A New Synthetic Data Set for Tax Policy Analysis

May 7, 2020; Version 1.0

Don Boyd*

Bio: Don Boyd is co-director of the State and Local Government Finance Project at the Center for Policy Research at the University at Albany's Rockefeller College and is a consultant to organizations that analyze state and local government finances, including the Pew Charitable Trusts and the Open Source Policy Center at the American Enterprise Institute. He is the principal of Boyd Research, an economic and fiscal consulting firm. Boyd holds a PhD in managerial economics from Rensselaer Polytechnic Institute.

Abstract: *Colleagues and I developed a synthetic federal income tax microdata file, with support from the Open Source Policy Center. We synthesized the file using random forests and constructed record weights that minimize differences between targets developed from the IRS public use file and corresponding weighted values from the synthetic file. The file is quite useful for some tax policy analysis purposes but less useful for others. We intend to improve file quality in future iterations. We are preparing the file and documentation for use with the Policy Simulation Library's Tax-Calculator federal income tax model, for free and without legal restrictions.*

Federal income tax policy affects every American. The Tax Cuts and Jobs Act made major changes and more are sure to come during tax and entitlement reform. Serious income tax policy analysis requires microdata based upon tax returns: aggregate data are not sufficient for analyzing distributional consequences of policies, and publicly available microdata such as the American Community Survey lack important tax-relevant information such as realized capital gains.

Most tax policy analysts outside of government do not currently have access to high-quality tax microdata. The best publicly available data, known as the IRS Public Use File (PUF), are based upon anonymized and blurred IRS administrative data but a license costs $10,000 and requires a legal agreement. Analysts with access to these data can conduct sophisticated tax policy analysis through the command-line interface to the popular and powerful federal Tax-Calculator model in the Policy Simulation Library (PSL) using a modified version of the data. Analysts without a license can still run models against the modified IRS data through the Tax-Brain web interface to Tax-Calculator, but they can only retrieve summary results, and these results are modified slightly by a validation server. PSL also includes a data file based on the Current Population Survey that is free and valuable for analysis of benefit policies but does not represent tax returns as well as the IRS-based data.

A free and unrestricted microdata file that is similar to tax-return data in characteristics and quality would have several benefits: analysts who do not have access to IRS-based data could conduct sophisticated analyses and examine detailed results; college students and professors in public policy, public finance, and similar programs would be able to learn on data that are close to the real thing; data files that represent individual states could be constructed and made freely available, allowing state analysts to analyze federal tax reform in their states and, with suitable models, analyze impacts of state tax reforms; and PSL developers could focus their data enhancement efforts on a PUF-based file rather than splitting efforts between PUF and CPS-based files as is done now.

For these reasons, my colleagues Max Ghenis[1], Dan Feenberg[2], and I set out to construct a synthetic tax microdata file based upon the existing PUF that can be used with Tax-Calculator – that is, a file that looks like tax microdata and approaches it in quality, but contains no actual tax returns and can be distributed for free without legal restrictions. With incubation support from the American Enterprise Institute's Open Source Policy Center (OSPC), we have constructed a beta synthetic file that meets IRS disclosure requirements, can be freely distributed, and is useful.

This *QN* explains what we did, examines its quality, and discusses next steps.

* *Quantitative Notes* is published by Open Research Group, Inc., a public benefit corporation. For more *QNs* or to arrange meetings with an expert, please contact experts@openrg.com. The author(s) did not receive funding from any source for the production of this *Quantitative Note*.

[1] Consultant to OSPC

[2] National Bureau for Economic Research

1. Constructing a synthetic data file

Synthetic data are created by means other than direct measurement - for example, by algorithm. They are intended to preserve key statistical characteristics of the true data such as means, variances, correlations, and patterns of missingness, while avoiding confidential information that must be suppressed. *Fully synthetic* data – the kind of data we constructed in this project – are data in which every value for every observation is synthesized: they include no real data, unlike partially synthetic data sets in which some variables or observations have real data and other values are imputed.

Several organizations have significant synthetic data projects. The Census Bureau has constructed a Synthetic Longitudinal Business Database and a synthetic Survey of Income and Program Participation, and is working on a synthetic version of the American Community Survey. The Scottish Longitudinal Study develops and maintains synthetic data on the Scottish population.[3] The Tax Policy Center is developing a synthetic public use tax data file that will go beyond our project because it will be based upon actual tax returns rather than upon PUF records, which are blurred and anonymized, but it will take longer to develop.

