Creating a Single-State Microdata Tax File Compatible with TaxData

Project report to Michael Strain, Alex Brill, and Matt Jensen
The American Enterprise Institute
Report prepared per Strain-Kaloyeros agreement
(dated September 3, 2019, signed September 17, 2019)

December 31, 2019

Don Boyd, Consultant

Yimeng Yin, Economic Researcher & Gang Chen, Principal Investigator
The Rockefeller College of Public Affairs
The Research Foundation for the State University of New York

Introduction and goals

Our goals for this project were to (1) create a single-state microdata tax file compatible with TaxData, (2) port the approach that produces the best results from R to Python so that it can be integrated with TaxData, (3) rigorously compare file quality under different approaches, and (4) conduct the work in a well-documented and transparent manner such that it can be extended to all fifty states.

This report gives our main conclusions about the single-state file we constructed for New York, approaches to creating 50-state files, areas for potential improvement, and lessons for AEI if it wishes to construct a 50-state file by apportioning weights from the PUF to individual states. The appendices contain further details.

Our single-state microdata tax file for New York

We have created a single-state microdata tax file for New York that is compatible with TaxData. It is in this password-protected Dropbox directory; we have provided Matt Jensen with the password. The file has all variables needed by TaxData, although the best method for applying TaxData enhancements will depend upon AEI goals, as discussed below. The file is targeted to data published by the IRS Statistics of Income branch in Historical Table 2 for New York for 2017, the latest available year.[1] The file

was derived by growing the national PUF sample to 2017 and reweighting it to resemble New York.

We constructed record weights in two steps: First, we scaled the national weights to the state level with proportionate adjustments that varied by marital status and income range so that the number of returns calculated with these scaled weights matched Historical Table 2 totals for New York in 2017 by marital status and income range. Second, we adjusted these scaled weights so that the file hit or came close to approximately 500 targets from Historical Table 2, while minimizing an objective function that penalized large changes in the scaled weights. In general, we tried to come within 0.5 percent of each target for high-priority variables (AGI, wages, real estate tax deductions, total income tax before credits, and the AMT), and for other variables we tried to reduce the discrepancy between the target and the calculated file value by 90 percent of the initial discrepancy calculated with the scaled weights. The details of the method are described in an appendix.

The resulting New York microdata tax file hits or comes close to important targeted values, although some variables, particularly small variables that are not prioritized, may be far from published aggregate values.

The appendix section *Comparison of the New York microdata tax file to Historical Table 2* provides detailed tables for seven variables that compare weighted file values and numbers of returns with nonzero values, by income range, to corresponding Historical Table 2 targets. The key conclusions from these tables are (1) we came extremely close to the targets for high-priority variables, (2) we came reasonably close to the targets for taxable interest income (a large variable that we did not prioritize), and (3) we are very far from the targets for unemployment insurance compensation, a small idiosyncratic variable that we did not prioritize. We speculate on reasons for this in the appendix but did not investigate the reasons. In general, we believe it is possible to improve upon our preliminary results.

The R code that created this file is in this GitHub repository. We have ported from R to Python the optimization code needed to create this file from a national PUF that has been grown to 2017.

We have also created a version of the New York microdata tax file that can be used as input to Tax-Calculator by making necessary changes to variable names and constructing simple versions of additional variables required by Tax-Calculator. We ran the Trump 2017 reform proposal on this file and compared it to 2017 law, to show the difference between national and New York state results. Our New York file shows that in

several income ranges a much greater percentage of New York tax returns would have tax increases under this tax cut than in the U.S. as a whole. This is consistent with what we would expect and with what had been estimated or assumed elsewhere, because the combination of the SALT cap, New York's high taxes, and New York's high number of itemizers means more New Yorkers would have faced tax increases. A table in the appendix section *Running a tax reform on the New York microdata tax file* shows the percentage of taxpayers with tax increases in the U.S. and in New York by AGI range.

Possible alternative approaches

The second major part of this project examined approaches to creating multi-state microdata tax files with a focus on the difference between creating stand-alone single-state files such as the New York PUF we created, and approaches that constrain state record weights so that every national record effectively is "shared" exactly among the states. The distinction between creating a set of stand-alone single-state files and files constrained to the national file is important.

If the PUF included the universe of all returns in the nation or were a rich sample drawn representatively from each state, then we might be able to subdivide the national file to produce 50 separate state files that represent each state accurately, by reverse-engineering which records are from which state. Unfortunately, the PUF sample was not stratified by state and is not as detailed as the non-public IRS data that generated Historical Table 2 summary statistics for states. The single-state approach we used for New York does not identify or estimate which PUF records are from New York but rather chooses state weights for every PUF record (which may be from anywhere in the nation) in a way that is consistent with what we know about New York from other data. If this single-state approach were used for every state, we might use more than 100 percent of some records and less than 100 percent of other records when developing the best-possible file for each state.

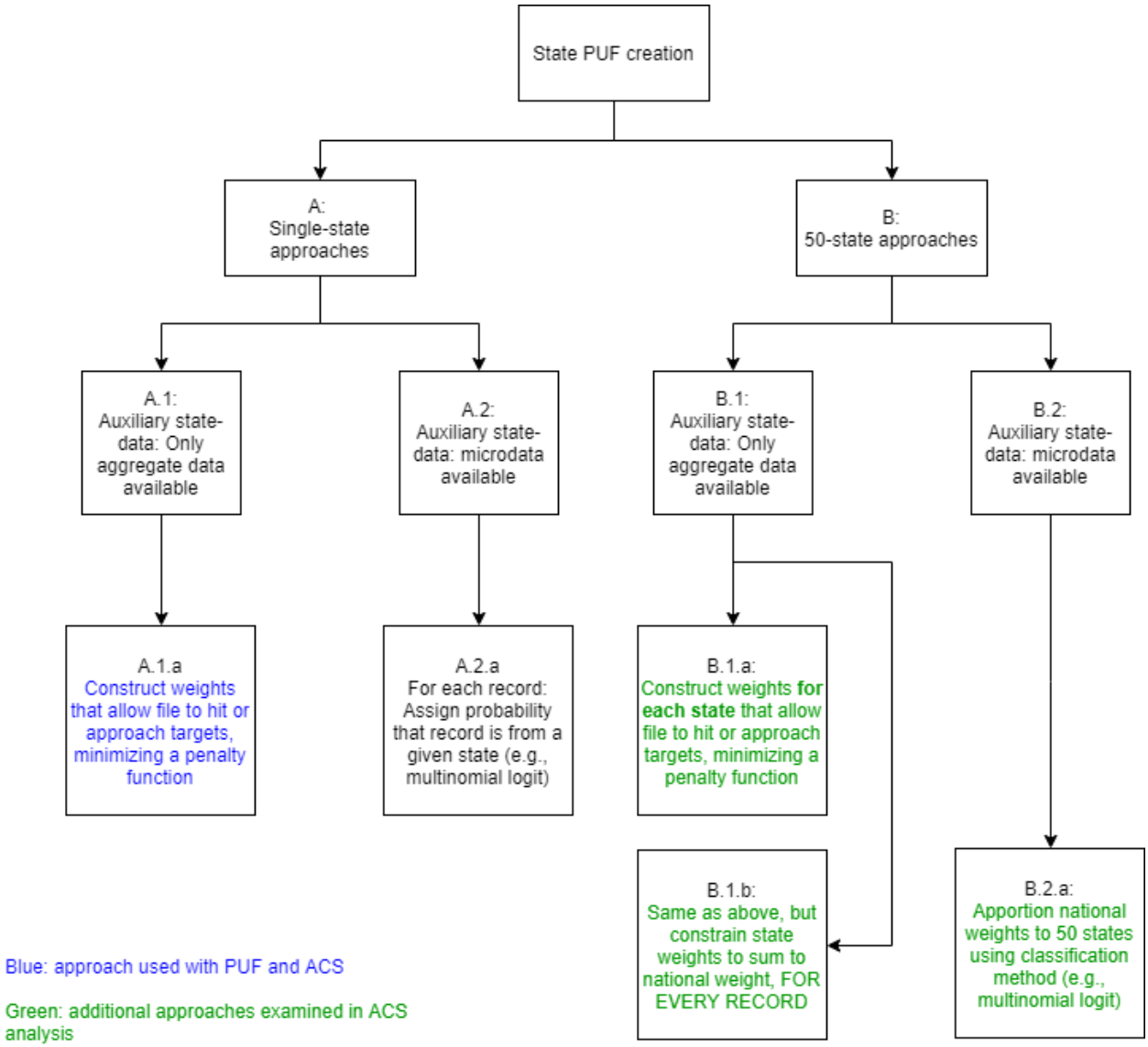
In the stand-alone approach a state file can be developed that fits known and estimated state data as well as possible, without regard to the state's relationship to the other 49 states. Every record in the national file can be used as much or as little as needed to attain the best fit with known or estimated data for the state. It is the approach a policymaker in a state would want if creating a file to analyze state tax policies or the impact of federal changes on the state, if not analyzing how their state is affected by policies in relation to all other states.

