Forum

SYNTHETIC INDIVIDUAL INCOME TAX DATA: PROMISES AND CHALLENGES

Claire McKay Bowen, Victoria L. Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Livia Mucciolo, Madeline Pickens, and Aaron R. Williams

Tax data are invaluable for research, but privacy concerns severely limit access. Although the US Internal Revenue Service produces a public-use file (PUF), improved technology and the proliferation of individual data have made it increasingly difficult to protect. Synthetic data are an alternative that reproduce the statistical properties of administrative data without revealing individual taxpayer information. This paper evaluates the quality and safety of the first fully synthetic PUF and demonstrates its performance in tax model microsimulations. The synthetic PUF could also be used to develop and debug statistical programs that could then be safely run on confidential data via a validation server.

Keywords: synthetic data, privacy, individual income taxes, validation server

JEL Codes: C15, C18, H24

I. INTRODUCTION

The US Internal Revenue Service (IRS) possesses invaluable data from individual income tax returns that could vastly expand our understanding of how tax policies affect behavior and how those policies could be made more effective. For decades, the IRS Statistics of Income (SOI) Division has released an annual

Claire McKay Bowen: Urban Institute, Washington, DC, USA (cbowen@urban.org); Victoria L. Bryant: Statistics of Income Division, Internal Revenue Service, Washington, DC, USA (Victoria.L.Bryant@irs.gov); Leonard Burman: Urban Institute, Washington, DC, USA (lburman@urban.org); Surachai Khitatrakun: Urban Institute, Washington, DC, USA (skhitatrakun@urban.org); Robert McClelland: Urban Institute, Washington, DC, USA (rmcclelland@urban.org); Livia Mucciolo: Urban Institute, Washington, DC, USA (LMucciolo@urban.org); Madeline Pickens: Urban Institute, Washington, DC, USA (mpickens@urban.org); Aaron R. Williams: Urban Institute, Washington, DC, USA (awilliams@urban.org)

Electronically published November 4, 2022

National Tax Journal, volume 75, number 4, December 2022.

 https://doi.org/10.1086/722094

public-use file (PUF), a privacy-protected stratified sample of individual income tax returns. Several organizations, including the Urban-Brookings Tax Policy Center (TPC), have developed PUF-based microsimulation models that help inform the public on potential impacts of policy proposals.

However, massive expansions in the amount of data available in external databases combined with increased computational power and more sophisticated technologies have made it increasingly difficult to protect from reidentification any anonymized microdata samples that are constructed using traditional statistical disclosure control (SDC) methods. For example, in 2019, computer scientists from Imperial College London and Université Catholique de Louvain examined simple deidentification and sampling. They estimated that 99.98 percent of Americans in an anonymized health data set could be uniquely identified with 15 detailed attributes that included ZIP code, date of birth, gender, and number of children (Rocher, Hendrickx, and De Montjoye, 2019). In another example, the US Census Bureau conducted a simulated attack on the 2010 Decennial Census, which used various traditional SDC techniques to protect against disclosure. They discovered they could reidentify one-sixth of the US population using publicly available data from sources such as Facebook (Leclerc, 2019). These examples show that reidentification attacks are a serious risk for data sets with detailed geographic information and detailed categorical variables.

Although these vulnerabilities do not directly apply to the PUF, because it lacks geographic information, has no sensitive categorical variables, and is not widely available, awareness of the growing threats to public-use microdata has led SOI to progressively restrict and distort more and more information in the PUF. This has made the PUF less useful for academic research and policy analysis. Continuing to remove or distort more variables in response to growing threats would further undermine the PUF's usefulness.

Fortunately, modern privacy protection methodologies can create a useful PUF and provide more robust protection against reidentification. In this paper, we describe our process for generating a fully synthetic public use file, which we call SynPUF, consisting of pseudo records of individual income tax returns that are statistically representative of the original data (Little, 1993; Rubin, 1993). The SynPUF for tax year 2012 is a proof of concept that we plan to extend and improve upon for later tax years.

Our methodology is an extension of Bowen et al. (2020), which created a synthetic version of tax year 2012 data on individuals who did not file a federal tax return, had no obligation to file, and were not claimed as a dependent based on information reported to the IRS from third parties. The synthetic nonfiler database (Supplemental SynPUF) replicated the characteristics of the underlying administrative data and protected individual information from disclosure. Microdata on nonfilers had never been released publicly before, so the Supplemental SynPUF will be useful for policy analysts who focus on low-income populations as well as those who want to study the entire income distribution rather than only those who file a tax return.

While our process is based on some of the lessons learned from the Supplemental SynPUF, not all lessons directly apply to the new SynPUF because individual taxpayer

data are much more complex and diverse than the nonfiler data. Our most important contribution is how we tackle these additional challenges in our synthetic data methodology, such as synthesizing the survey weights, selecting the order of variables for synthesis, and applying variable constraints. Many of these lessons will hold for creating future synthetic PUFs. An additional benefit is that the SynPUF may be produced more quickly than the traditional PUF.

Our ultimate goal is to generate a SynPUF that maintains strict data confidentiality and provides better data quality than the traditional PUF. For this purpose, we rigorously evaluate the SynPUF against both the administrative tax data and the PUF on various disclosure risk metrics and utility (or usefulness) measures. In Bowen et al. (2022), we show that the risk of disclosure in the SynPUF is extremely small, even before we add anonymizing noise to the actual values. After noise is added, the chance of any identifiable tax return information being reflected in the synthetic data is virtually nil. In this article, we demonstrate that the data quality of the SynPUF is high, although there is clearly room for improvement.

Finally, we are simultaneously developing a methodology whereby researchers can develop and debug statistical programs on the synthetic data and then submit them to a validation server, a system that allows the programs to run on the confidential data. The validation server modifies parameter estimates as necessary to protect confidentiality and preserve statistical validity. This is a valuable role of the synthetic data even when they are imperfect proxies for the real data.

This paper shows that it is possible to create a fully synthetic PUF that performs well in microsimulation modeling and in certain other tasks. Furthermore, we outline a list of planned quality improvements for the SynPUF that could yield a file that is more reliable for many purposes than the traditional PUF. Importantly, by automating much of the process of privacy protection, SOI will be able to produce and release these files faster than the traditional PUF and better protect taxpayer privacy. Because privacy protection is much more robust, the SynPUF may be safely distributed much more widely than the PUF. And, finally, the combination of a SynPUF and a validation server promises to allow much broader research access to confidential data.

II. DATA SYNTHESIS METHODOLOGY

In this section, we briefly describe how the SOI creates the confidential administrative tax data as well as the PUF that is derived from those data. We summarize our synthetic data generation process and how we address several of the challenges in synthesizing the complex tax data. For details on our methodology and testing, see Bowen et al. (2022).

A. Administrative Tax Data

The IRS processes all federal individual income tax returns filed in the United States and stores a selection of line items in the IRS Master File, a massive tax

database. In 2012, the file contained approximately 145 million unedited tax returns. However, the usefulness of the Master File for analysis is limited because it excludes some key variables, has data inconsistencies, and its size makes it unwieldy. To rectify those shortcomings, SOI creates an annual sample called INSOLE (Individual and Sole Proprietor). The INSOLE is a stratified sample of individual income tax returns that are edited to add items not captured in the Master File, including detail from most forms and schedules attached to Form 1040, and to be internally consistent (Burman et al., 2018). The INSOLE is the basis for the PUF and our synthetic data file.

The stratification design oversamples tax returns with high income (positive or negative). The highest-income returns are sampled at a 100 percent rate, and the lowest-income returns are sampled at a rate just less than 1 in 1,000. In addition, Congress requires some records be included in the INSOLE. Access to the INSOLE is limited to the Congressional Joint Committee on Taxation, Treasury's Office of Tax Analysis, IRS researchers, and a limited number of external researchers who are given access to the data for specific research projects.