We create synthetic data sequentially one variable at a time, as most synthetic data projects do. The basic steps are:

1. Start by creating one or more "seed" variables that can be used to predict other variables. These variables might later be discarded from the synthetic file if not needed or they might be retained if they contain no confidential information. For example, if the synthetic file is to have 100,000 records, we might start by creating a data set with 100,000 records each of which has only one variable - marital status – sampled with replacement from values in the actual data (the PUF).
2. Synthesize values for each additional variable in two steps:
   (a) Fit a model for the variable using as regressors actual PUF values of variables that have already been synthesized. For example, we might choose wages as our first variable and model it conditional only on marital status.
   (b) Using the model just estimated, predict synthetic values for the variable. To do this, we use synthesized values of the righthand side variables as predictors.
3. Repeat: Return to step 2 for the next variable and continue until all desired variables have been modeled and synthesized.
   (a) For example, if interest income is our second synthesized variable we would fit a model for interest income conditional upon wages and marital status using actual PUF data, and then predict synthetic values of interest income based upon the already-synthesized wage and marital status values.
   (b) We continue in this fashion for each new variable, fitting a model based upon actual PUF values for already-synthesized variables, and predicting synthetic values from this model, using already synthesized values as the predictors.

Thus, we model the joint distribution as a sequence of conditional marginal distributions.

In practice, we must make many important methodological decisions, including:

- Which variables to use as seeds? We used marital status and several important calculated tax variables such as adjusted gross income (AGI) and total deductions as seeds. After synthesis, we discarded all the calculated variables we used as seeds and calculated them from synthesized component variables.
- Which variable should be synthesized first, and second and so on? Existing research provides relatively little guidance. Some analysts have tried to define the sequence according to a presumed causal chain, which in our case could mean synthesizing income variables before synthesizing the charitable deduction, for example. Based on experimentation, we generally synthesized the largest most common variables first.
- How to fit models and predict values? Any prediction method is possible, including econometric approaches, tree-based methods such as Classification and Regression Trees (CART) and random forests, and other machine-learning methods. Several research papers suggested that CART and random forests are particularly effective. We found that random forests slightly outperformed CART.
- How many records to synthesize? We chose to synthesize approximately 800,000 records – about five times as many as are in the PUF – to make our weighting task easier.

We implemented our synthesis using the `synthimpute` Python package written by Max Ghenis. In addition we developed several test files using the R package `synthpop`.

After we constructed a synthetic file, we tested it to be sure that it would pass the non-disclosure requirement of the IRS Statistics of Income (SOI) branch. The requirement was that no synthesized record may match any unique PUF record, exactly, on every synthesized variable. Some records in the PUF are extremely simple, consisting mostly of zeros for

[3] This research group developed the R package `synthpop`, which we used for elements of this project.

almost all synthesized variables, and our prediction methods matched some of these records exactly. We investigated several approaches to eliminating these records such as increasing the number of leaves in CART trees and adding noise to predicted values. Ultimately we concluded that the records matched were simple, unimportant, and relatively few and so we dropped them, ensuring no violation of the requirement. SOI required expert certification that this requirement was met, which we obtained. SOI's acknowledgement that the file meets their disclosure requirement is not an endorsement of the data or its use for any particular purpose.

At this point, the synthetic file did not yet have record weights. After evaluating the quality of the unweighted synthetic file, we constructed weights and evaluated the weighted file.

2. Evaluating quality of the synthetic file

We evaluated the unweighted synthetic file by comparing it to the actual PUF. Before making these comparisons, we calculated adjusted gross income from the synthesized components of income to allow comparison to AGI in the PUF.[4] The most important comparisons we made were for:

- Summary statistics for individual variables, such as the percent of values that are zero, and the mean, median, standard deviation, skewness, and kurtosis. Most of the differences were small.
- Graphical comparisons of the distributions of variables in the two files. Figure 1 shows kernel density plots for four large income items and Figure 2 shows similar plots for four large deduction components. In all cases even though the distributions vary across variables the files are quite similar to each other. There is room for improvement, particularly for AGI and wages.
- Correlations between variables. In general, the correlations between pairs of variables in the synthetic file were quite close to correlations between the same pairs in the PUF, but there were some exceptions. Out of approximately 1,600 correlations, 10 differed by more than 0.115 points. The worst offender was the correlation between AGI and net long-term gains, which was 0.621 in the PUF and only 0.178 in the synthetic file, a difference of 0.442 points. The next largest correlation difference was 0.237 points and the remaining 8 large differences were all smaller. Most of these 8 large-difference correlations involved one or more variables that are unlikely to be major items of interest for many tax policy analysts (examples include the alternative tax foreign tax credit and the domestic production activities deduction).