In the constrained approach every PUF record must be allocated completely, with nothing left over or more than fully used. *This is important if AEI's primary goal is to analyze a federal tax reform and allocate the results of a national analysis to the 50 states completely, ensuring that the sum of the analyses for 50 states is always consistent with the national analysis.* Within the constrained approach, we make a distinction between (1) methods that choose state record weights by targeting specific state values while constraining state weights for each record to sum to the national weight for the record, and (2) methods that do not target state values but instead attribute a fraction of each record to each state, by predicting each record's probability of being from each state or by predicting each record's share of the national total, implicitly constraining state weights for each record to sum to the national weight.

The method of attributing fractions of each record's national weight to the 50 states might be implemented by predicting state probabilities with a multinomial logit model or by estimating relative frequencies. For an implementation of the relative frequencies approach, see Fisher, Robin, and Emily Y Lin (2015)[2]. Either approach requires an auxiliary sample of tax microdata with which to estimate probability models or relative frequencies. We do not currently have access to such data, but they do exist in the IRS and were used in the Fisher-Lin paper. Therefore, a variant of this approach could become feasible in the future and the method deserves evaluation.

Figure 1 below illustrates how we categorize methods. We used method A.1.a to create the New York PUF. Extended to all states, it becomes method B.1.a. Extended to all states with constraints that force the weights for individual states to sum to the national weight for every record, it becomes method B.1.b. The alternative approach of apportioning national weights to the 50 states, without targeting, is method B.2.a.

Figure 1 A categorization of approaches to creating state PUFs



We would expect the single-state approach (A.1.a) to yield the highest quality file for any given state, but that when extended to 50 states (B.1.a) the results of, for example, analyzing a particular tax reform would not necessarily add precisely to national results because the state weights would not add precisely to national weights for every record. If we force state weights to sum to national weights (B.1.b), we might have to sacrifice file quality in some areas of the data - we might find it harder to hit some state targets precisely, or we might find that file quality degrades in other ways.

If we take a completely different approach and directly apportion weights to states via a multinomial logit model or other classification or probabilistic approach (method B.2.a, assuming we have microdata available to build such a model), we might find that the

data are more faithful to details of the underlying microdata, at the same time that they may not be faithful to totals that are important to us in tax policy analysis.

Creating a laboratory environment to compare alternative approaches

How can we compare these different methods and assess their effects on file quality? We do not have state tax microdata available that are analogous to the PUF and so we cannot use those data to examine details of results or to build models. However, we can build an artificial laboratory environment in which we know "true" microdata and we can use this environment to examine methods.

We constructed such a laboratory environment using the American Community Survey (ACS) (code is here). The advantage of this approach is that we know the "true" microdata for a state when we use the ACS, whereas when we use the PUF we only have access to aggregate SOI Historical Table 2 summaries for a state. Thus, our ACS laboratory environment allows us to compare results from different estimation methods to true ACS microdata for a state. *The sole purpose of this ACS analysis is to gain insights about pros and cons of alternative methods - we are not using the ACS to construct data we might use.*

To build this environment, we started with the 2013-2017 5-year ACS, and selected 5 states of interest: California, Florida, Illinois, New York, and Texas. (We wanted the states to be reasonably diverse but beyond that the specific states were not of interest: our interest in this part of the project is in comparing methods, not in comparing or examining the states themselves.) We drew a random sample of 50,000 records for individuals aged 18 or older from these 5 states and included data for their person id, state id, record weight, age, sex, and marital status, plus 7 different income items. Thus, we had a data file with records for 50,000 records for whom we know their state, plus the other variables - enough data to allow us to explore different approaches, and small enough to work with quickly and efficiently.

We used this data set as follows:

- For the targeting methods (B.1.a or B.1.b):
    - We constructed targets for each state by summarizing weights and weighted amounts by income range. For example, one of our targets might be the number of people in California in the $25-50k total income range who have nonzero Social Security income, and another target might be the total value of public assistance income for people in Texas with

$10-25k of total income. (We constructed the same set of targets for each state, of course.)
    - We then pretended we did not know which records were from which state, and constructed a state file for each state by choosing a weight for each of the 50,000 records that allowed the file, when weighted, to hit or come close to the targets for the state. In method B.1.a, we chose these weights for each state in isolation. In method B.1.b, we constrained the state weights for each person to add up to the person's total weight. We used an optimization approach similar to that for the New York PUF, as detailed in the appendix.
- For the methods that required microdata (B.2.a):
    - Using the 50,000 microdata records, we built a multinomial logit model to predict the probability that each record was from each state. In this approach, for any person the probabilities for the 5 states sum to 1.
    - We then pretended we did not know which records were from which state and constructed a state file for each state by multiplying the weight for each of the 50,000 records by the relevant probability for each state. The resulting state weights for each person will sum to their total person weight.

We then compared the resulting state files with the true state data.