To produce the PUF, SOI draws a sample from the INSOLE and applies traditional SDC techniques, including deidentification, rounding, swapping, and aggregating records (Bryant et al., 2014; Bryant, 2017). Many INSOLE records are dropped to reduce the sampling rate and to protect privacy, making any individual return unlikely to be in the PUF (Burman et al., 2018). SOI first produced an annual PUF for tax year 1960. The most recent PUF is for tax year 2014.

B. Generating Synthetic Data

We describe our synthetic generation in three steps: a synthesis preparation step, a synthetic data generation step, and a postprocessing step. We classify any data preparation as our synthesis preparation step. Our synthetic data generation step encompasses the synthetic data generation process and any additional noise or modifications we introduced. In the postprocessing step, we include modifications to the synthesized data to ensure consistency among variables as well as tax return calculations.

1. Synthesis Preparation

In the first step, we derive a source file from the INSOLE called modINSOLE, which is designed to protect privacy and preserve information about the distribution of tax return data. The modINSOLE has more records than the PUF but fewer than the full INSOLE. The SynPUF has the same sample size as the modINSOLE. To generate the 2012 modINSOLE, we use tax returns for tax year 2012 that were filed in calendar year 2013. We combine strata with similar weights and characteristics, which reduces the number of strata from 98 to 25 (Bowen et al., 2022). To protect privacy, we adjust the sampling rates for high-income returns so that the likelihood

that any tax return is included in the file is 20 percent or less. By comparison, the PUF samples at a maximum rate of 10 percent. Thus, the modINSOLE preserves twice as many returns, and more variation in key variables, among the highest-income tax returns as in the PUF. By comparison, the INSOLE samples those returns at a 100 percent rate.

A key step in preparing the modINSOLE is deciding how to handle its survey weights. Leaving the original survey weights on a file can create disclosure risks. For example, a small survey weight for an observation in a synthetic data set derived from the INSOLE would indicate that the information used to create the observation is likely or mostly from a single extreme case. Furthermore, applying original weights to synthetic data without adjustment can lead to biased estimates and synthesized population totals that do not align with the population (Mitra and Reiter, 2006).

Most synthetic data applications either remove weights from the synthesis or simplify the problem. We simplify the problem by generating synthetic data within each stratum to preserve the survey weight relationship to our synthetic records. For any modINSOLE stratum that combines records from INSOLE strata with different weights, we repeatedly sample with replacement records from the stratum or strata with larger weights and recalculate the weights downward until all the sampling weights of all records within the new stratum are identical. This process allows us to vary the synthesis order and apply other synthesis strategies by stratum and implement a wider range of machine learning algorithms.

Overall, the sampling procedure discards fewer records from the INSOLE than the PUF does, which preserves more information to train our model. In 2012, the INSOLE contained 338,350 records, the PUF 172,415, and the modINSOLE and SynPUF contained 265,239.

2. Synthetic Data Generation

The second step is the generation of synthetic tax records. Below, we outline our synthetic generation model, including how we handle sparse categorical variables, how we deal with outliers, and how we choose the order in which we synthesize variables.

Categorical and discrete variables. Our synthesis procedure begins with 28 variables that are categorical (e.g., tax filing status), integer counts (e.g., number of dependents), or integer numeric variables (e.g., age in years). Instead of synthesizing categorical values sequentially, which may generate unobserved combinations of attributes, we simultaneously synthesize sets of related variables from their joint distribution. We first draw sets at random from the pool of observed combinations. We then check that the joint frequencies of these variables are sufficiently dense (i.e., counts are not ones or twos). When necessary, we combine categories and top-code

to avoid sparse combinations that might reveal that a unique individual is in the modINSOLE.[1]

Continuous variables. To synthesize continuous variables, we apply a sequence of regression trees. Regression trees are sometimes referred to as CART models. The application of regression trees in data synthesis dates to Reiter (2005), who proposed to generate partially synthetic data using a collection of nonparametric models developed by Breiman et al. (1984). Because this data-driven method is more flexible than parametric approaches, such as regression-based models, it can account for unusual variable distributions and nonlinear relationships that can be hard to identify and model explicitly. Evidence suggests that regression trees tend to outperform regression-based parametric methods (Goldstein et al., 2020; Drechsler and Hu, 2021).

Regression trees use predictor variables to sort data through a sequence of binary splits that end in homogenous groups known as nodes. A regression tree creates nodes with the smallest sum of squared errors (calculated as squared deviations from the mean). Therneau and Atkinson (2019) describe the procedure as follows:

1. Identify the variable and threshold that best splits the data into two groups. Split the data and calculate the sum of squared errors for each group.
2. For each subgroup, select the variable that best splits the data into two groups. Split the data and calculate the sum of squared errors for each group.
3. Continue this process until the subgroups reach a user-specified minimum size or tree depth, or the sum of squared errors stops decreasing.
4. If necessary to avoid overfitting, apply cross-validation and combine two or more nodes into one larger node.

To use an estimated tree to synthesize an observation, we use predictor variables for that observation to navigate the splits in a tree until a unique final node is identified. Traditionally, a prediction function (such as a mean for regression trees and mode for classification trees) is applied to this node to obtain a predicted value. Instead, as we describe below, we follow Reiter (2005) and synthesize a value by drawing an observation at random from the final node. This avoids the problem that conditional mean imputation may result in data with incorrect sample variances and covariances and too few values in the tails of marginal distributions (Little and Rubin, 2020).

We synthesize a variable with a regression tree model using all previously synthesized outcome variables as potential predictors. This is based on the insight that

[1] We also relegate some "nonessential" count variables to the end of the synthesis process (Bowen et al., 2022).

a joint multivariate probability distribution can be represented as the product of a sequence of conditional probability distributions. We define this mathematically as

$$f(X_1, X_2, \ldots, X_k|\theta) = f_1(X_1|\theta) \cdot f_2(X_2|X_1, \theta) \cdots f_k(X_k|X_1, X_2, \ldots, X_{k-1}, \theta),$$

where X_i for all $i = 1, \ldots, k$ are the variables to be synthesized, θ are vectors of model parameters, such as regression coefficients and standard errors, and k is the total number of variables.

The order in which we synthesize variables is important because those synthesized later in the sequence tend to be noisier (Goldstein et al., 2020). We tested several methods for determining the synthesis order and concluded that ordering based on the weighted sum of absolute values yields the highest quality data.

Because the ith regression tree model would need to consider $i - 1$ variables, and i can be as large as 150, we reduce the number of predictors by dropping those with very low variance (Kuhn and Johnson, 2013). We define a variable to be low variance if the most common value occurs in about 95 percent of observations, the second most common value occurs in about 5 percent of the observations, and unique values make up no more than 1 percent of the observations.

Dealing with outliers. Because a regression tree minimizes the sum of squared errors, outliers carry more weight in model estimation than values near the middle of the distribution. In testing, synthesized variables frequently have too many nonzero values and means too far from zero if the underlying variable is highly skewed, such as variables associated with capital gains.

In some cases, we apply a Yeo-Johnson transformation, which is a monotonic transformation that considers zero and negative values, before training a regression tree and inverting the transformation after predicting a value with the regression tree (Yeo and Johnson, 2000). The Yeo-Johnson is similar to a log transformation or a Box-Cox transformation, but it is applicable to variables with negative and zero values.

An alternative approach to dealing with skewed variables is to use predictive algorithms that minimize the sum of absolute deviations instead of the sum of squares. Least Absolute Deviation Regression Trees give equal weight to all values. In testing, we find that Least Absolute Deviation Regression Trees work best with variables with few nonzero values.

Smoothing. After drawing an observation at random from the final node, we add noise to protect against disclosing that a particular household is represented in the modINSOLE. This is necessary to preserve privacy because just sampling from the final nodes of a fitted regression tree reproduces observed values in the confidential

data. We also apply IRS rounding rules that are used to create the PUF (Burman et al., 2018).

The approach of randomly drawing from a final node and then adding noise was first developed in Reiter (2005). It is similar to using a kernel density estimator on the empirical univariate distribution of the final node and then drawing from that estimated distribution. When combined with CART, the method retains the relationships between the synthesized variable and the predictors, and approximates the conditional univariate distribution of that variable.