[4] We calculated AGI and selected other variables under 2011 law using NBER's taxsim tax calculator because the earliest tax year available in the PSL tax-calculator is 2013.

We learned many lessons from this analysis and in future syntheses we expect to improve upon the results from this initial file.

3. Weighting the synthetic file

After developing a satisfactory unweighted synthetic file we constructed record weights designed to produce totals similar to those calculated from the PUF.

We divided each file into mutually exclusive subsets by income range and marital status and chose weights for each record in each synthetic-file subset so that aggregate weighted values for targeted variables were close to corresponding values in the PUF subset. We targeted up to 128 values in each of 124 subsets for a total of approximately 16,000 targets. The variables we targeted and how we targeted them varied slightly by subset but generally for each variable we targeted the total number of returns with positive values (e.g., capital gains income or the medical expense deduction), the total number of returns with negative values, the sum of weighted positive values, and the sum of weighted negative values.

We chose weights in each subset that minimized a penalty function based on the squared difference between the summed weighted values in the synthetic file and the corresponding targets from the PUF, added up over all targets in the subset. It was not always possible to choose weights that made each difference zero, but we generally came close. To ensure that variables we considered especially important in tax analysis were favored when it was not possible to hit all targets, we assigned a priority factor of 100 for targets that involved the number of returns, AGI, wages, or tax before credits, and a priority factor of 1 for other variables. The results are quite good, but we believe there is room to improve our weighting procedure in future work.

4. Evaluating quality of the weighted file

We evaluated the weighed synthetic file in two ways: (1) we compared weighted sums by income range and marital status to corresponding sums from the PUF, and (2) we compared revenue and distributional results of tax policy analyses conducted with the synthetic file and with the PUF. We had to enhance the synthetic file slightly for these analyses to ensure that it was compatible with Tax-Calculator, and we made parallel modifications to the PUF so that we were comparing apples to apples.

4.1 Comparing the weighted 2011 synthetic file to the weighted 2011 PUF

Table 1 compares the weighted number of tax returns in our synthetic file to those in the modified PUF, by AGI range. The differences appear quite trivial, which is not surprising given

Figure 1. Distributions of important income variables in the synthetic file and in the PUF