Key conclusions from the ACS analysis

Our key preliminary conclusions from this work with the ACS are:

- The overall quality of the state files when using approaches that explicitly target known aggregates for a state (B.1.a and B.1.b) is much higher than when using the micro-data-based classification approach (B.2.a) that simply predicts probabilities without targeting known aggregates. Table 1 below, for example, shows the true distributions (percentiles) of total personal income in New York, and how the distribution of the national file (column 3) and the distributions produced by the three reweighting approaches (columns 4-6) compare to the true data. The percentiles produced by the two target-seeking approaches are very close to the true data (< 1% difference) while the percentiles from the classification approach differ from the target by a much larger margin.
- Comparing the results of B.1.a and B.1.b shows that incorporating constraints on the sum of record weights across state files into the target-seeking approach causes minimal deterioration in the quality of the resulting state files. Under certain circumstances, adding the record-weight constraints can even improve the quality of the state file (see the 25th and 50th percentiles in Table 1 below).

Table 1 Summary of performance of reweighting approaches

Comparing the performance of reweighting approaches on sample ACS data
Target variable: average personal income in New York

		% Difference from target, by weighting method			
Percentile	**Target: Average personal income**	**Initial ratio-adjusted state weights**	**B.1.a Stand-alone targeting**	**B.1.b Targeting with adding-up constraint**	**B.2.a Multinomial logit**
10th	$0	0.0	0.0	0.0	0.0
25th	8,496	9.3	0.2	0.0	8.6
50th	24,107	1.8	0.5	0.2	2.5
75th	51,463	-3.1	-0.4	-0.5	-1.6
90th	91,860	-9.2	0.0	0.0	-3.3

Details on weighting methods:

B.1.a: Stand-alone targeting. State aggregates are targeted. Each record's state weights are not constrained to its total weight.
B.1.b: Targeting with adding-up constraint. State aggregates are targeted. Each record's state weights must add to its total weight.
B.2.a: Multinomial logit. State probabilities are modeled but state aggregates are not targeted. Each record's state weights sum to its total weight.

- Although the micro-data-based classification approach produces state files with lower overall quality, we found that it has an advantage in approximating the state-specific characteristics of variables that are not explicitly targeted by the target-seeking approaches.

  Table 2 below summarizes the results of the three approaches for public assistance, which we intentionally did not target in the two targeting approaches. As the receipts of public assistance depend greatly on state-specific policies, the share of weighted records with positive public assistance and the average amount of public assistance vary greatly across the five states. The table shows that the classification approach based on the multinomial logit model we built better captured the variation across states than the two target-seeking approaches (except for Illinois). The better performance is largely attributable to the fact that the multinomial-logit-based classification, in contrast to the target-seeking approaches, assigns more weight to records that are truly from the target state when constructing state files from a national file, allowing it to improve the quality of all variables rather than focusing on optimizing the targeted variables. (The multinomial approach is fleshed out here. We are happy to discuss details.)

Table 2 Performance of reweighting approaches on a non-targeted variable (public assistance)

	Percentage of records (weighted) with positive public assistance				Average amount of public assistance (positive values only)			
		Difference from target (percentage points)				% Difference from target		
State	Target	B.1.a Stand-alone targeting	B.1.b Targeting with adding-up constraint	B.2.a Multinomial logit	Target ($)	B.1.a Stand-alone targeting	B.1.b Targeting with adding-up constraint	B.2.a Multinomial logit
CA	2.0	(0.3)	(0.3)	0.0	$4,548	(21.9)	(21.9)	(7.7)
FL	1.3	0.3	0.3	(0.1)	2,130	60.6	60.5	13.5
IL	1.8	(0.4)	(0.4)	(0.4)	2,660	28.2	28.3	31.1
NY	2.0	(0.4)	(0.4)	(0.2)	3,460	0.3	0.3	11.6
TX	0.8	0.8	0.7	0.3	2,481	41.1	41.0	(16.3)

Details on weighting methods:

B.1.a: Stand-alone targeting. State aggregates are targeted. Each record's state weights are not constrained to its total weight.
B.1.b: Targeting with adding-up constraint. State aggregates are targeted. Each record's state weights must add to its total weight.
B.2.a: Multinomial logit. State probabilities are modeled but state aggregates are not targeted. Each record's state weights sum to its total weight.

We caution that our conclusions are specific to the data and methods we used. We have not yet examined the Fisher/Lin method mentioned earlier.

Potential areas for improvement and further study

This work is preliminary and there are several areas for potential improvement:

- Reweighting vs. weights from scratch: We call the method in which we establish initial weights (the ratio-adjusted weights), and then adjust them to come close to targets, “reweighting”. An alternative potential method is to create “weights from scratch,” in which we do not establish initial weights but simply choose a set of weights that satisfy the known targets. We have experimented with these two methods in other contexts and do not yet have a formal conclusion on when one would be better than the other. Our intuition is that in this context, reweighting is better because we believe there is information of value on relationships among records and variables in our initial weights, which are based on the PUF and that it makes sense to penalize large differences. That said, this is an area worth exploration. Constructing weights from scratch is computationally more intensive than reweighting.
- Simulated annealing as an alternative weighting method: All the reweighting approaches we have used tend to produce weights that are nonzero – every PUF record is included in every state. By contrast, in the real world, only a small fraction of PUF records would be from any given state. Alternative approaches might choose whether a weight is from a given state and if so, what the weight should be. One such method we have experimented with is simulated annealing.

It produces weights that are either zero or a specific non-zero amount, much like true weights would be if we knew them. This is intuitively appealing, but we have no idea at this point whether it would produce a better file, and it is worth further study. Simulated annealing is computationally more intensive than what we have done.

- Choice of penalty function: We use a penalty function that is often used for reweighting microdata files and has been used in some contexts by the U.S. Treasury. It has conceptually appealing properties and has the practical value of being continuous, twice differentiable, and quickly solvable. It differs from the function used in TaxData and from other possible objective functions. It would be useful to examine alternative penalty functions. Our intuition is that this is not likely to have major impacts on results.
- Tolerance setting: This is probably the most important area of potential improvement. We established tolerances around each constraint target that are based on a variable's priority: +/- 0.5% for high-priority variables, and 10% of the initial percentage difference for low-priority variables so that they will reduce the initial discrepancy by 90% (for example, a low-priority variable that initially was 75 percent away from the target would have to be brought within 7.5 percent of the target). We believe this could be improved by running the optimization multiple times, starting with tight tolerances and loosening tolerances only for targets that are extremely hard to hit. We believe this is likely to result in a higher-quality file.
- Extend our ACS comparative framework to use noisy test data: In our ACS framework, the targets we set for individual states are entirely consistent with the underlying microdata we use to try to hit those targets (because the targets are constructed from the underlying data). This is different from and easier than the problem we face with the PUF, where the PUF sample is drawn from slightly different data than are used to construct Historical Table 2, and targets can, conceivably, be inconsistent with the PUF or quite hard to hit. For some tests, we may want to make our ACS laboratory environment more like the PUF situation by introducing noise into the ACS targets.
- Analyze the Fisher-Lin approach and other alternatives: We can use our ACS comparative framework to examine other approaches. The most important issue to examine, in our opinion, is how would file quality be affected in the "B.2.b" (see Figure 1) approach where we build a model of the probabilities or relative shares of each record that should be attributed to different states if we use an approach other than multinomial logit modeling. For example, how would file quality be affected if we used the Fisher-Lin relative frequencies approach? How would it be affected if we used machine learning (CART/RF/other) classification methods? (Implementing any of these methods for the PUF would require access

to state microdata that is reasonably similar to the PUF microdata we have nationally. This is not currently possible with the PUF.)

