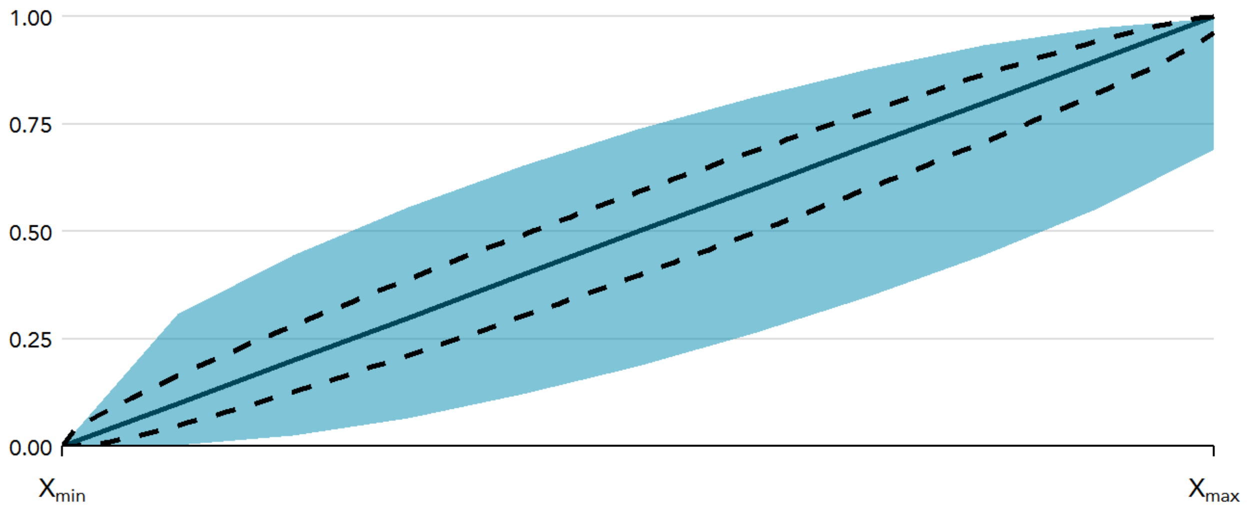
For the synthetic individual income tax return database, selecting a sample size that is at least an order of magnitude smaller than the underlying population obscures the nature of the underlying distribution. To illustrate, suppose the actual distribution of data in the administrative dataset is uniform within an interval that includes 100 records. The actual distribution is the solid line in Figure 2. An ideal synthesis would draw *n* independent observations from the uniform distribution within this interval. [5] An intruder might attempt to infer the underlying distribution by ranking the observations from smallest to largest and plotting the empirical distribution function. The intruder could glean little information about the underlying distribution from this plot, especially if *n* is much smaller than 100.

The probability that the k^{th} 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 k^{th} 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.

Figure 2 shows the underlying uniform distribution (solid line) and the 95 percent confidence interval for samples of size 10 and 100. The blue shaded area corresponds to the sample of size 10 and the dashed line represents the confidence interval for a sample of size 100. It is clear that many underlying distributions could be consistent with the sample distribution.

FIGURE 2

95 percent Interval Around Points Drawn from a Uniform Distribution Function with 1-in-10 Draw versus 1-in-1 draw (n = 100)



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

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

b. 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 an 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

$$\mu = \frac{x_m}{100} \tag{1}$$

and the variance is

$$\sigma^2 = \sum_1^{100} \frac{(x_i - \mu)^2}{100} \tag{2}$$

$$= \frac{99(-0.01x_m)^2 + (0.99x_m)^2}{100} = \frac{\big(0.99(0.01 + 0.99)\big)x_m^2}{100}$$

$$= \frac{0.99x_m^2}{100}$$

Just 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 $\sqrt{100\sigma^2/.99}$.

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:

$$E(\bar{x}) = \mu = \frac{x_m}{100} \tag{3}$$

$$Var(\bar{x}) = \frac{\sigma^2}{10} \tag{4}$$

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

$$se(\bar{x}) = \frac{\sigma}{\sqrt{10}} = \frac{\sqrt{0.99}x_m}{10\sqrt{10}} = 0.0315\, x_m \tag{5}$$

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

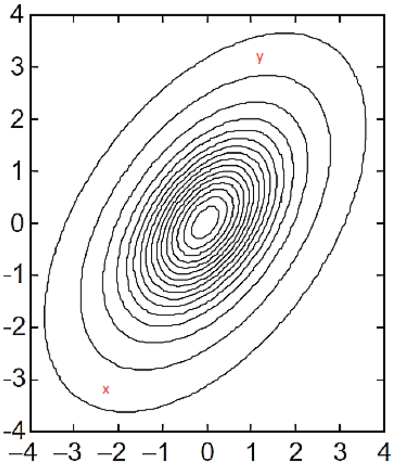
In the other extreme case where all the values are approximately the same, the simulated values will be very close to the outlier values, but there is no disclosure because those values are not unique.

c. Attribute Disclosure

There are two types of tax return attribute disclosures of concern. One is revealing information about particular taxpayers based on unique combinations of attributes. The second is simply revealing that a person has filed a tax return, which the IRS treats as an impermissible disclosure.

The synthesis methodology described here protects against both types of disclosure. As shown above, the synthesis will prevent an intruder from inferring any particular values on any individual's tax return, even if the intruder possesses extensive information about the taxpayer's other attributes. The probability of inferring rare combinations of attributes will be even smaller. For example, in a bivariate distribution, rare pairings are in the relatively flat part of the distribution along with an enormous number of other equally improbable combinations that do not occur in the original dataset. Thus, observing a point such as *x* on Figure 3 below (contour lines for a bivariate normal distribution with ρ = 0.5) tells us virtually nothing about whether a point like *x* exists in the original data. And the figure vastly understates the sparseness of the distribution in the tails. There is more of a disclosure risk 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



Notes: Simulated data of normal random variables 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. If the IRS revealed that a person had not filed an income tax return, this would reveal information about that person's income or evidence that the person might have violated the law, both of which would be meaningful disclosures that could potentially harm the individual. However, it would be impossible to 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.

Synthesis should also make it highly unlikely that an intruder will infer that a particular person filed a return. The highest risk of identification in a non-synthetic sample would be for extreme outliers—i.e., those with extremely high incomes or with rare combinations of attributes. As noted above, the synthesis methodology effectively addresses this source of disclosure risk.

The synthetic Supplemental Public Use File database, described in the next section, is constructed from a 1-in-1000 sample. Applying the same attribute disclosure reasoning, it would be impossible to infer from it that a particular individual was in the Supplemental Public Use File dataset.

7. Synthesizing the Low Income Supplement Sample Data

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

We began with a definition of nonfilers from Cilke (2014), "Any U.S. resident that does not appear on a Federal income tax return filed [for] a given year." We were not interested in people who are required to file but do not, so we excluded those with incomes above twice the filing threshold for married couples filing jointly (which varies depending on whether either or both spouses are age 65 and older). Our sample is thus comprised of people who don't file a Federal income tax return for a given year and do not appear to have an income tax filing requirement.[6]

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 is comprised of individuals whose SSN (Social Security number) or ITIN (individual taxpayer identification number for those without SSN) ends in one of ten four-digit combinations. The last four digits are randomly assigned at birth and range from 0001 to 9999. Thus, the sample is a 10 in 9,999 (or approximately one in 1,000) random sample.[7]

[6] Some self-employed people may not owe income tax but still be required to file a 1040 because they owe SECA payroll taxes (Langetieg, Payne and Plumley, 2017). We retain those people in the sample.

[7] The sample is called the Continuous Work History Sample (CWHS) and has been maintained by the IRS for many decades, although some of the ten digits were not selected in earlier years of the panel. The last four digits of SSNs and of ITINs are randomly assigned, but 0000 is never assigned. Thus, there are only 9,999 possible four-digit endings.

TABLE 1
Percentage of Nonfilers with Specific Information Return Types (2010)

Type	Description	Percent of nonfilers with one or more Information Return Types	Percent of nonfilers with only one Information Return Type
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

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

We deleted records for those who appear in the information return dataset 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, we ended up with an administrative dataset with about 26,000 observations.

We synthesized the data using a customized version of CART from the R package synthpop (Raab et al. 2019). synthpop contains multiple methods for creating partially-synthetic and fully-synthetic datasets and for evaluating the utility of synthetic data. It does not include any tools for evaluating the confidentiality or privacy of a synthetic data set.