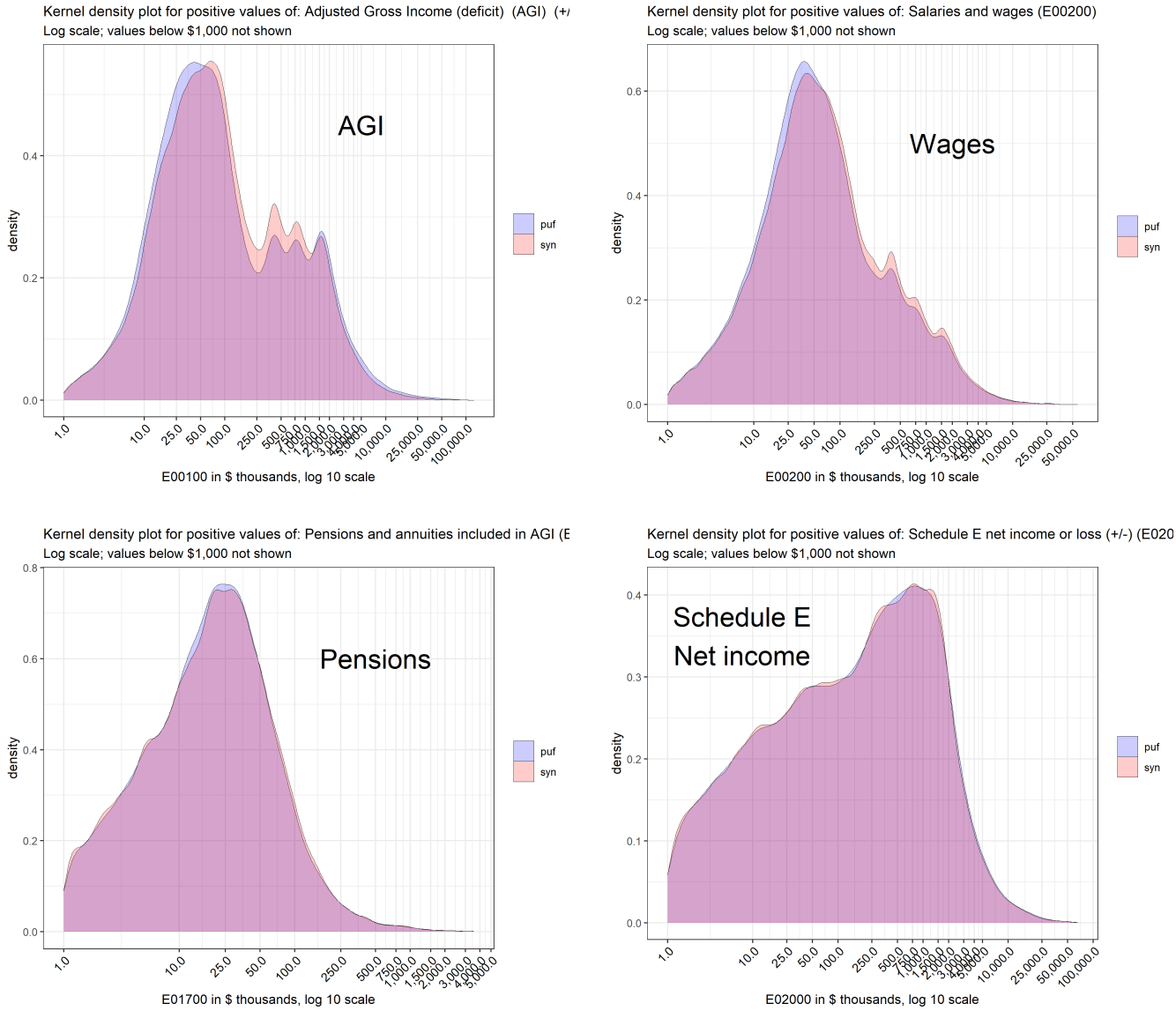
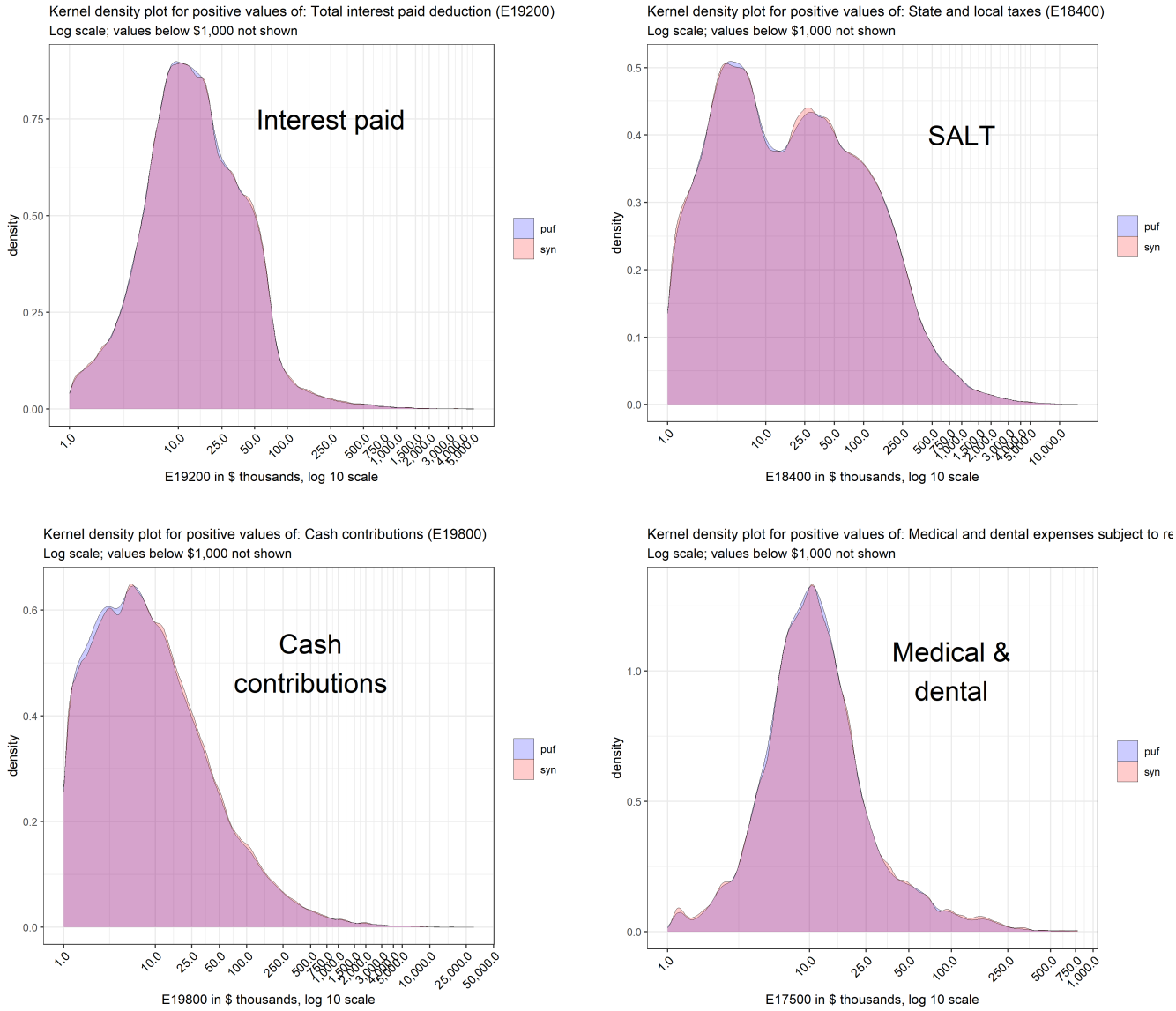