- There may be ways to combine the target-seeking and model-building approaches: We have not given this enough thought yet.

Lessons learned and implications for AEI

Our preliminary New York file shows that we can produce a high-quality single-state file by reweighting the national PUF to hit Historical Table 2 targets, but that it will have difficulty in hitting some targets.

Our ACS analysis suggests that if AEI wants a 50-state file that sums precisely to the national file, it probably will be better to (a) extend the New York target-seeking approach to all 50 states (and other IRS areas) and constrain the sum of state weights for each record to the national weight than to (b) use a multinomial logit approach (even if appropriate state-labeled microdata are available to build such a model).

We reach this conclusion tentatively because (a) the PUF problem is complex (500+ constraints per state) and large (approximately 8 million weights needed for 50 states) – far more so than our ACS analysis – and complexity and scale often lead to surprises, and (b) as noted earlier our ACS analysis was done without adding noise to the targets. We suspect the conclusion would hold if we added complexity and noise to our ACS analysis, but we have not yet done that.

As noted above, we have not examined alternative model-building and record-sharing approaches such as the Fisher-Lin approach, and it would make sense to explore additional approaches before committing firmly to a specific approach to a 50-state file.

Finally, there are additional issues to consider in moving from a single-state file to a 50-state file. We list three of the most-important issues below.

How should we balance file quality with the TaxData workflow? For example, one approach might be to match state PUF files with individual state CPS files to gain the highest possible quality in CPS enhancements. Another approach, far more compatible with the TaxData workflow, would be to apportion records after CPS matching has occurred. This is much simpler, but we suspect it will lead to low-quality CPS enhancements. It may be the best way to go in the near term, coupled with suggestions that users should not focus on state-estimates that are driven by CPS enhancements.

What should we do about file extrapolations? State economies change over time in different ways. For example, we know that Maine is aging far more rapidly than Utah is aging, and that Texas is growing far more rapidly than West Virginia, which has had population declines recently. It would be a substantial amount of work to try to capture these differences in a file extrapolated, say, to 2028, although there are some approaches that might capture reasonably expected changes in a practical way. Unless capturing these changes is crucial to an analysis, the best solution might be to estimate 50 state weights for each record for the most recent available data year (2017 as of now) and then maintain each state's share of each record's total weight in future years. It would make sense to caution users that estimated interstate policy-impact differentials might become less accurate as the number of years from the file base year increases.

What should we do if we just can't nail some states with confidence? With 50 very different states, some states may be very hard to hit given the relatively small size of the PUF sample and the requirement to constrain state weights to the national total for every record. If we find that quality is unacceptably low for some states, we might want to consider some combination of (a) providing warnings, (b) relaxing the adding-up constraint, (c) possibly combining states, or (d) other options.

All these issues and potential improvements deserve thought. However, our overall conclusion is that building a high-quality 50-state PUF is possible.

Appendix

Accessing the single-state microdata tax file for New York

We have placed two versions of a New York microdata tax file in this Dropbox directory, which has been shared with Matt Jensen. Both files are targeted to the latest SOI data year, 2017:

- "puf_ny_2017.csv", which is at 2017 income levels, follows the variable naming conventions of the SOI PUF, and is an appropriate starting point for enhancements that might be done by TaxData. It includes variables needed for TaxData and Tax-Calculator, plus: AGI_STUB, which indicates the 2017 AGI group a record falls into; wtus_2017, which is a weight that allows the file to hit U.S. total AGI in 2017 as reported in SOI Historical Table 2, and weight_state, which is the New York weight we constructed for the file.
- "puf_ny_2017_tcversion.csv", which follows Tax-Calculator naming conventions and input requirements and can be used as input to Tax-Calculator. The required prime-spouse variables were calculated with a 50-50 split, for the purpose simply of creating a file that is acceptable to Tax-Calculator. Ultimately, it would be desirable to have a state-file that reflects the full set of adjustments done by TaxData.

This file is useful for comparing results to IRS Historical Table 2 and other expectations, and for exploring differences in impacts between the nation and a state. However, it is not ready for real-world state-specific analysis: we have identified many issues in this report that could lead to a higher-quality file.

Creating a single-state microdata tax file for New York

This section describes how we created a single-state PUF – it fits in section A.1.a box of Figure 1.

In order to create a state data file, we need to have "targets" for what the file should look like. The best data available for targeting individual states are in SOI's Historical Table 2, which are available for more than 20 years in varying formats. They are available for every year from 2011 (the year of the PUF) through 2017 (latest available). In 2017, for each state and for the U.S. as a whole, these data provide the following information for each of 10 AGI ranges and for the state as a whole:

- Number of returns by marital status and in total
- Number of exemptions

- For each of approximately 66 continuous variables, including wages, interest income, dividends, net capital gains, major itemized deductions, tax before credits, and more:
    - Dollar amount of the variable
    - Number of returns for which the variable was nonzero
- In addition, the tables include counts for several indicators such as the number of farm returns and the number of returns that were filed electronically.

In aggregate, the Historical Table 2 data provide nearly 1,500 potential targets per state in 2017, although some are more important than others and we target a large subset rather than all values.

The data in Historical Table 2 are constructed by the IRS from a sample. The published U.S. totals for these data match the sums of the individual states from these data, but do not match exactly the U.S. totals that TaxData uses to target the PUF in future years, because TaxData targets are based on data summarized from the universe of returns rather than from a sample. Although the Historical Table 2 data do not match the universe exactly, they are close.

We decided to construct a NY PUF that targets the 2017 values directly rather than targeting interim years. We believe this was the most efficient way to do this. If there is a follow-on project, and if it is important to target earlier years, we should discuss this. We used the following steps:

1. Prepare PUF and Historical Table 2 data
    a. Determine PUF variables that are needed by Tax-Calculator and that we will need to grow to years beyond the 2011 PUF base year. Based on our experience creating synthetic data, we chose 71 variables to include.
    b. Construct a slimmed-down 2011 PUF that has only these variables
    c. Read and parse electronic versions of SOI Historical Table 2 for 2011 and 2017 to get values from these tables for each state. The two years are not in the same format, so we convert the 2011 table to the current format.
    d. For 2017 we used:
        i. 149 variables for each AGI range and geographic area, consisting of:
            - 132 targets for 66 continuous variables – in each case yielding a target for the amount (e.g., the amount of interest income in a given AGI range and state) and a target for the number of nonzero returns (e.g, number of returns with interest income), plus

- 17 other variables (e.g, number returns by marital status, number of exemptions)

ii. 11 income ranges (10 mutually exclusive ranges plus a total), and
iii. 53 geographic areas (50 states, DC, other areas, and U.S. total)

This yields a total of 1,490 potential targets per state (149 variables x 10 mutually exclusive income ranges). (Currently we are using about a third of these variables – targeting about 50 values for each of 10 income ranges, yielding slightly more than 500 targets per state.)