We use CART to partition the sample into relatively homogeneous groups, subject to the constraint that none of the partitions be too small, to protect against overfitting (Benedetto, e*t al.,* 2013). In testing on the Supplemental Public Use File 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 Public Use File dataset, we start with the administrative dataset with 26,000 observations described above. First, we split the data into two parts. The first part is observations from the confidential data that have zeros for all seventeen tax variables.[8] The second part is observations with at least one non-zero tax variable. For the sample with all zeros for tax variables, we randomly assign gender based on the proportions in the zero subsample (see below), synthesize age based on gender, and finally assign zeros to all tax variables.

For the sample with at least one non-zero value for a tax variable, we choose gender—a binary variable—as X_1. We do not synthesize gender, but randomly select gender based on the underlying share in the confidential dataset. With 51 percent female and 49 percent male in the administrative dataset, the assigned gender for each row in the synthetic dataset will have a 51 percent probability of female and a 49 percent probability of male. Due to the random assignment of gender, the distribution of gender in the synthetic dataset 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 ages (X_2) to each record conditional on gender. Since 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 the differences are likely to be small given the sample size. Age is top coded at 85 after synthesis.[9]

For continuous variables, we start with the variable with the most non-zero values—wage income (X_3), and then order the remaining variables, ($X_4, X_5, \ldots, X_{19}$), in terms of their correlations with wage, from most highly to least correlated.[10] CART partitions the data into relatively homogeneous wage groups within each gender/age group, and randomly selects a wage value from all the values in that wage group. For all non-zero values, we replace values with a random value drawn from a normal distribution with a mean equal to the observation being replaced and variance equal to the optimal variance from a kernel density estimator estimated on the corresponding percentile of the distribution of the variable being replaced. This is a computationally efficient way to approximate a kernel density estimator and has the desirable feature that the error is much larger for the sparse

[8] Note that this peculiarity is limited to the information return dataset of nonfilers, where a sizable percentage 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 non-zero value—otherwise there is no reason to file a tax return.

[9] Based on Census data, the age 85 cut-off groups together about two percent of the adult population (three percent of females and one percent of males). The percentages are probably a bit higher for nonfilers since people whose income comes mostly or entirely from Social Security generally do not have a filing requirement.

[10] Ordering from the variable with the most non-zero observations to the variable with the fewest non-zero observations is the norm for creating synthetic data, but we found that the correlation-order with wage worked better in preliminary tests.

parts of the distribution than dense parts of the distribution. It worked well for the Supplemental Public Use File database.[11]

Further, the CART algorithm mixes up values across observations. This means, even without smoothing empirical distributions, uncommon combinations of zeros and non-zeros within a synthesized record may be an artifact of the synthesizer and not an attribute of the underlying confidential data.

No noise is added to values of 0, which make up the majority of values for all continuous variables in the Supplemental Public Use File data. We don't consider zeros to be a disclosure risk because the variable with the most non-zero values is 73 percent zeros. Many of the variables are zero for almost every record. By default, synthpop does not smooth values if the frequency of a single value exceeds 70 percent.

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 over-fit data, which can increase out-of-sample prediction error. So, most trees are reduced based on a penalty for the number of final nodes in the tree (Kuhn and Johnson, 2016). We only reduce our trees in extreme cases because our minimum bucket size is large (50) and the default parameter for penalizing complex trees in synthpop is small.

8. Measures of How Well the Synthetic Supplemental Public Use File Data Protects Privacy

Our methodology is designed to protect confidentiality *ex ante*. However, we also use a set of 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 synthesis, by adjusting smoothing methods and parameters like the minimum size of the final nodes in the CART synthesizer.

We focus on three different types of metrics. The first type counts the number of unique donors to each row in the synthetic data. The second type examines the frequency and uniqueness of synthesized rows in the confidential data. The third type applies a formal privacy framework called ℓ-diversity to the CART synthesizer.

a. Unique donors

[11] The nonfiler database is relatively homogenous so percentile smoothing works well. The file of individual income tax returns has some records with extremely large values. We expect to use a standard kernel density estimator that draws from the distribution of records in the neighborhood of each value. This will automatically add enough noise to make it impossible to distinguish one outlier from another (or to determine if any outlier observation is in the administrative dataset), as described above.

Before smoothing, the CART synthesizer combines actual values for tax variables from many different rows in the confidential data into a single row in the synthetic data. This process of shuffling up real values (before smoothing) from many records into a new synthesized record is an important part of our method. It's easy to imagine a synthesizing method that is too precise and creates synthetic rows from a few rows or one row in the confidential data–though it's difficult to imagine implementing such a precise synthesizer. As a check, we trace the number of unique rows in the confidential data that "donate" values to the synthetic data. This definition is a little loose because nothing is donated in our method; rather, we trace every observation with a unique identifier through the synthesis process to ensure that the data are adequately shuffled.

We found the minimum number of unique donors to each observation of the synthetic dataset. There are 19 variables in the Supplemental Public Use File dataset (17 tax, 2 demographic). The maximum number of possible "donors" to any given row of the synthetic data is thus 19. The minimum number of unique donors in the Supplemental Public Use File dataset is 15.

b. Duplicates

We examined several metrics of the frequency and uniqueness of synthesized rows in the confidential data. As we showed above, no rows are actually duplicated. Instead the rows appear duplicated because the combination of synthesized variables from the synthesizer combine into a row that has the same values as a row in the confidential data.

The simplest metric 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 frequent rows that occur in the confidential data that would be expected to appear as replicated in the confidential data by chance.

c. 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.

d. Row-wise Squared Inverse Frequency

Finally, we relax the uniqueness requirement and instead use 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. It then squares the inverse of this metric so that rows that appear once are assigned a value of 1, rows that appear twice are assigned a value of ¼, rows that appear three times are assigned a value of 1/9, and so on. With all of these measures, it is important to remember that the confidential data being synthesized come from a 10/9,999 sample of tax records. So even if low probability rows are included in the data, it is unlikely that they are unique in the population.

The results for duplicates, unique-uniques, and row-wise inverse frequency were all very small and thus are not reported below.

e. ℓ-diversity of final nodes in the CART algorithm

We were concerned that the CART algorithm could generate final nodes that lack adequate heterogeneity. Too little heterogeneity in the final nodes could result in 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 non-sensitive 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 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 percent 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.

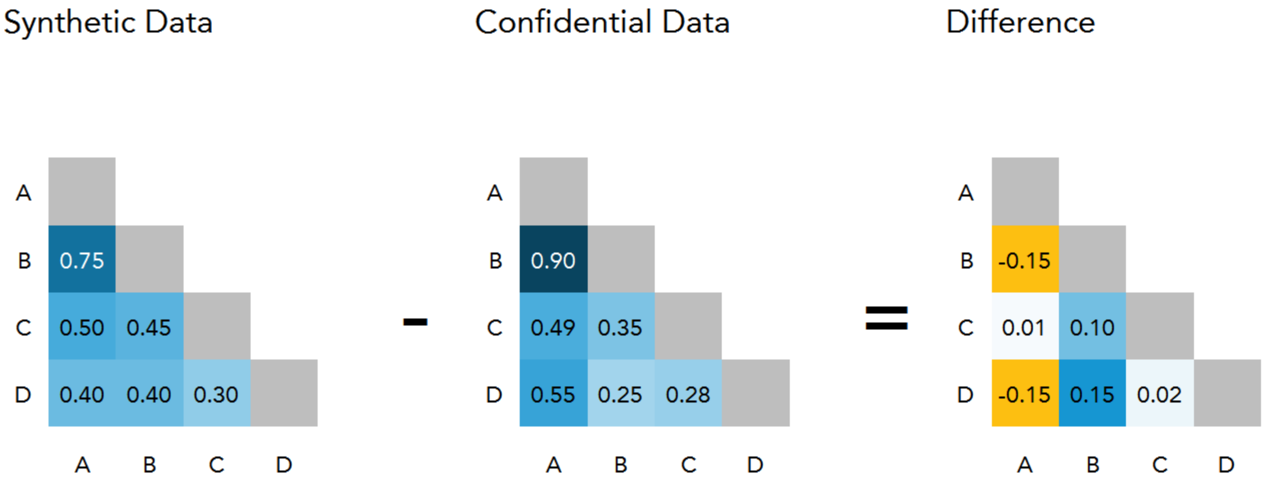
9. Measures of the Quality of the Synthetic Supplemental Public Use File Data

a. General utility metrics

Correlation fit measures how well the synthesizer recreates the linear relationships between variables in the confidential dataset. Correlation fit is the lower triangle of a Pearson's linear correlation matrix from the synthetic data minus the lower triangle of a Pearson's linear correlation matrix 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 matrices from the synthetic and confidential data sets.