Figure 2. Distributions of important deduction variables in the synthetic file and in the PUF



that we gave return targets high priority when we weighted the file.

Table 1. Comparison of weighted number of returns in the synthetic file to number in the PUF

AGI Range	# of returns in millions			
	Modified PUF	Synthetic File	Diff.	% Diff.
Negative	1.4	1.4	0.0	-1.0
≥ $0 to < 25k	74.5	74.5	0.0	0.0
≥ $25k to < 50k	35.2	35.2	0.0	-0.1
≥ $50k to < 100k	31.9	31.9	0.0	-0.1
≥ $100k to < 200k	15.6	15.6	0.0	-0.1
≥ $200k to < 1m	4.7	4.7	0.0	1.0
≥ $1m	0.3	0.3	0.0	-0.1
Total	163.6	163.5	0.0	0.0

Table 2 compares the sum of weighted AGI in the two files by income range. The results are not quite as good but generally fall within 0.3 percent except for two income ranges. We consider this acceptable for a first effort.

Table 2. Comparison of weighted adjusted gross income in the synthetic file to the PUF

AGI Range	AGI in $ billions			
	Modified PUF	Synthetic File	Diff.	% Diff.
Negative	-62.0	-60.6	1.4	-2.2
≥ $0 to < 25k	760.4	760.5	0.2	0.0
≥ $25k to < 50k	1272.9	1271.6	-1.3	-0.1
≥ $50k to < 100k	2275.2	2272.1	-3.1	-0.1
≥ $100k to < 200k	2086.8	2081.5	-5.3	-0.3
≥ $200k to < 1m	1573.5	1589.4	15.9	1.0
≥ $1m	831.1	829.8	-1.4	-0.2
Total	8737.9	8744.3	6.4	0.1

4.2 Tax policy analysis with the synthetic file

To examine potential tax reforms we had to extrapolate the synthetic file forward from 2011 to 2013 because 2013 is the first data year that Tax-Calculator works with. We extrapolated the PUF in the same manner.

Table 3 shows our analysis of a simple across-the-board tax cut, versus 2017 law, on these two files. The first two numeric columns of the table show the estimated impacts in billions of dollars for each file. The third column shows the difference between the two estimates, also in billions of dollars, and the fourth column shows the percentage difference. The final two columns show the percentage tax cut in each file relative to its baseline value. The total tax cut in the PUF was $237 billion, or 21.0 percent. The synthetic PUF was extremely close, with a total tax cut of $237.6 billion, also 21.0 percent. The differences across income ranges generally are minor. This across-the-board cut was an easy test for the synthetic file.

Table 4 shows the results for a more-complex tax reform that created winners and losers. It eliminated all standard and itemized deductions, decreased regular income tax rates, and increased capital gains tax rates and pass-through-income tax rates. This was much more difficult for the synthetic file. Although the results on the bottom line were within 0.9 percent, the results for individual income ranges, particularly the two highest income ranges, are not as good. However, the percentage changes from baseline in the two rightmost columns are quite close. When we drilled down into different filing statuses, some income ranges had larger discrepancies.