2. Prepare a national 2017 PUF from which we can make any state 2017 PUF:
    a. Compute per-return values and per-return national growth rates from the 2011 and 2017 Historical Table 2 data for approximately 66 variables. The growth rates are for the U.S. in aggregate (not by state and not by income range).
    b. Map these growth rates for Historical Table 2 variables to PUF variables to construct 2017 growfactors for the PUF variables.
    c. Apply these 2017 growfactors to all relevant variables on the 2011 PUF to create a 2017 national PUF for which components of income and deductions have been grown. (As with TaxData, each growfactor is specific to one or more PUF variables and is applied to all returns in the same way. Adjustments do not vary by return type or income range.) Note that at this point we do not yet have grown AGI, which must be calculated, nor do we have grown weights.
    d. Run this grown 2017 national file through Tax-Calculator using 2017 law to get calculated variables that will be needed when we target individual states. These calculated variables include AGI (c00100) and income tax before credits (c05800). This produces a file with 2017 income and deduction levels but that still has 2011 weights.
    e. Calculate the single growth rate needed, applied equally to all individual weights in the PUF, so that the sum of weighted AGI will equal the U.S. total AGI in Historical Table 2 for 2017.

    This results in a 2017 national PUF we can use to construct state 2017 PUFs. For any state for which we want a state-specific PUF for 2017, we will need to construct state-specific weights for each record.
3. To create a state-specific 2017 PUF, construct state-specific record weights for each record in a two-stage process:
    a. Construct initial state weights by proportionately adjusting national weights within each of the 10 AGI ranges and 4 filing statuses so that the weighted number of returns in each of these 40 subsets of the 2017 PUF equal the

number of returns reported in Historical Table 2 for the state. We call these initial weights “ratio-adjusted weights”.

b. Choose which values to target for each income range. Currently we are targeting 51 variables per income range, chosen judgmentally:
    i. The amounts for 23 large income, deduction, or tax variables
    ii. The number of returns that have nonzero values for these 23 variables
    iii. The number of returns for each of 4 return types
    iv. The number of personal exemptions
    v. This yields 510 targets per state.
c. Define tolerances for each of these targets (i.e., how close to these targets we will try to come).
    i. It will not be possible to hit each of these targets exactly, and for some it will not be possible to even come close. Therefore, set tolerances around each of these targets, judgmentally. The goal is to make these tolerances relatively tight, and to be tightest for the most important variables. This is an evolving area of our analysis and we expect to improve it further. The steps below describe our current method, but it will improve over time.
    ii. Define a set of high priority variables for which we know we want the tolerances to be extremely tight in each of the 10 income ranges. Currently, we have defined this as total AGI, total wages, total tax before credit, and total real estate taxes, as well as the number of personal exemptions and the number of returns in each of the 4 filing statuses.
    iii. Set tolerances for each group. Currently we use:
        - A 0.5% tolerance for the high-priority variables, and
        - 10% of the initial discrepancy as a tolerance for other variables.

        For example, suppose that using the initial “ratio-adjusted weights” for a state, the weighted sum for a variable not in our high-priority group such as the medical expense deduction is 75% different from the Historical Table 2 value for the state. In this case we would set a tolerance around this variable of 10% x 75%, or 7.5%. That is, we would like the optimization software to choose weights that get us within 7.5% of the target value (an improvement from the initial 75% discrepancy).

iv. We envision in the future running the optimization iteratively, adjusting tolerances with each iteration, until we can get the tolerances as close as practical.

d. Set up the optimization problem. This involves:
   i. Choosing an objective function to minimize - a function that penalizes large changes in weights relative to the “ratio-adjusted weights” in step 3.a above. We minimize the following function:

$$min \sum w_i \left(x_i^2 + x_i^{-2} - 2 \right)$$

Where:

- i indexes records (if there are 10,000 records in a given AGI range, then i runs from 1 to 10,000).
- w[i] is the ratio-adjusted weight (the initial weight) for record i, and
- x[i] is the ratio of the new weight that we wish to choose for record i to the initial weight - when it is 1, the function is at its minimum of zero; x is the vector of variables we solve for

In other words, we penalize new weights that are far in either direction from the initial weights, and we penalize a given difference by more if it is on a record with a large weight than if it is on a record with a small weight.

   ii. Setting constraint bounds using the tolerances above
   iii. Defining derivatives, the Jacobian, and the Hessian structure.
   iv. Scaling the problem for numerical stability

e. Run the optimization on each of the 10 AGI ranges.
f. Gather the 10 resulting sets of x vectors, multiply the x vectors by the initial ratio-adjusted weights to compute the new weights, and construct the state file.

We have ported Step 3, which is the hardest step, to Python.

Comparison of the New York microdata tax file to Historical Table 2

The first section below shows an excerpt of Historical Table 2 data; the remaining sections compare our New York file to Historical Table 2 targets, for each of several

variables. (The excerpted Table 2 targets match the corresponding Table 2 column in the comparison tables for variables shown in the excerpt.)

For each variable there are two tables: a comparison of the total value by AGI range and a comparison of the number of returns that have nonzero values. The variables fall into two broad groups:

1. High-priority variables we targeted with a 0.5 percent tolerance: adjusted gross income, salaries and wages, real estate tax deductions, income tax before credits, and the alternative minimum income tax.
2. Lower-priority variables where the tolerance was 10% of the discrepancy between the target and the initial file value: Taxable interest income and unemployment insurance compensation.

The tables below show that for the high-priority variables we hit our targets within the chosen tolerance.

For the two lower-priority variables, the results vary: The variances for taxable interest are all below 11 percent and many are much lower.

But for unemployment insurance the variances are much larger, particularly in the three lowest income ranges. And the number of returns with unemployment insurance is off by 18.9 percent, or 53.7 thousand returns. The initial values for unemployment insurance using our ratio-adjusted weights were very far from the targets; most were several hundred percent from the targets.

For example, in AGI group # 3 ($10k to <$25k), the New York target for unemployment insurance compensation in 2017 was $318.154 million as the screenshot below from the SOI spreadsheet shows.