First, we rank the differences between each pair of variable from smallest to largest. Variable pairs with large distances 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 synthesis-wide number that measures how well the synthesis is capturing linear relationships.

FIGURE 4
Example Calculation of Correlation Fit



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.

$$Correlation\ Fit = \frac{\sqrt{\sum_{i=2}^{n}\sum_{j=1}^{i}\left(S_{ij} - O_{ij}\right)^2}}{\binom{n}{2}}$$

The *Kolmogorov-Smirnov test* is a nonparametric test of the equivalence of univariate probability distributions. For synthetic data, the Kolmogorov-Smirnov 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.

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

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

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

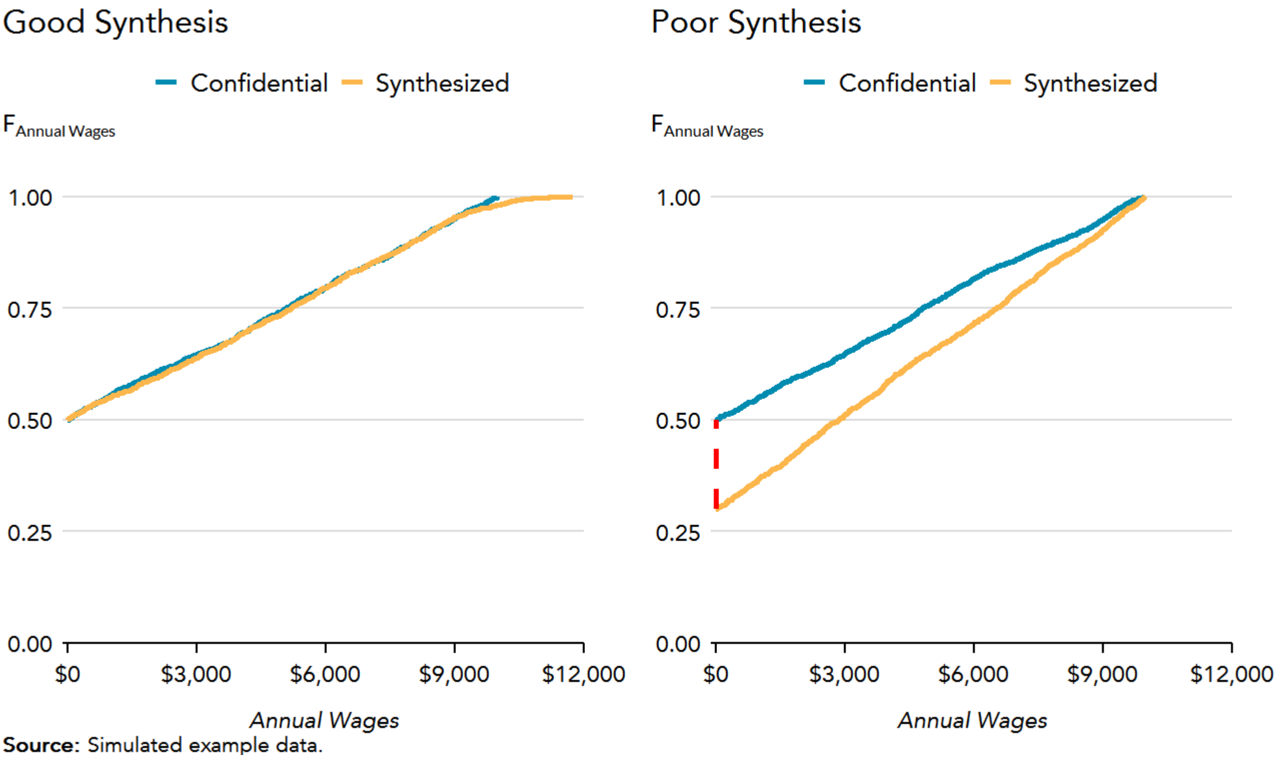
$$F_{m,2} = \frac{1}{m}\sum_{i=1}^{m} I_{(-\infty,x_i](X_i)}$$

The Kolmogorov Smirnov statistic for the above samples and ECDFs is

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

This Kolmogorov Smirnov test essentially finds the largest absolute vertical distance between the two ECDFs and measures 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

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

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 that estimates whether a model can distinguish between the confidential and the synthetic data. Woo et al. (2009) introduced and Snoke et al. (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 data set 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 data set belongs to the synthetic data set. The propensity scores are modeled with a logistic regression. The predictors are all variables in the combined data without interactions. Interactions up to a specified maximum order of interactions are possible, but estimation struggles to converge. Alternatively, a CART model can be used to estimate the propensity scores.
3) Calculate the probability expected if the data did not distinguish the synthetic data from the original data. The probability expected is the proportion of synthetic data in the combined data. In most cases this will be 0.5 because the confidential data set 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. Let $\widehat{p_i}$ be the estimated propensities. Let p_0 be the probability expected (typically 0.5).

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

We focus 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 dataset. Let n_2 be the number of observations in the synthetic dataset. Let $N = n_1 + n_2$.

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

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.

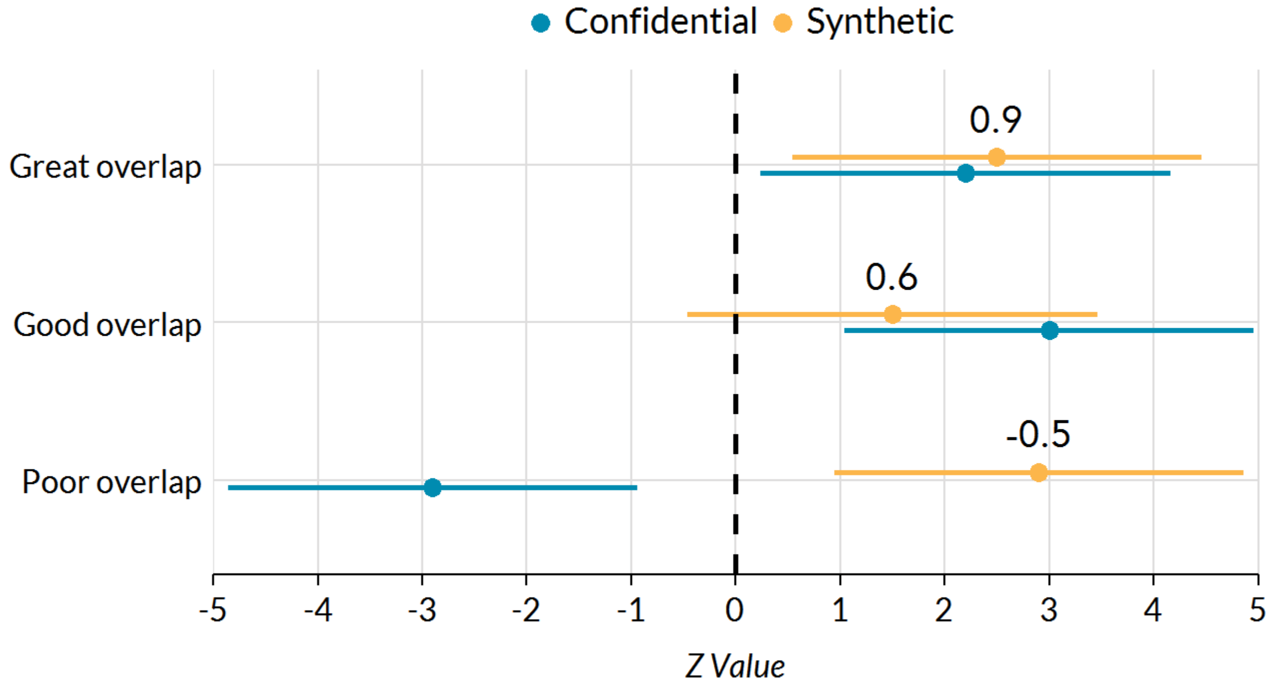
b. Specific utility metrics

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 “o” and “s” denote the confidence interval bounds for the original and synthetic data:

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

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 don't 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 Public Use File dataset will be used for tax microsimulation. We built a tax calculator to compare calculations of AGI, 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 turns off the standard deduction, and it lowers the personal exemption to $500. This unorthodox combination of rules is necessary to get useful calculations using the Supplemental Public Use File data, which come from a population that pays federal income tax only through withholding by payers of wages and other income.