We follow Reiter (2005) in adding noise drawn from a zero-mean normal distribution, but we differ in how we set the variance of the distribution (Bowen et al., 2022). In Reiter's approach, the variance for the noise added to all observations in each node is set to the variance of the observations in that node. In our approach, the variance of the added noise is separately determined for each observation. We increase the variance with the sparsity of the distribution, adding more noise to observations in parts of the distribution where individual values are distinct (such as for very high-income households) and less noise in parts of the distribution where values are common and unidentifiable. This enhances privacy and increases the quality of data. In the Reiter (2005) approach, a single outlier in a final node can dramatically increase the variance for noise added to observations drawn from that node. In our approach, the noise added to each observation is unaffected by those outliers, reducing the variance of the added noise for many observations.

3. Postprocessing-Adding Constraints

Without any constraints, our approach may generate impossible values. In some cases, such as taxpayers with extremely high or low incomes, this preserves privacy, but in other cases it can create values that cannot logically be observed in tax data. In these cases, we impose various consistency checks and constraints in a way designed not to bias the synthetic data. Because synthesized variables are used as predictors for synthesizing other variables, these constraints are often imposed midsynthesis to maintain logic and consistency in the data.

Many tax variables have bounds imposed by law. For example, net capital losses may only take values in the range of $0–3,000. We impose univariate constraints on the minima and maxima of continuous variables as applicable. There are also constraints that vary by filing status or other variables. For example, educator expenses have a maximum of $500 for married taxpayers filing jointly and $250 for taxpayers with other filing statuses. A variable may need to simultaneously satisfy multiple such constraints.

Some variables must be greater than or equal to another variable or variables. As an example, total dividends must be greater than or equal to qualified dividends, total Individual Retirement Arrangement (IRA) distributions must be greater than or equal to taxable IRA distributions, and total pensions and annuities must be greater than or equal to taxable pensions and annuities.

To impose these linear constraints, we may synthesize component parts (e.g., taxable and nontaxable IRA distributions) and sum them or model some variables as proportions of other variables. For example, we model the wage split between primary and secondary taxpayers on married filing jointly returns as a function of total reported wage and salaries and other variables.

If a constraint cannot be applied with postprocessing or synthesizing components, we apply one of two different approaches during the synthesis step: hard bounding and z-bounding. For some variables, where the law imposes a boundary, such as capital losses, those are simply imposed. This is called hard bounding. However, this method can cause truncation bias when there is not an underlying approximately unbounded distribution generating the data.

To mitigate the bias caused by hard bounding, we apply the z-bounding technique (Drechsler, 2011). When values are out of bounds after applying IRS rounding rules, we repeat the synthesis until the synthesized value satisfies all relevant constraints or until it reaches a maximum number of iterations, z. If after z iterations the value still fails to meet the requirements, we hard bound the value.

Several variables are computed as a function of other variables reported on the income tax return. Those variables include various tax credits, itemized deductions, adjusted gross income (AGI), taxable income, and alternative minimum tax liability. These are all calculated based on previously synthesized values. We also cap several variables to protect privacy or to preserve variable relationships between variables in the INSOLE. Finally, we reweight the synthetic data to match SOI published totals by income group to guarantee that key aspects of the synthetic data match the US population distribution.

III. EVALUATION OF SYNPUF

Synthetic data sets are evaluated in two dimensions: privacy and utility (or usefulness). Bowen et al. (2022) describe in detail the privacy metrics we used to assess the SynPUF. Based on those metrics, the data set is safe from disclosure risk. There still may be risks associated with synthetic data, but we show in earlier work that those kinds of inferential risks are minimized by the combination of synthesis, sampling, and smoothing (Burman et al., 2018).

Of course, protecting privacy does not guarantee that a data set is useful. A totally safe data set might consist of all zeros or random numbers, which would be useless for analysis. Our goal is to produce a synthetic data set that is useful for policy analysis and protects against disclosure of individual taxpayer information. There is no generally accepted summary statistic of synthetic data quality but rather an array of standard metrics, which we report below. We compare weighted counts, means, and standard deviations in both the SynPUF and the PUF with the INSOLE. We also compare correlations between the SynPUF and the INSOLE. We find that the SynPUF generally performs as well or better than the PUF, although there is obvious room for improvement.

A. Data Quality Metrics

A simple measure of quality is whether it reproduces counts of tax returns with nonzero values. The SynPUF and the PUF both meet this test well. Figure 1 shows scatter plots of weighted counts in the SynPUF and PUF plotted against the corresponding counts in the INSOLE. Each dot represents a variable such as wages and salaries or interest income. The diagonal line represents equivalent weighted counts; distance of the dots from the line reflects the size of count errors. The figure shows that for most variables, the weighted counts in the three data sets are close. The SynPUF has fewer notable deviations from the INSOLE than the PUF, and the discrepancies appear to be smaller.

The SynPUF also closely matches all the weighted means in the INSOLE (Figure 2). The PUF also matches fairly closely but has more notable discrepancies than the SynPUF.

However, both the PUF and the SynPUF do less well at capturing the higher moments of univariate distributions. For instance, several variables have much smaller standard deviations in the synthetic data than in the INSOLE although some have markedly higher standard deviations (Figure 3). The PUF appears to perform even worse by this metric, although the standard deviations are much more likely to be understated in the PUF than in the SynPUF. This is probably because some traditional SDC measures, such as grouping records into a single aggregated record, reduce variability for highly skewed variables such as capital gains. We also examined higher moments (not shown), and those also tended to be understated in the synthetic data. Correcting these shortcomings is a high priority for our future work (see Section IV).

We would also like the SynPUF to capture the correlation between all pairs of variables. Figure 4 compares the correlations for all variables that are included in

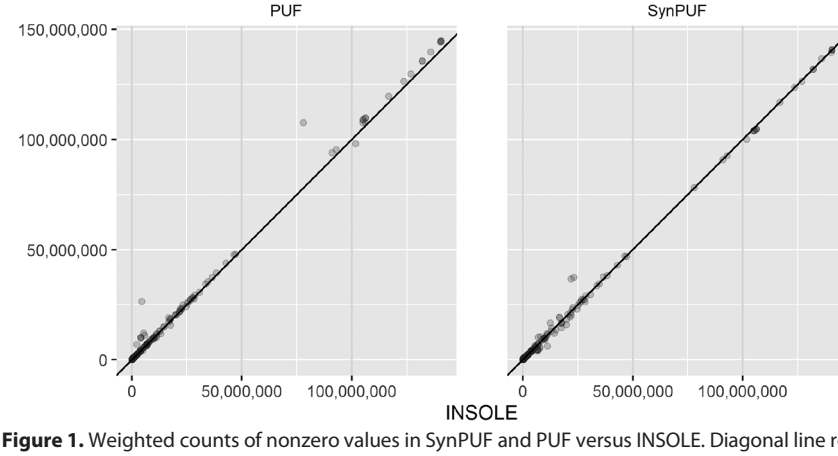

Figure 1. Weighted counts of nonzero values in SynPUF and PUF versus INSOLE. Diagonal line represents equivalence. Difference between PUF/SynPUF and INSOLE statistics is the vertical distance between the point and the diagonal line, representing equivalence. Data are from Urban Institute, Tax Policy Center.