Figure 2 Unemployment compensation excerpt from Historical Table 2 for New York, 2017

Table 2. Individual Income and Tax Data, by State and Size of Adjusted Gross Income, Tax Year 2017
[Money amounts are in thousands of dollars]

Item	All returns	Under $1 [1]	$1 under $10,000	$10,000 under $25,000
Unemployment compensation: [9] Number	284,400	640	13,140	69,560
Amount	1,431,098	3,800	48,604	318,154

However, the amount on the file using our initial ratio-adjusted weights was $1.32 billion, a difference of about 315 percent. Our optimally chosen record weights reduced

this discrepancy to 31.5 percent (10 percent of the initial 315 percent discrepancy), as the unemployment insurance table much further below shows, but it is still quite far. We suspect that with more effort we could improve upon this, but we would need to investigate reasons. There were major changes in the economy between 2011, the base year of our data, and our target year of 2017 and the problem might be that New York fared much better than the nation over this period. (The values on our file reflect national-average growth rates.) Another possibility is that there are definitional differences between unemployment insurance as reported on the PUF for 2011 and unemployment insurance as reported in Historical Table 2 for 2017. In general, we looked for these issues, but we did not have the resources or time to be comprehensive about this.

In any event, we suspect we could improve upon this with more work, but in the short term this would be an issue we'd raise to data users' attention so that they know now to give much credibility to analyses of unemployment insurance.

Historical Table 2 data

IRS SOI Historical Table 2 provides data by AGI range for the 50 states, the District of Columbia, other areas in the U.S., and the U.S. as a whole, for selected variables. The data for recent years are provided in two formats: csv files that are easy to work with, and spreadsheets that are easier to look at. The screenshot below is an excerpt of the 2017 spreadsheet for New York.

Figure 3 Excerpt from Historical Table 2 for New York, 2017

Table 2. Individual Income and Tax Data, by State and Size of Adjusted Gross Income, Tax Year 2017
[Money amounts are in thousands of dollars]

Item	All returns	Size of adjusted gross income									
		Under $1 [1]	$1 under $10,000	$10,000 under $25,000	$25,000 under $50,000	$50,000 under $75,000	$75,000 under $100,000	$100,000 under $200,000	$200,000 under $500,000	$500,000 under $1,000,000	$1,000,000 or more
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
NEW YORK											
Number of returns	**9,694,910**	**112,910**	**1,345,790**	**2,005,200**	**2,159,160**	**1,326,250**	**848,000**	**1,309,360**	**451,700**	**83,060**	**53,480**
Number of single returns	5,013,270	72,990	1,115,810	1,198,920	1,190,480	673,230	330,010	329,920	81,790	12,760	7,360
Number of joint returns	2,980,010	29,780	99,170	323,880	436,000	386,460	385,850	858,250	350,390	66,790	43,450
Number of head of household returns	1,486,660	5,810	112,170	450,950	477,470	223,760	105,780	93,660	13,470	2,230	1,370
Number of electronically filed returns	8,926,080	86,850	1,190,650	1,847,360	2,004,070	1,229,250	786,670	1,225,610	426,630	78,640	50,350
Number of computer prepared paper returns	389,510	13,450	77,110	85,780	79,180	45,570	27,540	40,960	14,850	2,930	2,150
Number of returns with paid preparer's signature	6,221,770	78,790	791,470	1,251,430	1,300,610	841,630	568,080	923,930	344,580	71,300	49,950
Number of returns with direct deposit	5,848,040	25,010	618,070	1,365,760	1,528,240	868,070	529,260	745,930	146,620	13,240	7,840
Number of exemptions	17,469,320	165,350	1,272,370	3,375,680	3,800,340	2,385,170	1,713,800	3,148,060	1,215,230	235,600	157,750
Number of dependent exemptions	5,360,410	30,830	245,440	1,168,070	1,220,090	674,370	480,560	981,080	413,330	85,820	60,820
Total number of volunteer prepared returns [2]	231,430	1,190	66,610	79,740	59,420	15,920	5,220	3,170	** 170	**	**
Number of volunteer income tax assistance (VITA) prepared returns	142,040	840	44,860	50,730	37,050	6,360	1,340	810	** 40	**	**
Number of military volunteer prepared returns	2,040	**	** 340	970	340	150	120	110	** 10	**	**
Number of tax counseling for the elderly (TCE) prepared returns	87,360	360	21,410	28,040	22,030	9,400	3,760	2,250	** 110	**	**
Number of volunteer prepared returns with earned income credit	43,930	50	18,360	17,810	7,670	** 40	0	0	0	**	**
Number of refund anticipation check returns [3]	1,212,290	2,790	140,040	386,830	378,240	162,350	71,870	64,330	5,630	160	40
Number of elderly returns [4]	2,288,700	47,890	258,770	394,020	427,970	339,430	254,110	395,110	130,260	24,470	16,680
Adjusted gross income (AGI) [5]	**836,755,149**	**-23,906,390**	**6,998,061**	**34,317,550**	**78,562,853**	**81,527,861**	**73,482,730**	**179,003,932**	**129,588,367**	**56,532,599**	**220,647,586**
Total income: [6] [7] Number	9,671,350	89,360	1,345,780	2,005,200	2,159,160	1,326,250	848,000	1,309,360	451,700	83,060	53,480
Amount	847,994,803	-23,730,466	7,238,458	34,928,052	79,579,879	82,474,271	74,201,091	180,935,718	131,745,862	57,786,501	222,835,437
Salaries and wages in AGI: Number	7,957,650	28,950	904,070	1,560,740	1,893,590	1,158,770	735,030	1,157,870	403,520	71,870	43,240
Amount	549,679,189	1,221,150	4,950,724	25,567,251	65,312,304	65,054,349	55,967,004	134,967,516	92,248,432	34,332,789	70,057,670
Taxable interest: Number	3,343,880	43,320	207,740	301,630	483,390	499,840	437,650	869,710	370,830	77,400	52,360
Amount	13,515,254	822,159	86,258	187,128	302,417	358,480	367,317	1,007,250	1,067,258	737,488	8,579,499

High-priority variables targeted with a 0.5 percent tolerance

Adjusted gross income: Amount and number of returns

AGI range	Adjusted gross income (AGI) (Billions of dollars) Historical Table 2	Calculated value	Difference	% Difference
Under $1	-23.9	-24.0	-0.0	0.2
$1 - < $10k	7.0	7.0	0.0	0.3
$10k - < $25k	34.3	34.2	-0.1	-0.3
$25k - < $50k	78.6	78.9	0.4	0.5
$50k - < $75k	81.5	81.9	0.4	0.5
$75k - < $100k	73.5	73.9	0.4	0.5
$100k - < $200k	179.0	179.9	0.9	0.5
$200k - < $500k	129.6	130.2	0.6	0.5
$500k - < $1m	56.5	56.8	0.3	0.5
$1m+	220.6	221.2	0.5	0.2
Total	836.8	840.1	3.4	0.4

AGI range	Number of returns with adjusted gross income (AGI) (Thousands of returns) Historical Table 2	Calculated value	Difference	% Difference
Under $1	89.4	89.0	-0.4	-0.4
$1 - < $10k	1,345.8	1,339.2	-6.6	-0.5
$10k - < $25k	2,005.2	1,995.3	-9.9	-0.5
$25k - < $50k	2,159.2	2,154.1	-5.0	-0.2
$50k - < $75k	1,326.2	1,328.5	2.2	0.2
$75k - < $100k	848.0	849.1	1.1	0.1
$100k - < $200k	1,309.4	1,306.1	-3.3	-0.3
$200k - < $500k	451.7	449.5	-2.2	-0.5
$500k - < $1m	83.1	83.4	0.4	0.4
$1m+	53.5	53.7	0.3	0.5
Total	9,671.4	9,647.9	-23.4	-0.2