We will now focus on the utility of the synthetic Supplemental Public Use File data to illustrate.

TABLE 2
Count of Genders by Data Source

Gender	Original	Synthetic
Female	13,669	13,567
Male	13,274	13,263

TABLE 3
Count of Age Groups by Data Source

Age Group	Original	Synthetic
1-17	856	815
18-24	2,299	2,296
25-34	2,811	2,734
35-54	6,136	6,143
55-64	3,956	3,946
65+	10,885	10,896

TABLE 4
Count of Age Groups by Gender and Data Source

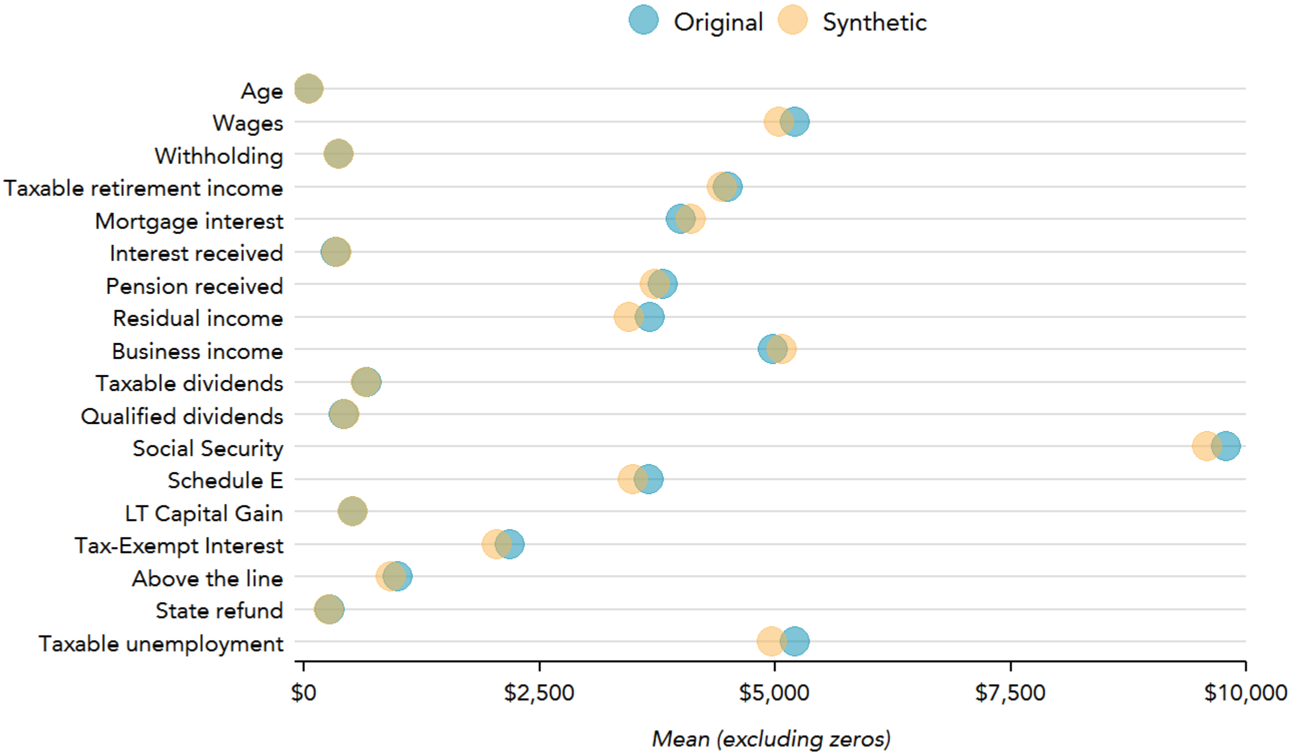
Age Group	Gender	Original	Synthetic
1-17	Female	414	369
1-17	Male	442	446
18-24	Female	1,030	1,002
18-24	Male	1,269	1,294
25-34	Female	1,093	1,036
25-34	Male	1,718	1,698
35-54	Female	2,514	2,571
35-54	Male	3,622	3,572
55-64	Female	1,913	1,897
55-64	Male	2,043	2,049
65+	Female	6,705	6,692
65+	Male	4,180	4,204

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

c. Summary statistics

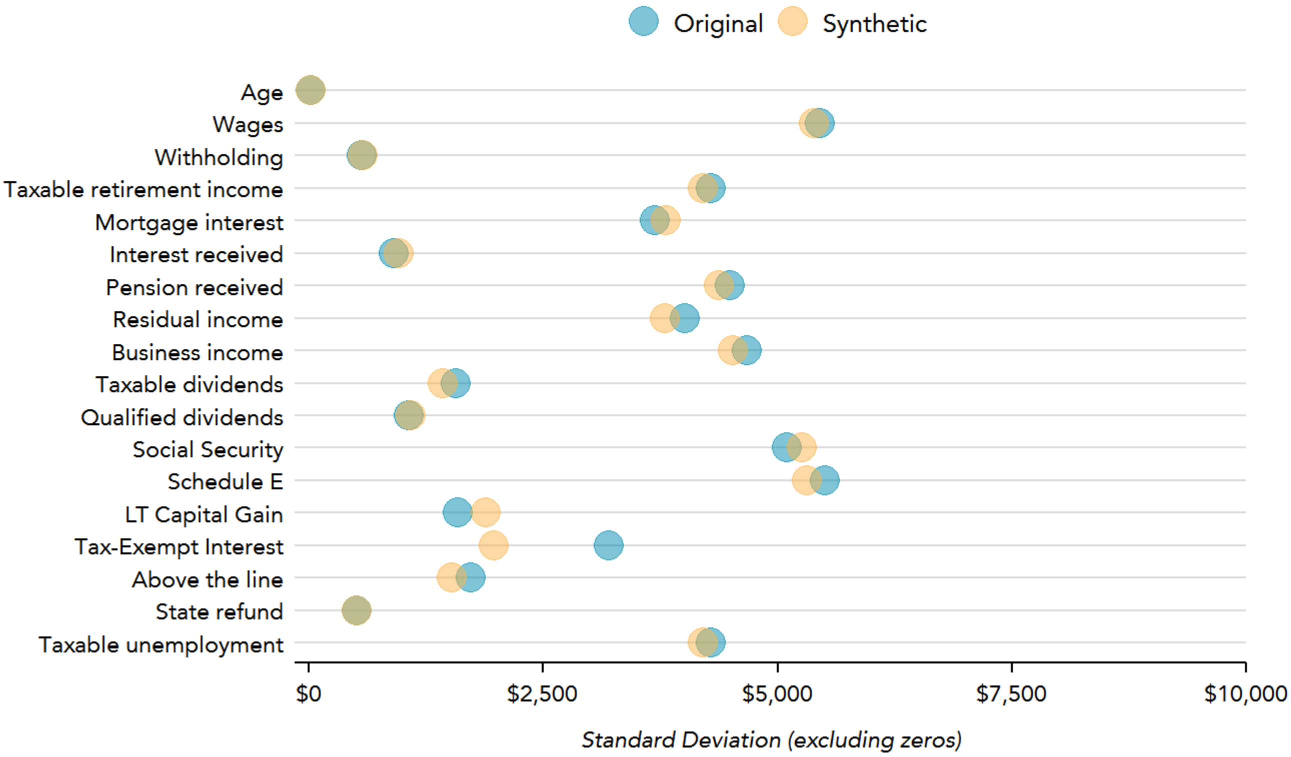
The synthesis recreates the univariate distribution of the tax variables. Figures 7, 8, 9, and 10 respectively compare the mean, standard deviation, skewness, and kurtosis of the tax variables in the synthetic data set with the tax variables in the confidential data set. The four figures exclude any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