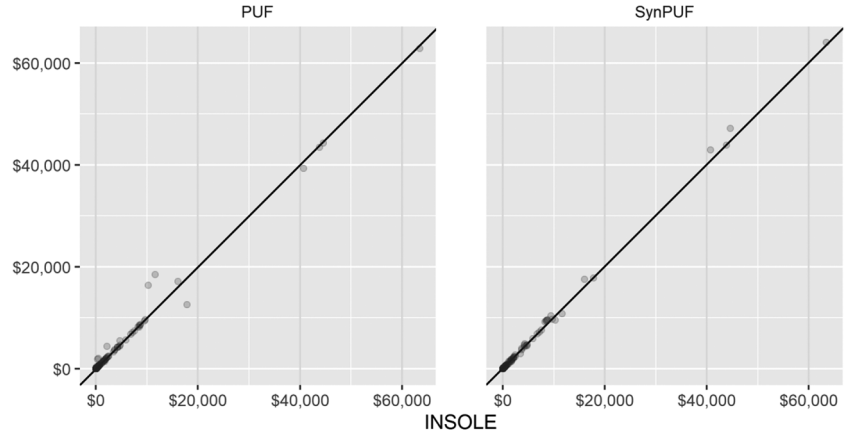

Figure 2. Weighted variable means in SynPUF and PUF versus INSOLE. Diagonal line represents equivalence. Difference between PUF/SynPUF and INSOLE statistics is the vertical distance between the point and the diagonal line, representing equivalence. Data are from Urban Institute, Tax Policy Center.

the SynPUF with the corresponding correlations in the INSOLE. The correlations are very similar between the two data sets. While there are a few outliers, 96 percent of correlation differences are less than 0.01 and about 56 percent of differences are less than 0.001.

We examined multivariate correlations by regressing selected variables against an assortment of other variables. In most cases, confidence intervals for coefficients on right-hand-side variables (not really explanatory variables because we are not estimating a structural model) overlap between the SynPUF and the administrative data

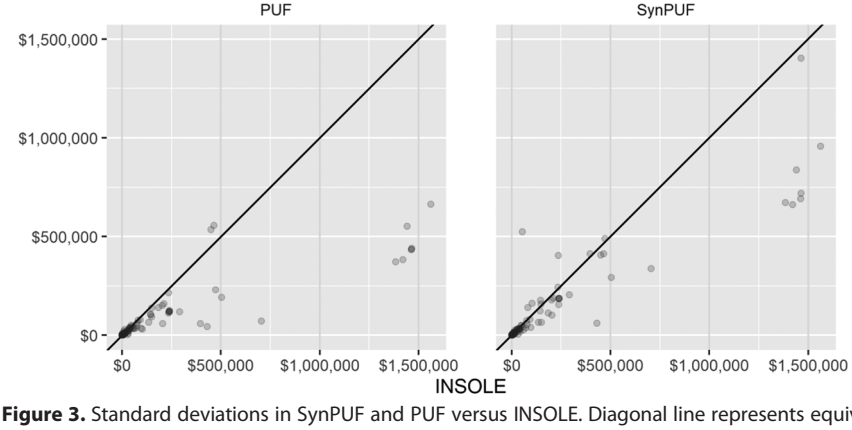

Figure 3. Standard deviations in SynPUF and PUF versus INSOLE. Diagonal line represents equivalence. Difference between PUF/SynPUF and INSOLE statistics is the vertical distance between the point and the diagonal line, representing equivalence. Data are from Urban Institute, Tax Policy Center.

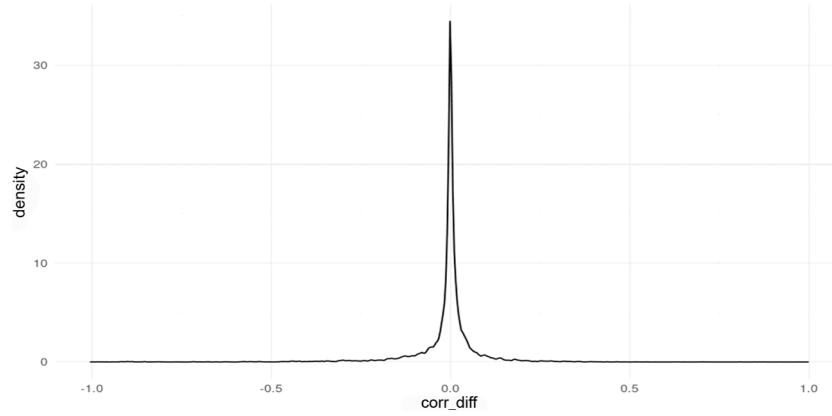

Figure 4. Density of pairwise correlation differences between SynPUF and INSOLE. Data are from Urban Institute, Tax Policy Center.

and signs are almost always the same (Bowen et al., 2022). There is no clear winner in the statistical horserace between the SynPUF and the PUF. In some cases, estimates based on the SynPUF are closer to those in the administrative data, while in others, the PUF appears to be better. As reflected in Figure 4, one-variable regressions tend to produce very close results (regression of one variable and an intercept against another is equivalent to measuring a correlation). Regressions with more right-hand-side variables are less likely to produce the same statistical inferences for all variables.

We compared the distribution of AGI and taxable income by AGI group in the SynPUF and PUF to the distribution derived from the INSOLE. Both the PUF and SynPUF capture the weighted distribution of AGI by income group well, although the PUF slightly underestimates AGI and the SynPUF slightly overestimates it at the very top of the distribution (Figure 5).

The SynPUF undercounts the number of filers with positive taxable income in the bottom 60 percent, while the PUF slightly overcounts filers with positive taxable income at all income levels (Figure 6). In total, the SynPUF comes closer to the total count than the traditional PUF.

Figure 7 shows the distribution of total income. The synthetic data overstates negative income in the bottom 60 percent and positive income in the top 1 percent. The rest of the distribution of the synthetic data closely matches the distribution of the confidential data.

Perhaps not surprisingly, income tax liability is also overstated at the top of the distribution when we apply 2012 tax law (Figure 8). Here the bottom 60 percent is less of an issue because their total income tax liability is close to zero. But the progressive income tax magnifies the difference in taxable income into the difference in income tax liability of the synthetic and confidential data for the top 1 percent.

This difference appears to be mostly driven by capital gains (Figure 9). Panel A shows that the synthesizer matches the underlying distribution of capital gains

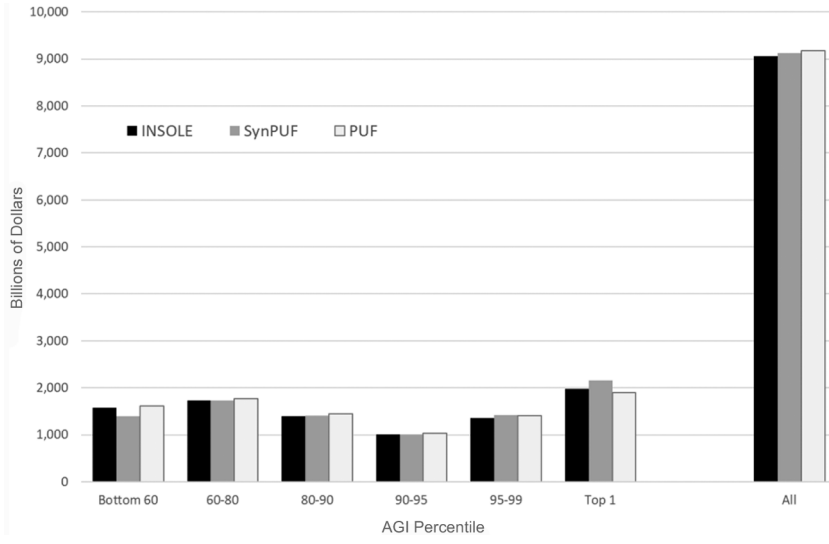

Figure 5. Adjusted Gross Income (AGI), tax year 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

income more closely than in the PUF. However, Panel B shows that taxes on capital gains are too high on the synthetic file. This is because the synthesizer is assigning too much capital gains at the higher tax rates and too little at lower tax rates. The SynPUF overstates capital gains by about $64 billion, or 16 percent, in the top 1 percent.

B. Tax Policy Simulations

The TPC and other organizations use the PUF for microsimulations of policy changes. Thus, it is important that the SynPUF produces reliable estimates for model simulations. We adapted the tax calculator in the open-source Policy Simulation Library to reflect 2012 tax law.[2] We used the model to simulate some illustrative policy changes.