I draw two lessons from this analysis. First, our synthetic file performs very well for plain-vanilla reforms, but is less faithful to the PUF if we examine impacts of complex reforms on small subsets of taxpayers. It is important to have this in mind when using the data and important for us to provide guidance to potential users, to help them understand what analyses the file is best for, and what it is least suited for. Second, in conducting this analysis we learned a great deal about how we can improve the file and we expect to do so in future iterations.

5. Conclusions and next steps

We have produced a synthetic file that is useful for some kinds of analyses. I believe it can be extremely useful in university public policy programs and other activities in which students and policy analysts are learning how to structure and examine reforms, where it is not essential to have results that are precise enough in their details for policymaking. We plan to make the file available for use with Tax-Calculator and TaxBrain later this year, with these audiences fore in our minds. We have learned many lessons about how to improve upon this initial effort. Our next effort will be useful for more policy analyses than the current file. In addition, we want to create state-specific synthetic files from a national synthetic file. These files could be used by policy analysts in individual states to study the impacts of federal income tax reforms on their states and, with adaptation, to examine state income taxes. We currently are raising funds for improvements and enhancements along these lines.

Modeling Notes

We used the `synthimpute` python package to generate a synthetic version of the Public Use File, estimating relationships and predicting values using random forests with 200 trees. For comparative purposes, we also developed several test files with the R package `synthpop` using CART methods. We found that random forests outperformed CART slightly, primarily in areas of the data that had relatively few returns.

We constructed the synthetic file by choosing weights that minimized the sum of squared differences between targets based on the PUF and the corresponding sum of weighted

Table 3. Impact of an across-the-board rate cut: Synthetic File vs. PUF
Across-the-board rate cuts compared with 2017 law as baseline, Regular tax before credits, $ billions

AGI Range	Estimated Impacts				% change from baseline	
	Modified PUF	Synthetic File	Difference in Estimated Impacts	Difference as a % of PUF estimate	Modified PUF	Synthic File
Negative	0.0	0.0	0.0	N/A	N/A	N/A
$\geq$ \$0 to $<$ 25k	-6.0	-6.1	0.0	0.7	-45.6	-45.6
$\geq$ \$25k to $<$ 50k	-30.0	-30.0	0.0	-0.1	-39.3	-39.3
$\geq$ \$50k to $<$ 100k	-70.8	-71.0	-0.2	0.2	-33.2	-33.3
$\geq$ \$100k to $<$ 200k	-70.3	-70.4	-0.1	0.1	-26.5	-26.6
$\geq$ \$200k to $<$ 1m	-44.9	-45.2	-0.3	0.7	-14.0	-14.0
$\geq$ \$1m	-15.0	-15.0	0.0	0.1	-6.2	-6.2
Total	-237.0	-237.6	-0.6	0.2	-21.0	-21.0

Table 4. Impact of a complex tax reform: Synthetic File vs. PUF
Complex winners-losers reform compared with 2017 law as baseline, Regular tax before credits, $ billions

AGI Range	Estimated Impacts				% change from baseline	
	Modified PUF	Synthetic File	Difference in Estimated Impacts	Difference as a % of PUF estimate	Modified PUF	Synthic File
Negative	0.0	0.0	0.0	N/A	N/A	N/A
$\geq$ \$0 to $<$ 25k	-1.9	-2.0	0.0	1.9	-14.7	-14.9
$\geq$ \$25k to $<$ 50k	-16.3	-16.4	-0.1	0.6	-21.4	-21.6
$\geq$ \$50k to $<$ 100k	-40.8	-41.1	-0.3	0.8	-19.1	-19.3
$\geq$ \$100k to $<$ 200k	-41.2	-42.1	-0.9	2.1	-15.6	-15.9
$\geq$ \$200k to $<$ 1m	-11.7	-10.1	1.6	-13.3	-3.6	-3.1
$\geq$ \$1m	17.1	17.8	0.7	3.8	7.0	7.4
Total	-94.9	-94.0	0.9	-0.9	-8.4	-8.3

values in the synthetic file, with judgmentally chosen priority adjustments for each difference as noted in the text. We used the R package `nloptr` to choose the weights. We encountered some challenges in minimizing this function and found that the method-of-moving-asymptotes (MMA) algorithm worked well, with the number of iterations limited to 500.