FIGURE 7
Means from Original and Synthetic Data



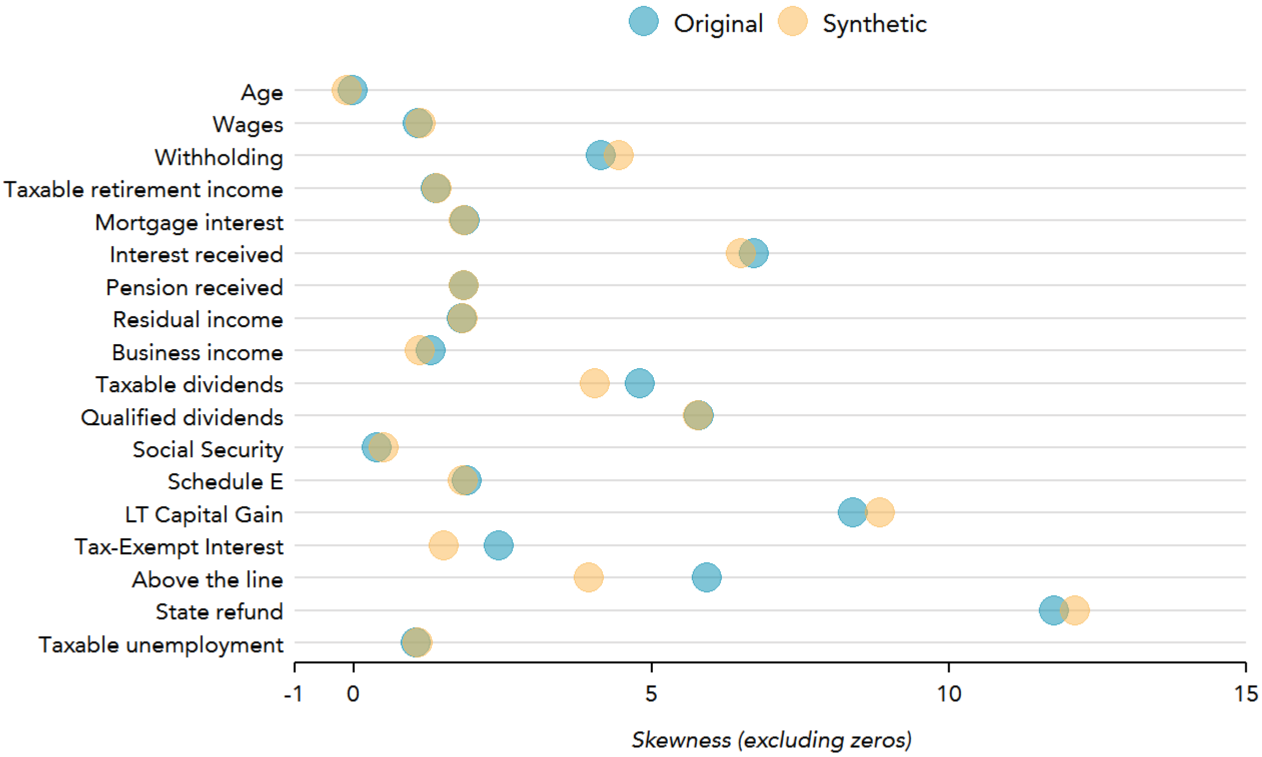
Note: Calculations exclude all zeros.

FIGURE 8
Standard Deviations from Original and Synthetic Data



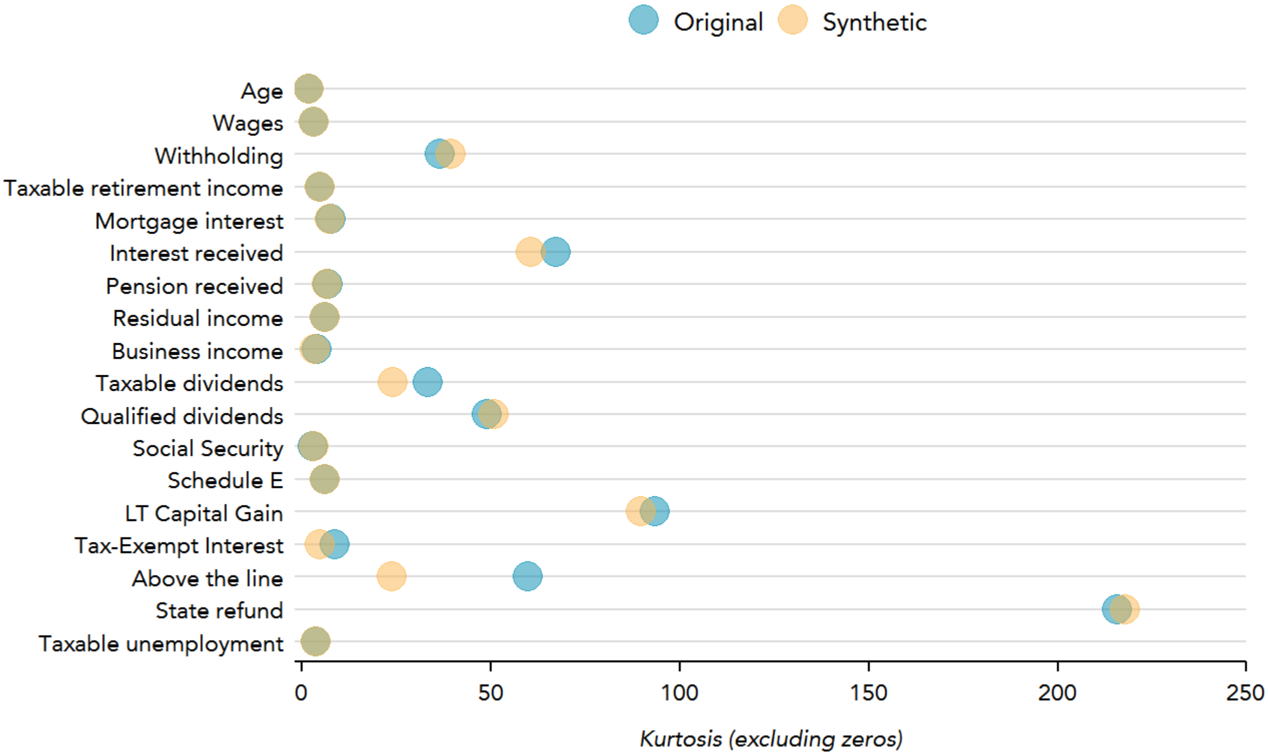
Note: Calculations exclude all zeros.

FIGURE 9
Skewness from Original and Synthetic Data



Note: Calculations exclude all zeros.

FIGURE 10
Kurtosis from Original and Synthetic Data

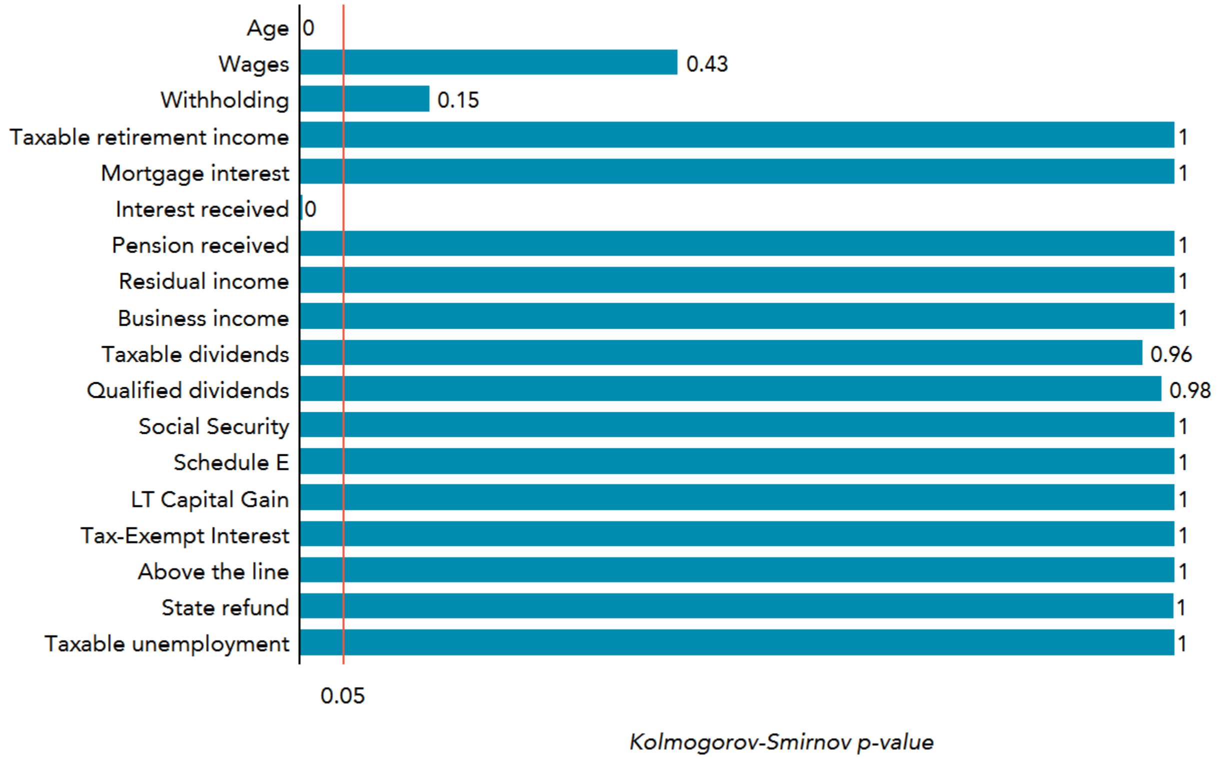


Note: Calculations exclude all zeros.