We modeled two sets of tax law changes on both the PUF and SynPUF.[3] The two sets of reforms were modeled to resemble 2000 and 2013 tax law, with all amounts adjusted to 2012 dollars. These comparisons are informative because Congress and President Obama were negotiating about whether to extend all or some of the Bush Administration income tax cuts, which were set to expire in 2012. Some of the income tax cuts were extended, including increases in the standard deduction for married couples filing a joint return, increases in the child tax credit, and income tax rate cuts and marriage penalty relief for all but the top income tax brackets. We simulated the

[2] See https://taxcalc.pslmodels.org/usage/overview.html for a description of the model.

[3] Technical problems prevented us from simulating the policy changes on the modINSOLE. We plan to include comparisons with the administrative data in an updated version of Bowen et al. (2022).

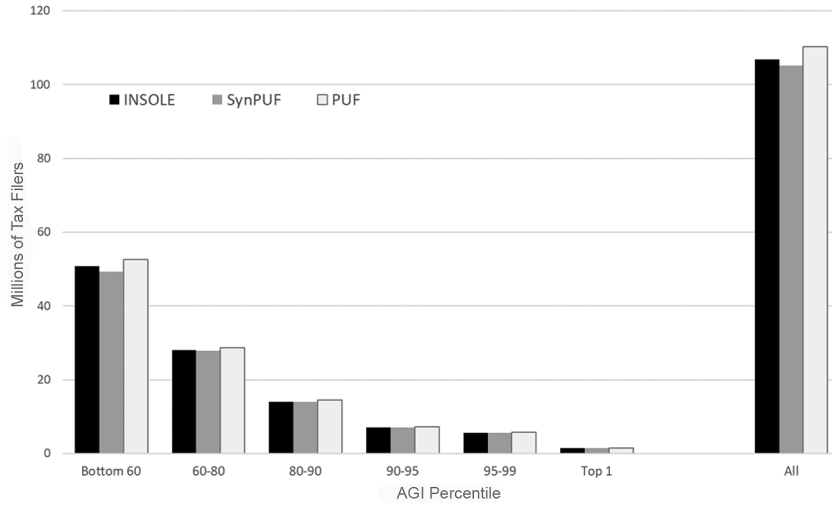

Figure 6. Number of filers with positive taxable income, tax year 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

effect of those changes — that is, applying 2013 law in 2012 — as well as repealing all of the Bush income tax cuts.

1. 2000 Tax Law Changes

We set the individual income tax rates and brackets to 2000 tax law by lowering the number of tax rates from six to five. This meant the rate schedule would have changed

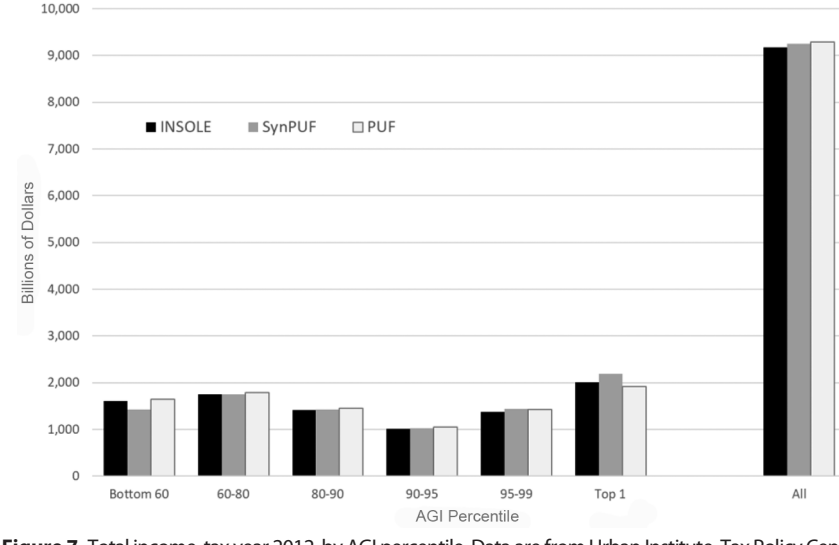

Figure 7. Total income, tax year 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

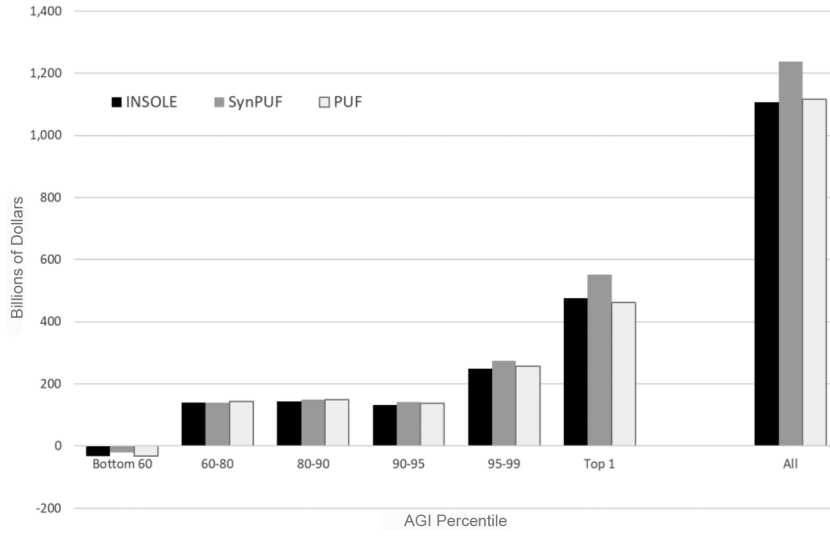

Figure 8. Income tax liability, tax year 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

from 10/15/25/28/33/35 to 15/28/31/36/39.6. The income tax thresholds were reduced for married filing jointly returns in the four lowest tax brackets, undoing the marriage penalty relief enacted in 2001. Similarly, we changed the standard deduction from $11,900 to $9,900 for married filing jointly returns. We also modified the capital gains rate from a single rate of 15 percent to two rates, 10 and 20 percent. The simulation would also tax dividends at the same rate as ordinary income, instead of at the capital gains rate. We reintroduced the personal exemption phaseout and the phaseout of itemized deductions, which had been repealed in 2012. The maximum child tax credit would fall from $1,000 to $500. The maximum child and dependent care tax credit would fall from $6,000 to $4,800. The maximum credit rates would fall from 35 to 30 percent, and the phaseout threshold would be reduced.

If the Bush tax cuts had been repealed in 2012, the overwhelming share of the tax increase would have fallen on the top 10 percent of tax filers (Figure 10). This is apparent on both the PUF and the SynPUF, but the tax increase appears to be larger in the SynPUF in the top 10 percent because too much income is assigned to households in the top bracket.

As noted earlier, too much capital gains are assigned to households in the top 1 percent who are subject to the highest capital gains tax rate. The top 1 percent in the SynPUF pays almost $7 billion more in additional capital gains tax than in the PUF (Figure 11). Not surprisingly, a similar pattern occurs in the distribution of income tax rate increases (not shown).