Salaries and wages: Amount and number of returns

AGI range	Salaries and wages amount (Billions of dollars) Historical Table 2	Calculated value	Difference	% Difference
Under $1	1.2	1.2	0.0	0.3
$1 - < $10k	5.0	4.9	-0.0	-0.5
$10k - < $25k	25.6	25.6	0.1	0.2
$25k - < $50k	65.3	65.5	0.2	0.3
$50k - < $75k	65.1	65.1	0.1	0.1
$75k - < $100k	56.0	56.2	0.3	0.5
$100k - < $200k	135.0	135.6	0.7	0.5
$200k - < $500k	92.2	92.7	0.5	0.5
$500k - < $1m	34.3	34.2	-0.2	-0.5
$1m+	70.1	69.7	-0.4	-0.5
Total	549.7	550.9	1.2	0.2

AGI range	Number of returns with salaries and wages (Thousands of returns) Historical Table 2	Calculated value	Difference	% Difference
Under $1	28.9	28.9	-0.1	-0.2
$1 - < $10k	904.1	908.6	4.5	0.5
$10k - < $25k	1,560.7	1,555.6	-5.1	-0.3
$25k - < $50k	1,893.6	1,886.0	-7.6	-0.4
$50k - < $75k	1,158.8	1,154.6	-4.1	-0.4
$75k - < $100k	735.0	731.9	-3.1	-0.4
$100k - < $200k	1,157.9	1,152.9	-5.0	-0.4
$200k - < $500k	403.5	401.5	-2.0	-0.5
$500k - < $1m	71.9	71.9	0.0	0.1
$1m+	43.2	43.0	-0.2	-0.5
Total	7,957.6	7,935.0	-22.7	-0.3

Real estate tax itemized deductions: Amount and number of returns

AGI range	Real estate taxes amount (Billions of dollars) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.0	0.0	0.0	Inf
$1 - < $10k	0.2	0.2	-0.0	-0.5
$10k - < $25k	0.6	0.6	-0.0	-0.5
$25k - < $50k	1.5	1.5	-0.0	-0.5
$50k - < $75k	2.2	2.2	-0.0	-0.5
$75k - < $100k	2.5	2.5	-0.0	-0.5
$100k - < $200k	7.6	7.6	-0.0	-0.5
$200k - < $500k	4.9	4.9	-0.0	-0.5
$500k - < $1m	1.7	1.6	-0.0	-0.5
$1m+	2.3	2.3	-0.0	-0.5
Total	23.6	23.5	-0.1	-0.5

AGI range	Number of returns with real estate taxes (Thousands of returns) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.0	0.0	0.0	Inf
$1 - < $10k	31.3	31.4	0.2	0.5
$10k - < $25k	93.1	93.6	0.5	0.5
$25k - < $50k	250.4	251.6	1.3	0.5
$50k - < $75k	354.7	356.5	1.8	0.5
$75k - < $100k	358.7	360.5	1.8	0.5
$100k - < $200k	861.8	866.1	4.3	0.5
$200k - < $500k	355.6	357.4	1.8	0.5
$500k - < $1m	69.1	69.5	0.3	0.5
$1m+	47.4	47.7	0.2	0.5
Total	2,422.1	2,434.2	12.1	0.5

Income tax before credits (regular + AMT): Amount and number of returns

AGI range	Income tax before credits amount (Billions of dollars) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.0	0.0	-0.0	-100.0
$1 - < $10k	0.0	0.0	-0.0	-0.5
$10k - < $25k	0.8	0.8	-0.0	-0.5
$25k - < $50k	5.2	5.2	-0.0	-0.5
$50k - < $75k	8.0	7.9	-0.0	-0.5
$75k - < $100k	8.3	8.2	-0.0	-0.5
$100k - < $200k	23.9	23.8	-0.1	-0.5
$200k - < $500k	25.7	25.6	-0.1	-0.5
$500k - < $1m	14.3	14.2	-0.1	-0.5
$1m+	59.0	59.3	0.3	0.5
Total	145.3	145.1	-0.2	-0.1

AGI range	Number of returns with income tax before credits (Thousands of returns) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.9	0.0	-0.9	-100.0
$1 - < $10k	125.4	126.0	0.6	0.5
$10k - < $25k	1,249.4	1,243.2	-6.2	-0.5
$25k - < $50k	2,055.7	2,051.6	-4.1	-0.2
$50k - < $75k	1,313.6	1,310.9	-2.7	-0.2
$75k - < $100k	842.8	840.8	-2.1	-0.2
$100k - < $200k	1,304.6	1,299.2	-5.4	-0.4
$200k - < $500k	451.1	449.3	-1.8	-0.4
$500k - < $1m	83.0	83.4	0.4	0.5
$1m+	53.5	53.7	0.3	0.5
Total	7,480.1	7,458.2	-21.9	-0.3

Alternative minimum income tax: Amount and number of returns

AGI range	Alternative minimum tax amount (Billions of dollars) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.0	0.0	-0.0	-100.0
$1 - < $10k	0.0	0.0	-0.0	-100.0
$10k - < $25k	0.0	0.0	-0.0	-100.0
$25k - < $50k	0.0	0.0	-0.0	-89.2
$50k - < $75k	0.0	0.0	-0.0	-28.7
$75k - < $100k	0.0	0.0	-0.0	-0.5
$100k - < $200k	0.3	0.3	0.0	0.5
$200k - < $500k	2.5	2.5	-0.0	-0.5
$500k - < $1m	0.9	0.9	-0.0	-0.5
$1m+	1.6	1.6	-0.0	-0.5
Total	5.4	5.3	-0.1	-1.1

AGI range	Number of returns with alternative minimum tax (Thousands of returns) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.5	0.0	-0.5	-100.0
$1 - < $10k	0.1	0.0	-0.1	-100.0
$10k - < $25k	0.3	0.0	-0.3	-100.0
$25k - < $50k	0.4	0.5	0.0	0.5
$50k - < $75k	2.2	2.2	0.0	0.5
$75k - < $100k	10.5	10.6	0.1	0.5
$100k - < $200k	117.8	117.2	-0.6	-0.5
$200k - < $500k	368.4	366.6	-1.8	-0.5
$500k - < $1m	56.6	56.3	-0.3	-0.5
$1m+	12.8	12.8	-0.1	-0.5
Total	569.6	566.0	-3.6	-0.6