d. CDF analysis

The Kolmogorov-Smirnov tests and zero coverage also suggest that we do a good job recreating the univariate distributions of the tax variables. Age failed the Kolmogorov-Smirnov test because of top coding but passes without top coding. Interest received fails the Kolmogorov-Smirnov test because of rounding but passes without rounding. No other variables failed the Kolmogorov-Smirnov test and as figure 11 shows, none of the p-values is near common cutoffs of 0.01, 0.05, or 0.1.

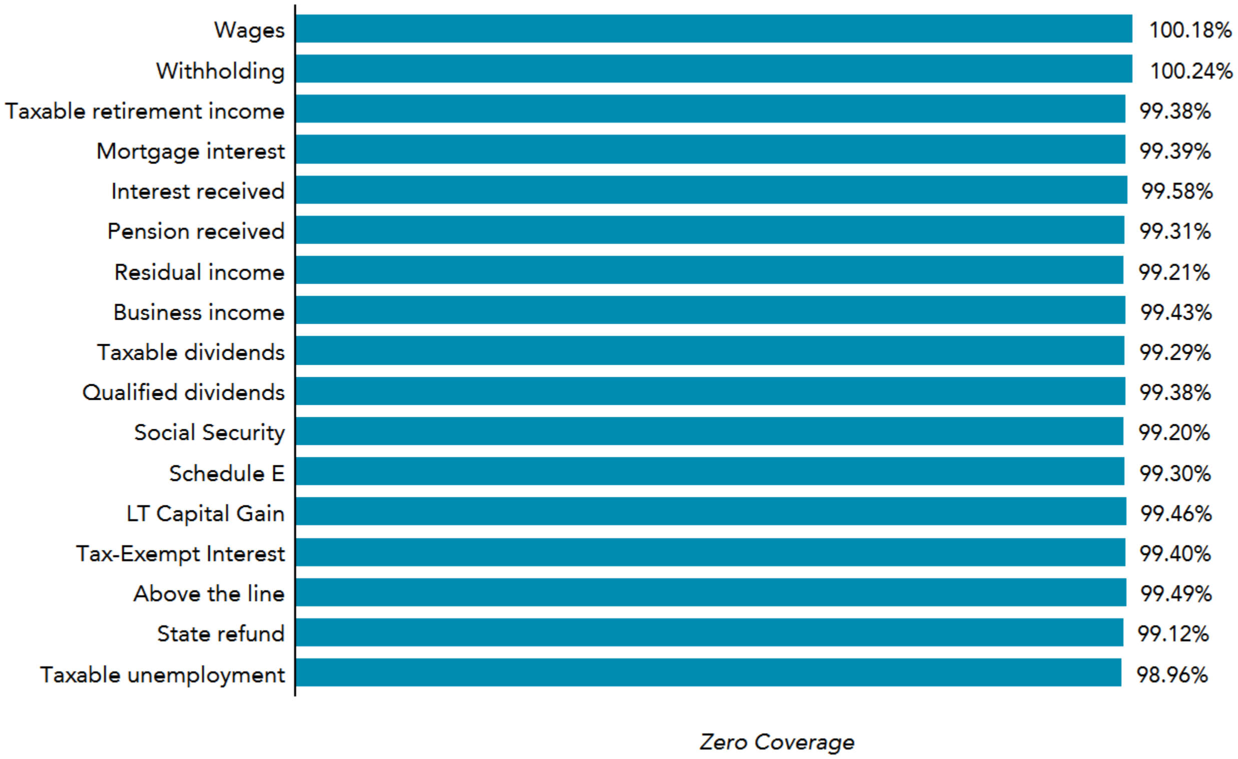
FIGURE 11
P-Values from Two-Sample Kolmogorov-Smirnov Tests on Original and Synthetic Data



Note: Calculations exclude rows with zeros for all seventeen tax variables.

The synthesizer also did a good job recreating the share of zero values. As figure 11 shows, all variables are within about 1 percent of the correct number of zeros.

FIGURE 12
Percentage of Values that are Zeros in the Synthetic Data Relative to the Original Data

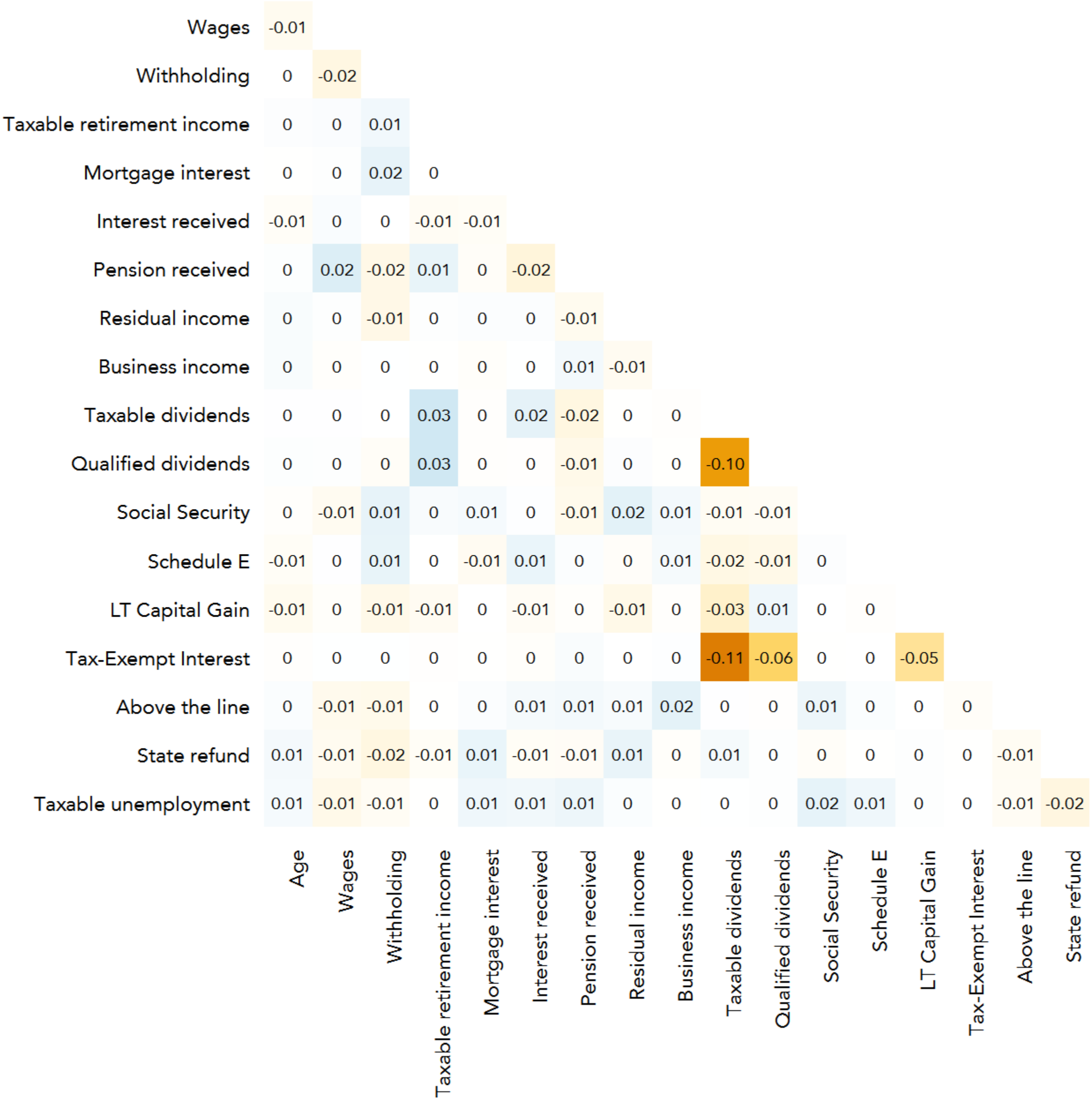


Note: Calculations exclude rows with zeros for all seventeen tax variables.