In contrast, the synthesis seems to understate the effect of cutting the child tax credit to 2000 law levels for households with very low incomes (Figure 12). This could be because the synthesis produces too few qualifying children or because earnings for

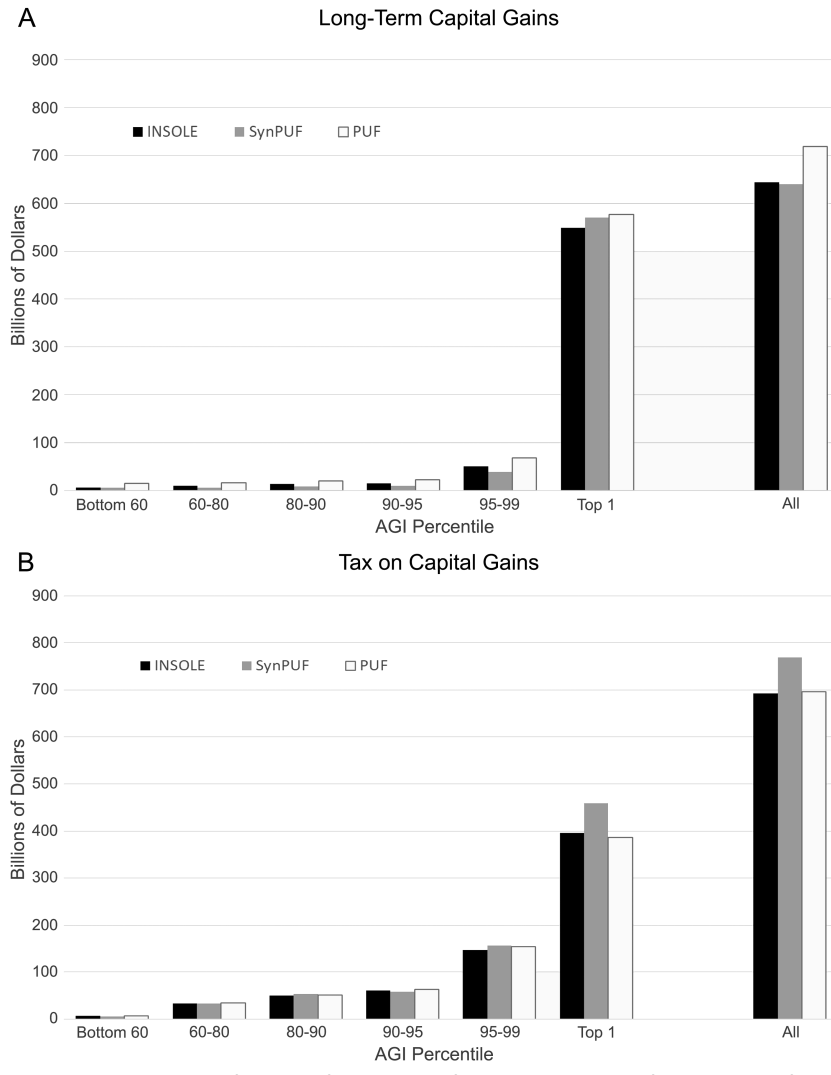

Figure 9. Long-term capital gains and tax on capital gains, tax year 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

low-income households are underestimated. Because the refundable portion of the child tax credit phases in with earnings, that would reduce refundable credits, which are most important for households with very low incomes.

2. 2013 Tax Law Changes

In 2013, income taxes were increased for very high-income taxpayers compared with 2012, although not 2000, law. We simulated the effect of those changes in 2012

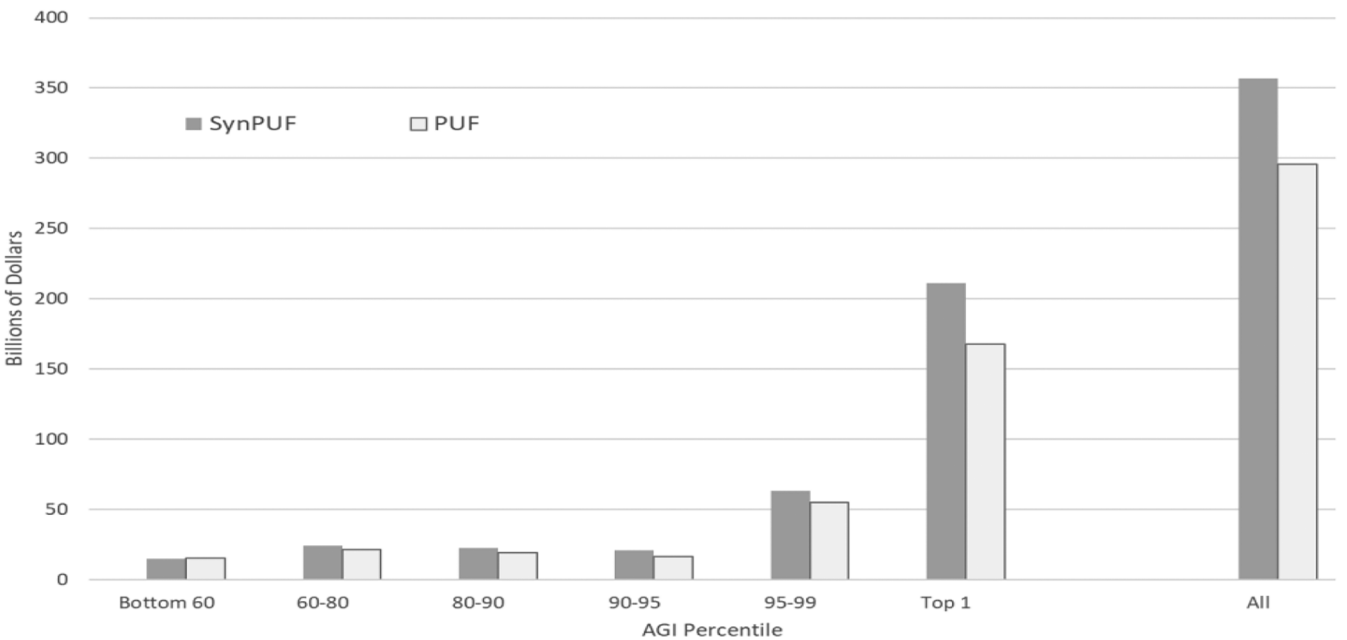

Figure 10. Total change in tax from repealing the Bush tax cuts in 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

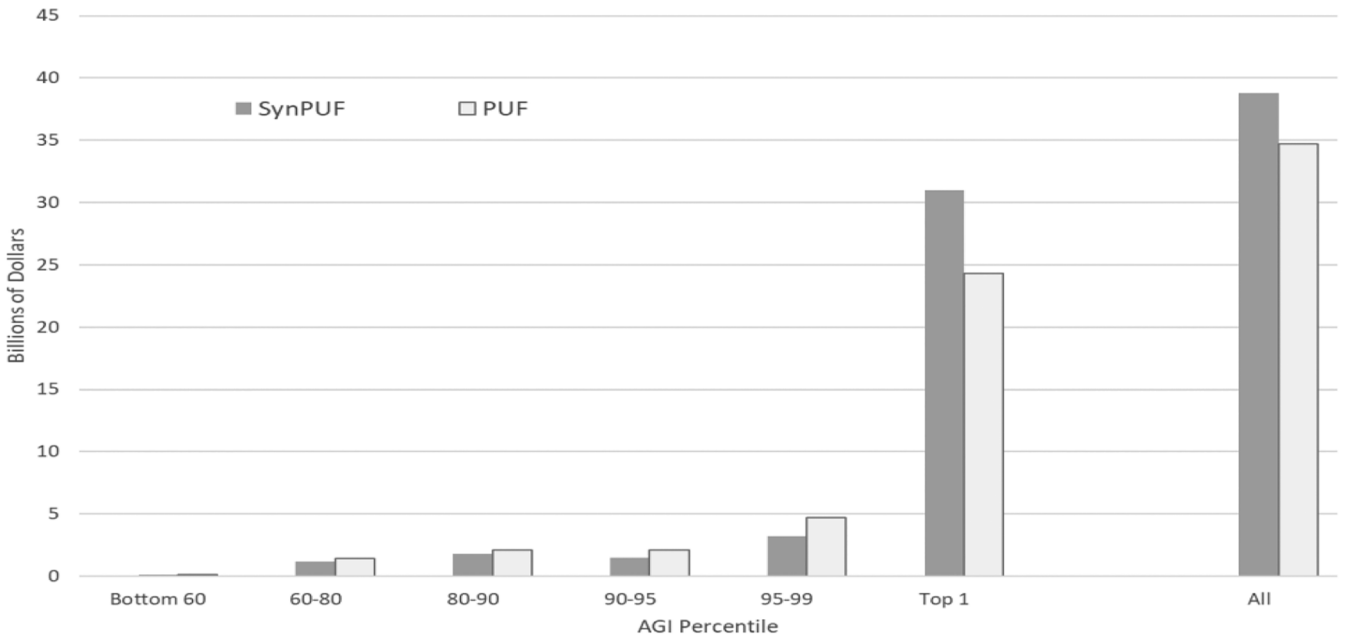

Figure 11. Total change in tax from restoring tax rates on long-term capital gains to 2000 levels, in 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