Lower-priority variables targeted for 90% improvement in discrepancy

Taxable interest income: Amount and number of returns

AGI range	Taxable interest amount (Billions of dollars) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.8	0.8	0.0	2.0
$1 - < $10k	0.1	0.1	0.0	7.6
$10k - < $25k	0.2	0.2	0.0	10.7
$25k - < $50k	0.3	0.3	0.0	9.3
$50k - < $75k	0.4	0.4	0.0	6.1
$75k - < $100k	0.4	0.4	0.0	5.1
$100k - < $200k	1.0	1.0	0.0	4.1
$200k - < $500k	1.1	1.1	0.0	2.6
$500k - < $1m	0.7	0.8	0.0	2.8
$1m+	8.6	8.1	-0.5	-5.5
Total	13.5	13.2	-0.3	-2.0

AGI range	Number of returns with taxable interest (Thousands of returns) Historical Table 2	Calculated value	Difference	% Difference
Under $1	43.3	44.3	1.0	2.4
$1 - < $10k	207.7	213.1	5.4	2.6
$10k - < $25k	301.6	308.0	6.4	2.1
$25k - < $50k	483.4	488.3	4.9	1.0
$50k - < $75k	499.8	501.2	1.4	0.3
$75k - < $100k	437.6	438.3	0.6	0.1
$100k - < $200k	869.7	870.5	0.8	0.1
$200k - < $500k	370.8	371.8	0.9	0.3
$500k - < $1m	77.4	77.5	0.1	0.1
$1m+	52.4	52.4	0.0	0.0
Total	3,343.9	3,365.5	21.6	0.6

Unemployment insurance compensation: Amount and number of returns

AGI range	Unemployment compensation amount (Billions of dollars) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.0	0.0	0.0	37.7
$1 - < $10k	0.0	0.1	0.0	86.8
$10k - < $25k	0.3	0.4	0.1	31.5
$25k - < $50k	0.4	0.4	-0.0	-8.9
$50k - < $75k	0.2	0.2	0.0	8.8
$75k - < $100k	0.1	0.1	-0.0	-5.9
$100k - < $200k	0.2	0.2	0.0	3.9
$200k - < $500k	0.0	0.1	0.0	14.0
$500k - < $1m	0.0	0.0	-0.0	-8.1
$1m+	0.0	0.0	0.0	20.3
Total	1.4	1.6	0.1	9.2

AGI range	Number of returns with unemployment compensation (Thousands of returns) Historical Table 2	Calculated value	Difference	% Difference
Under $1	0.6	0.9	0.2	35.6
$1 - < $10k	13.1	23.4	10.3	78.0
$10k - < $25k	69.6	87.7	18.2	26.1
$25k - < $50k	82.7	93.8	11.2	13.5
$50k - < $75k	43.4	48.6	5.2	12.1
$75k - < $100k	27.6	30.8	3.2	11.7
$100k - < $200k	38.4	42.7	4.3	11.3
$200k - < $500k	8.2	9.2	1.0	11.8
$500k - < $1m	0.7	0.7	0.1	12.9
$1m+	0.2	0.2	0.0	26.9
Total	284.4	338.1	53.7	18.9

Running a tax reform on the New York microdata tax file

This project has focused on methods for creating a state microdata file, and on creating a preliminary file for New York. Its purpose is not to use that file for real-world policy analysis, although that is a logical eventual goal. To get an early look at how such a file might be used in such an analysis, we ran the file through Tax-Calculator, comparing a plan that implemented the Trump 2017 income tax proposal, which preceded the TCJA, to 2017 law. Among other things, that plan capped the state and local tax deduction. This was one of several features that would lead to winners and losers and was likely to create a relatively greater number of losers in a high-tax high-itemizers state like New York, than in the nation as a whole.

The table below shows the percentage of tax returns that would face a tax increase (i.e., losers) in New York and in the nation. While this is hardly a comprehensive look at how our file would fare in real-world tax-policy analysis, it is certainly consistent with our expectations: we have relatively more losers in New York, particularly in income ranges with AGI of $75k or more. It was not practical in this project to explore tax reform analyses more extensively but if we develop a file that reflects each of the 50 states, it would be important to do this.

Percentage of returns with tax increase

AGI range	US	NY
Under $1	0.0	0.0
$1 - < $10k	0.1	0.1
$10k - < $25k	1.4	1.3
$25k - < $50k	5.1	5.1
$50k - < $75k	9.6	9.8
$75k - < $100k	14.4	17.0
$100k - < $200k	19.1	25.4
$200k - < $500k	19.7	14.8
$500k - < $1m	6.9	10.0
$1m+	18.4	21.5
Total	7.5	8.6

Additional notes on approaches to creating state PUFs

The problem of constructing a state microdata file when national microdata and auxiliary state data are available is not unique to tax data. Geographers, survey researchers, economists and others have faced this problem in other contexts. (For a review see Hermes, Kerstin, and Michael Poulsen. “A Review of Current Methods to Generate Synthetic Spatial Microdata Using Reweighting and Future Directions.” *Computers, Environment and Urban Systems* 36, no. 4 (July 2012): 281–90. https://doi.org/10.1016/j.compenvurbsys.2012.03.005.) A common approach to the problem is to use all the national records but adjust their weights so that the file, when summarized with the adjusted weights, is consistent with summary data for the region of interest.

Our task is particularly demanding for a few reasons:

- We want the state microdata file to be consistent with many state targets if possible (more than 500 such targets for a single state).

- The data used to construct the state-level summaries differs from the data used to create the national microdata file and so the national file is not completely consistent with state targets.
- In some scenarios we may have additional constraints we want to impose. In particular, we may want to have one constraint per record in the national file, requiring that the state weights based on that record must sum to the national weight for the record.
- As a result, some state targets may be difficult to hit.
- We may need to place tolerances around each target so that instead of attempting to hit some targets exactly, we aim to have our results fall within a range around the target.
- The problem can become quite large. For example, under one way of framing the problem, for a national tax microdata file with 160k records (the approximate size of the PUF) we could need to solve for 800k variables (50 weights for every record - 1 per state), with the following constraints:
    - 500 targets per state:
        - 50 targets per income range, times
        - 10 income ranges
    - and 50 states
    - yielding 25,000 targets
    - plus, an additional 160k constraints, one per record, requiring that the 50 state weights for every national record must sum to the national weight

  This creates a problem that could have 800k variables and 185k constraints. Furthermore, if we were to solve this problem for a synthetic PUF that has 5 times as many records as the IRS PUF, as we may want to in the future, the problem would have 4 million variables and 825k constraints (the 25k state targets, plus 800k adding-up constraints).

  Thus, it is important to pay close attention to software characteristics and to how the problem is set up, so that it is solved efficiently and in a way that does not cause numerical difficulties, and so that it takes advantage of opportunities for parallelism where appropriate.

Endnotes

[1] In recent weeks this link has not been working well and 2017 data are not easy to find. We suspect SOI will fix this shortly, but if you want the 2017 Historical Table 2 data sooner, please let us know.
[2] Fisher, Robin, and Emily Y Lin. “Re-Weighting to Produce State-Level Tax Microsimulation Estimates.” Technical Paper. United States Department of the Treasury, Office of Tax Analysis, June 2015. https://www.treasury.gov/resource-center/tax-policy/tax-analysis/Documents/TP-6.pdf.