e. Correlation fit

The synthesizer does a good job recreating the linear relationships between variables. Overall, the correlation fit was 0.0013. Figure 13 shows 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 since these variables have very few non-zero 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 13
Correlation Differences (Synthetic minus Original)



	Age	Wages	Withholding	Taxable retirement income	Mortgage interest	Interest received	Pension received	Residual income	Business income	Taxable dividends	Qualified dividends	Social Security	Schedule E	LT Capital Gain	Tax-Exempt Interest	Above the line	State refund
Wages	-0.01																
Withholding	0	-0.02															
Taxable retirement income	0	0	0.01														
Mortgage interest	0	0	0.02	0													
Interest received	-0.01	0	0	-0.01	-0.01												
Pension received	0	0.02	-0.02	0.01	0	-0.02											
Residual income	0	0	-0.01	0	0	0	-0.01										
Business income	0	0	0	0	0	0	0.01	-0.01									
Taxable dividends	0	0	0	0.03	0	0.02	-0.02	0	0								
Qualified dividends	0	0	0	0.03	0	0	-0.01	0	0	-0.10							
Social Security	0	-0.01	0.01	0	0.01	0	-0.01	0.02	0.01	-0.01	-0.01						
Schedule E	-0.01	0	0.01	0	-0.01	0.01	0	0	0.01	-0.02	-0.01	0					
LT Capital Gain	-0.01	0	-0.01	-0.01	0	-0.01	0	-0.01	0	-0.03	0.01	0	0				
Tax-Exempt Interest	0	0	0	0	0	0	0	0	0	-0.11	-0.06	0	0	-0.05			
Above the line	0	-0.01	-0.01	0	0	0.01	0.01	0.01	0.02	0	0	0.01	0	0	0		
State refund	0.01	-0.01	-0.02	-0.01	0.01	-0.01	-0.01	0.01	0	0.01	0	0	0	0	0	-0.01	
Taxable unemployment	0.01	-0.01	-0.01	0	0.01	0.01	0.01	0	0	0	0	0.02	0.01	0	0	-0.01	-0.02

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

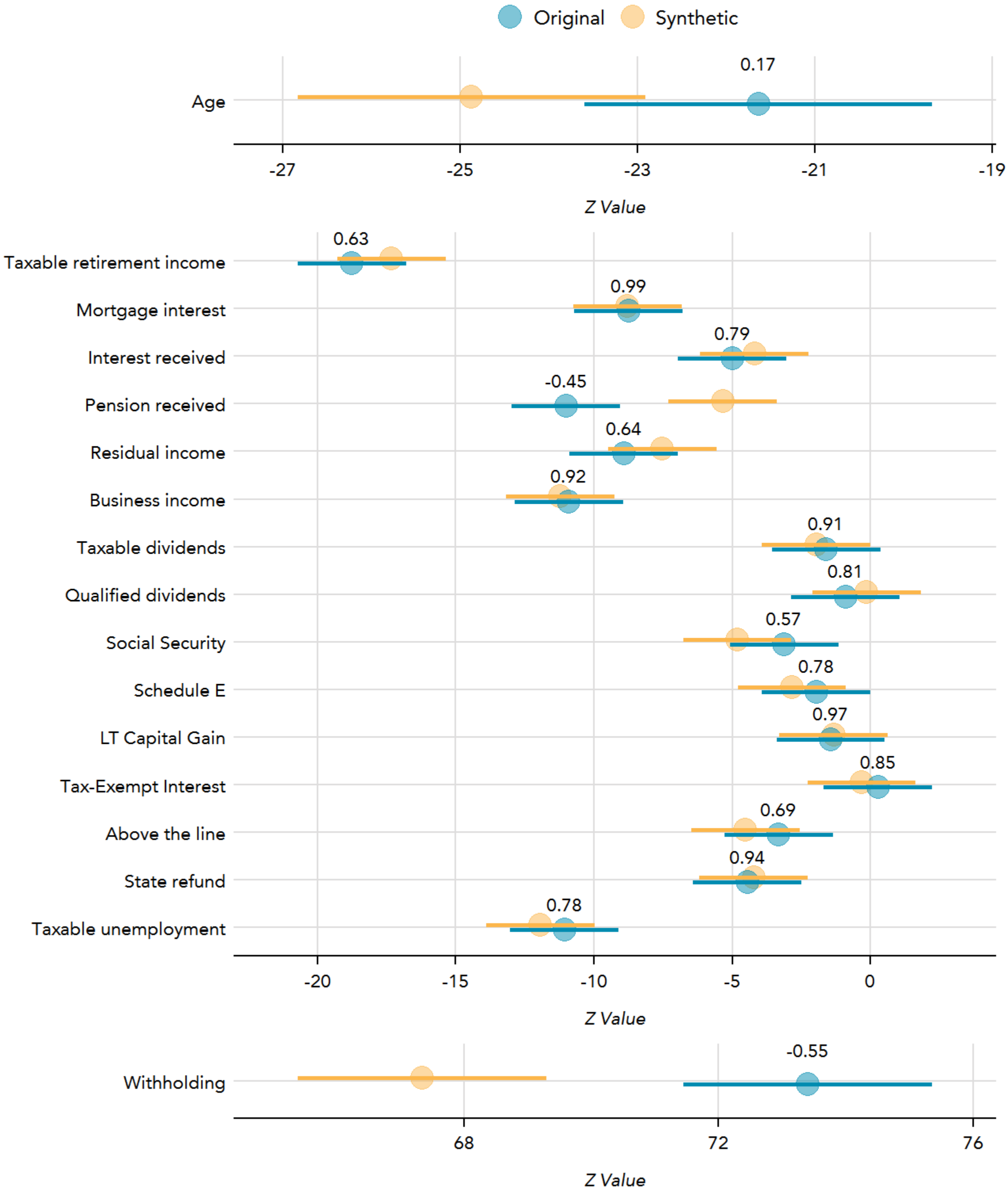
f. pMSE

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 it is difficult to distinguish between the confidential and synthetic data. This all suggests high general utility.

g. Confidence interval overlap

The synthesizer performs adequately for our measure of regression confidence interval overlap. Figure 14 compares the coefficient estimates and error bars 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 estimates with negative overlaps are at least directionally correct and are comparable in magnitude.