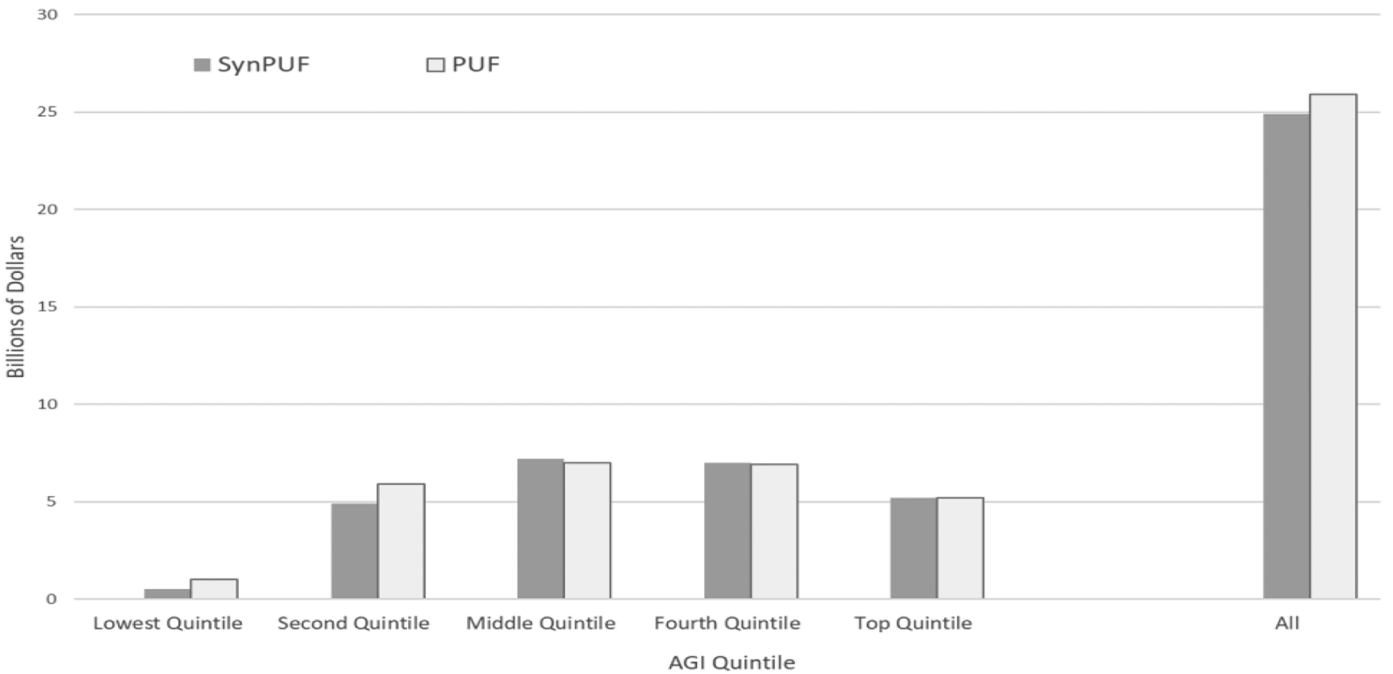

Figure 12. Total tax change from restoring the child tax credit to 2000 levels, in 2012, by quintile. Data are from Urban Institute, Tax Policy Center.

by creating a new higher top tax bracket with an ordinary income tax rate of 39.6 percent and a capital gains tax rate of 20 percent. The personal exemption phaseout and phaseout of itemized deductions, which had been repealed in 2012, were restored, although at higher income levels than would have applied under 2000 law.

Simulations on the SynPUF and the PUF both show that these tax changes were concentrated at the top of the income distribution, but the point estimates differ significantly (Figure 13). As in the earlier simulations, SynPUF appears to overestimate tax changes because too much income is taxed at the top bracket rates on ordinary income and capital gains.

Overall, these figures show that while our synthetic data have a similar distribution as both the underlying data and the PUF, there are marked shortcomings. We plan a number of measures to improve the synthesis, as described in the next section.

IV. CONCLUSIONS AND PLANNED FUTURE WORK

This project has demonstrated the feasibility of producing a fully synthetic public use file of individual income tax return information. The SynPUF matches fairly closely on key characteristics of the underlying data, such as univariate means and most correlations between variables. The data also adhere to hundreds of logical constraints that reflect the complexity of the tax code. Based on preliminary tests, the data set appears promising for microsimulation modeling, but certain aspects, such as the correlation between capital gains and other variables, need refinements.

The synthetic data have several advantages over the traditional PUF. First, the synthesis is designed to provide a robust protection against disclosure of individual data for every variable and record. Traditional SDC methods require identifying privacy risks associated with particular records or variables and designing targeted approaches to mitigate those risks.

Second, the process of creating a synthetic PUF can be nearly automated, saving staff time and other resources. SOI will still want to carefully assess the utility and privacy of the resultant file and probably submit it to trusted users for beta testing before release, but it should be feasible to create synthetic PUFs within months. By comparison, release of the traditional PUF has lagged production of the INSOLE by many years. As of this writing, the most current PUF is for tax year 2014. Importantly, given resource constraints at the IRS, the synthesis process could require much less staff time than the current methods.

Third, a synthetic data file could safely include more variables than are currently released in the PUF. These variables allow analysts at Congress's Joint Committee on Taxation and Treasury's Office of Tax Analysis to analyze the effects of public policies for a wide range of policies. Including the full suite of INSOLE variables in the SynPUF would allow analysts outside of government to model the same range of policies as the official scorekeepers. Such independent vetting of the effects of current and proposed policies better informs the public and can strengthen confidence in the information released from official sources.

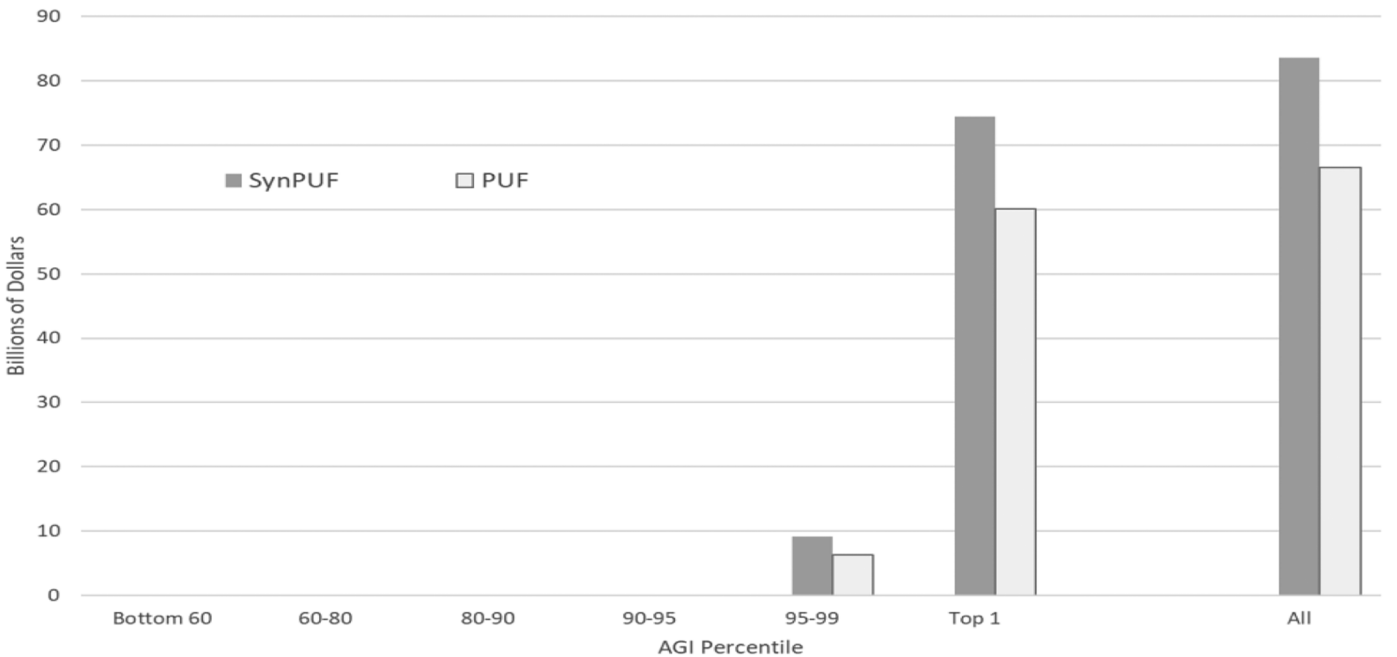

Figure 13. Total tax change from imposing 2013 law in 2012, by AGI percentile. Data are from Urban Institute, Tax Policy Center.