FIGURE 14
Regression Confidence Interval Overlap

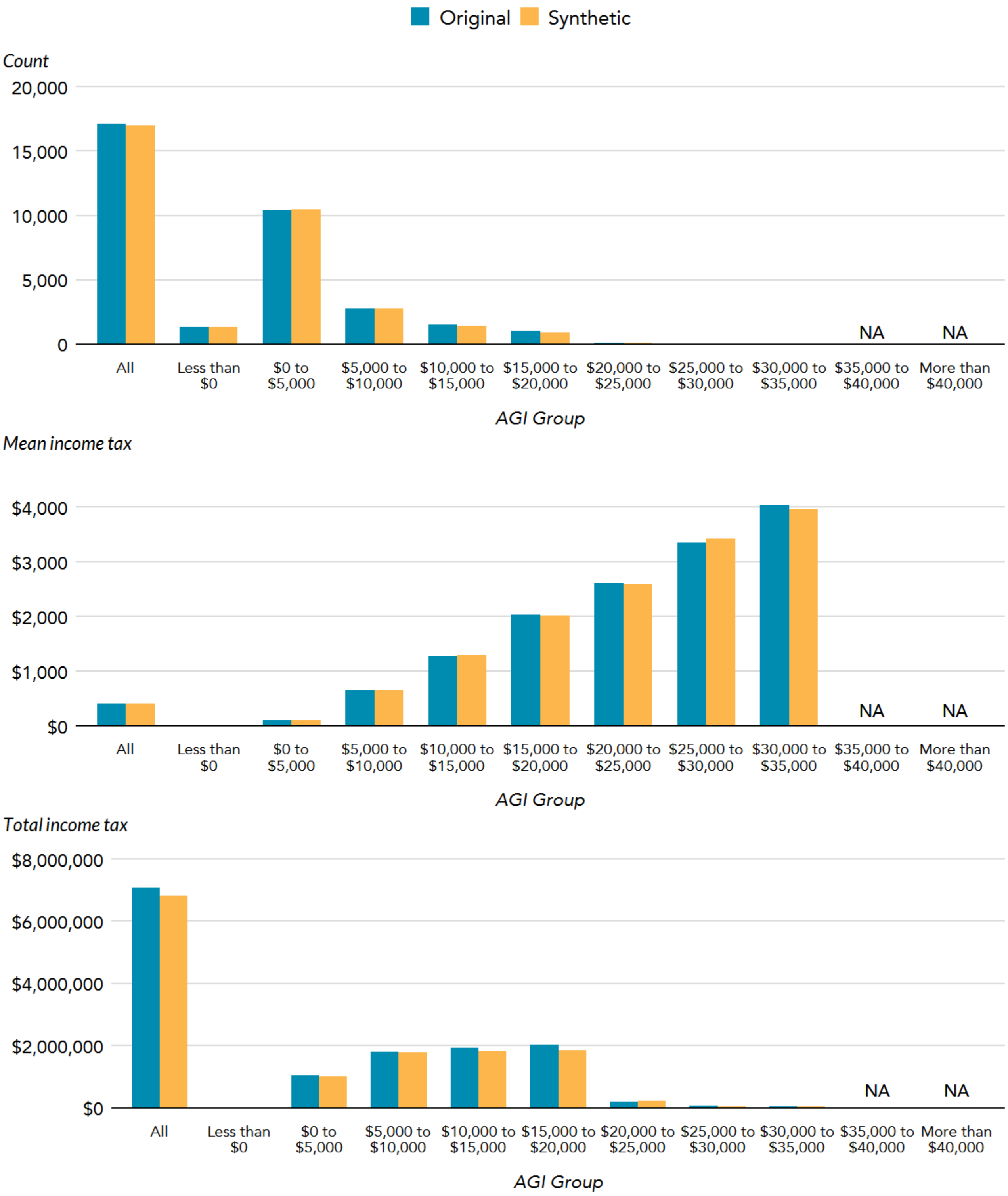


Note: Calculation excludes rows with zeros for all seventeen tax variables.

h. Tax calculator

The main use of the synthetic Supplemental Public Use File dataset will be tax microsimulation. The synthetic file performs well in our simple tax calculator and approximates the results from the confidential data set. Figure 14 compares results for the original and synthetic data sets across different Adjusted Gross Income (AGI) groups for count, mean tax, and total tax.

FIGURE 15
Tax Calculator Results for the Original and Synthetic Data



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

10. Conclusions and Planned Future Work

This paper outlines a method to create a fully synthetic database that would not allow an 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 or had not filed a tax return beyond statistical relationships between variables. We have implemented this for the Supplemental Public Use File database and found that the resulting synthetic dataset replicates the characteristics of the underlying administrative data while protecting individual information from disclosure.

Our next step is to create a synthetic dataset of the much more complex and diverse individual income tax return data. We do not know, a priori, how well the synthesis methodology used for the Supplemental Public Use File data will replicate the underlying distributions of these data. We plan to test a range of synthesis methods, including random forests (which performed less well than CART for the Supplemental Public Use File data, but could outperform CART for the individual income tax return data). At a minimum, our goal is to find a method that will create a synthetic file that protects the privacy of individuals and reproduces the conditional means and variances of the administrative data. We hope it will be useful for estimating the revenue and distributional effects of tax law changes, and also be useful for exploratory statistical analysis.

Perhaps most importantly, because the synthetic dataset will have the same structure as the underlying administrative data, it will serve a valuable purpose as a "training dataset" that researchers could use to develop and debug complex statistical programs in R or Stata. If we are successful in establishing a validation server, then researchers will be able to submit their programs to run on a subset of the restricted data. This could significantly expand research access to a vital information resource.

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, 5262:239–46. Berlin, Heidelberg: 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.

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

Bryant, Victoria L. 2017. "General Description Booklet For the 2012 Public Use Tax File." Individual Statistics Branch Statistics of Income Division Internal Revenue Service. 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." Tax Policy Center.

Cilke, James. 2014. "The Case of the Missing Strangers: What We Know and Don't Know About Non-Filers." 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 - PODS 202–10. San Diego, California: ACM Press, 2003. 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, no. 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, no. 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, 4978:1–19. Berlin, Heidelberg: 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, no. 337: 7–18.

Fienberg, Stephen E, and Jiashun Jin. 2009. “Statistical Disclosure Limitation for Data Access.” Encyclopedia of Database Systems.

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, no. 1: 75–89.

Fuller, Wayne A. 1993. “Masking Procedures for Microdata Disclosure Limitation.” Journal of Official Statistics 9, no. 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, 8744:185–99. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-11257-2_15.

Internal Revenue Service. 2019. “2012 Supplemental Public Use File.”

Karr, A. F, C. N Kohnen, A Oganian, J. P Reiter, and A. P Sanil. 2006. “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality.” The American Statistician 60, no. 3: 224–32. https://doi.org/10.1198/000313006X124640.

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.” https://www.census.gov/ces/pdf/CES-WP-11-04.pdf.

Kuhn, Max, and Kjell Johnson. 2016. Applied Predictive Modeling. Springer.

Langetieg, Patrick, Mark Payne, and Alan Plumley. 2017. “Counting Elusive Nonfilers Using IRS Rather Than Census Data.” IRS Research Bulletin (Publication 1500).

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, A, D Kifer, and J Gehrke. 2006. “L-Diversity: Privacy beyond k-Anonymity.” 22nd International Conference on Data Engineering (ICDE’06): 1–47.

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, no. 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, 4302:177–88. Berlin, Heidelberg: 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, D.C.: Congressional Budget Office. https://www.cbo.gov/publication/53125.

National Research Council, 1993. *Private lives and public policies: Confidentiality and accessibility of government statistics*. 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, no. 3: 67–97.

Reiter, Jerome P. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." Journal of Official Statistics 18, no. 4: 531–43.

———. 2005a. "Using CART to Generate Partially Synthetic Public Use Microdata," Journal of Official Statistics 21, no. 3: 441-462.

———. 2005b. "Estimating Risks of Identification Disclosure in Microdata." Journal of the American Statistical Association 100, no. 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, no. 1. 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, no. 1: 41–55.

Rubin, Donald B. 1993. "Discussion: Statistical Disclosure Limitation." Journal of Official Statistics 9, no. 2: 461–68.

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

Skinner, Chris, Catherine Marsh, Stan Openshaw, and Colin Wymer. 1994. "Disclosure Control for Census Microdata." Journal of Official Statistics 10, no. 1: 31–51.

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.

Sweeney, Latanya. 2002. "K-Anonymity: A Model for Protecting Privacy." International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 5: 557–70.

Templ, Matthias, Bernhard Meindl, and Alexander Kowarik. 2019. "Introduction to Statistical Disclosure Control (SDC)." Comprehensive R Archive Network. https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf

Therneau, Terry M., and Elizabeth J. Atkinson. 2018. "An Introduction to Recursive Partitioning Using the RPART Routines." Comprehensive R Archive Network. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Vilhuber, L. and Abowd, J. M. 2016. Usage and outcomes of the synthetic data server. Presentation at society of labor economics meetings, Cornell University, Labor Dynamics Institute

Wicklin, Rickj. 2016. "How to Visualize a Kernel Density Estimate." The Do Loop. https://blogs.sas.com/content/iml/2016/07/27/visualize-kernel-density-estimate.html

Winkler, William E. "Examples of Easy-to-Implement, Widely Used Methods of Masking for Which Analytic Properties Are Not Justified." Research Report Series. Washington, DC: U.S. Census Bureau, n.d. 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, no. 1. 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." SSRN Electronic Journal. 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, 2316:135–52. Berlin, Heidelberg: Springer Berlin Heidelberg, https://doi.org/10.1007/3-540-47804-3_11.