Finally, because the SynPUF does not include any actual tax records, SOI could safely distribute it to a broader research audience than the traditional PUF.

While there are exciting opportunities that will become possible with the development of the SynPUF, we have identified several promising avenues to improve the quality of synthetic data. For example, we believe that separately synthesizing positive and negative values of variables that can take a negative value will increase the SynPUF utility. Developing and applying machine learning methods such as random forests may also increase the overall quality. We also may be able to improve the SynPUF for specific types of analyses by reweighting it to match published SOI aggregates by income and filing status and optimizing the synthesis so that it produces accurate microsimulation analyses for more policies. We would also like to apply the synthesis methodology to later years of tax data and to develop complementary data sets, including more demographic information and more geographical detail, as well as data sets based on information compiled from information returns, such as detailed data for head of household and spouse, health insurance coverage, pensions and IRAs, home sales, executive compensation, and so on. Finally, two long-term goals are to develop synthetic panels of individual tax data and synthetic corporate tax data.

Some of the synthetic data sets we plan to create are relatively straightforward applications of our methodology, but some present new challenges. For instance, although there are examples of synthetic panel data (e.g., the US Census Bureau's Synthetic Longitudinal Business Database), tax data have many more records and variables. This makes preserving both privacy and utility very difficult. Corporate tax data also present unique challenges because large corporations in concentrated industries may be easy to identify in any data set that is accurate at the industry level. Despite these challenges, synthetic corporate tax data should still be valuable as training data for a validation server even if data privacy protections limit their use for standalone analyses.

As noted, we are building a validation server, which is an essential complement to the synthetic data. Synthetic data are unreliable for estimating statistical relationships that they were not specifically designed to reflect. In particular, they are unlikely to accurately reflect the response to quasiexperiments, such as those generated at kinks and notches in marginal tax rate schedules, because they are smoothed out in the synthesis process. A validation server would allow researchers to measure such responses. It could produce estimates for a wide range of models based on the underlying administrative data with parameter estimates altered and standard errors adjusted commensurately to protect privacy and preserve statistical validity (Barrientos et al., 2021; Taylor et al., 2021).

Researchers would use synthetic data that have the same record layout as the confidential data for exploratory analysis and to test and debug complex statistical programs before submitting those programs to the validation server. Vilhuber and Abowd (2016) describe a system that provides access to the confidential version of the Survey of Income and Program Participation and produces statistical output after a privacy review by a US Census Bureau staff member. Our goal is to create a

similar framework that would modify statistical outputs to guarantee privacy and preserve statistical validity, eventually without requiring human review.

ACKNOWLEDGMENTS AND DISCLAIMERS

We are grateful to Lillian Hunter for preparing the figures and help with manuscript preparation, to our advisory boards for helpful comments and advice, and to those who tested our synthetic supplemental PUF. The work described in this paper would not have been possible without generous support from the Urban Institute's funders, including Arnold Ventures, the Alfred P. Sloan Foundation, and the National Science Foundation. The findings and conclusions are those of the authors and do not necessarily reflect positions or policies of the IRS, the Urban Institute, or its funders.

DISCLOSURES

The authors (other than Bryant) have received financial support for this research from Arnold Ventures, the Alfred P. Sloan Foundation, and the National Science Foundation.

REFERENCES

Barrientos, Andrés F., Aaron R. Williams, Joshua Snoke, and Claire McKay Bowen, 2021. "A Feasibility Study of Differentially Private Summary Statistics and Regression Analyses for Administrative Tax Data." *arXiv*, preprint arXiv:2110.12055.

Bowen, Claire McKay, Victoria Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Philip Stallworth, Kyle Ueyama, and Aaron R. Williams, 2020. "A Synthetic Supplemental Public Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications." In *International Conference on Privacy in Statistical Databases*, 257–270. Springer, Cham.

Bowen, Claire McKay, Victoria Bryant, Leonard Burman, Robert McClelland, Livia Mucciolo, and Aaron R. Williams, 2022. "Synthetic Individual Income Tax Data: Promises and Challenges." Tax Policy Center, Washington, DC.

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

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

Bryant, Victoria L., John L. Czajka, Georgia Ivsin, and Jim Nunns, 2014. "Design Changes to the SOI Public Use File (PUF)." *Proceedings, Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* 107, 1–19.

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." In Plumley, Alan (ed.), *The IRS Research Bulletin (Publication 1500)*, 223–246, https://www.irs.gov/pub/irs-soi/18resconburman.pdf.

Drechsler, Jörg, 2011. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Vol. 201. Springer, New York.

Drechsler, Jörg, and Jingchen Hu, 2021. "Synthesizing Geocodes to Facilitate Access to Detailed Geographical Information in Large-Scale Administrative Data." *Journal of Survey Statistics and Methodology* 9 (3), 523–548.

Goldstein, Ross, Michael E. Woolley, Laura M. Stapleton, Daniel Bonnéry, Mark Lachowicz, Terry V. Shaw, Angela K. Henneberger, Tessa L. Johnson, and Yi Feng, 2020. "Expanding MLDS Data Access and Research Capacity with Synthetic Data Sets." Maryland Longitudinal Data System Center, Baltimore, MD.

Kuhn, Max, and Kjell Johnson, 2013. *Applied Predictive Modeling*. Springer, New York.

Leclerc, Philip, 2019. "Results from a Consolidated Database Reconstruction and Intruder Re-identification Attack on the 2010 Decennial Census." Presented in the workshop "Challenges and New Approaches for Protecting Privacy in Federal Statistical Programs," June 6–7, sponsored by National Academy of Sciences, Washington, DC.

Little, Roderick J. A., 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9 (2), 407–426.

Little, Roderick J. A., and Donald B. Rubin, 2020. *Statistical Analysis with Missing Data*, 3rd ed. Wiley, Hoboken, NJ.

Mitra, Robin, and Jerome P. Reiter, 2006. "Adjusting Survey Weights When Altering Identifying Design Variables Via Synthetic Data." In Domingo-Ferrer, Josep, and Luisa Franconi (eds.), *Privacy in Statistical Databases*, Vol. 4302, 177–188. Springer, Berlin.

Reiter, Jerome P., 2005. "Using CART to Generate Partially Synthetic Public Use Microdata." *Journal of Official Statistics* 21 (3), 441–462.

Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre De Montjoye, 2019. "Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models." *Nature Communications* 10 (1), 1–9.

Rubin, Donald B., 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2), 461–468.

Taylor, Silke, Graham MacDonald, Kyle Ueyama, and Claire McKay Bowen, 2021. "A Privacy-Preserving Validation Server Prototype." Technical Paper. Urban Institute, Washington, DC.

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

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

Yeo, In-Kwon, and Richard A. Johnson, 2000. "A New Family of Power Transformations to Improve Normality or Symmetry." *Biometrika* 87 (4), 954–959.