# Data Processing for PWBM's Tax Module

John Ricco*

January 09, 2020

**Abstract**

Penn Wharton Budget Model's Tax Module consists of calculators for federal individual income taxes, payroll taxes, corporate income taxes, and estate taxes. Each calculator takes as input a simulated dataset of tax returns, which are used as the basis for calculating liabilities and projecting tax revenues. This paper details how PWBM combines public-use administrative records, survey data, and summary tables published by the IRS to simulate tax return microdata.

*Senior Analyst, Penn Wharton Budget Model. Email: jricco@wharton.upenn.edu

# Contents

# 1 Overview

A major component of the Penn Wharton Budget Model (PWBM) is an integrated tax microsimulation model known as the Tax Module. It consists of a set of integrated tax calculators that provide micro-level estimates of individual income tax, payroll tax, corporate income tax, and estate tax liabilities for representative samples of families, businesses, and decedents. It simulates "micro-dynamic" behavioral feedback—for example, business income shifting across legal entity type—expected to arise in response to changes in tax policy. The Tax Module forms the basis for PWBM's projections of federal tax revenues under current law and counterfactual tax reform proposals, and it calculates various tax functions used in PWBM's overlapping generations (OLG) dynamic model.

This paper explains the steps taken to prepare each dataset used for tax calculation in the Tax Module. For individual income and payroll taxes, we create an augmented version of the IRS's Statistics of Income Public Use File (PUF), a representative sample of actual tax returns filed with the IRS. However useful, the raw PUF has some shortcomings as a data source for the Tax Module. First, it becomes available for purchase on a substantial lag (usually about six years) and thus requires a method to "historically age" the data forward in accordance with more recent publicly available aggregate tax data. Second, the IRS uses an anti-disclosure aggregation procedure to mask the information of more than 1,300 tax returns containing extreme values; these "aggregate returns" need to be disaggregated to ensure statistical representation and create adequate heterogeneity at the top of the income distribution. And finally, the PUF lacks key demographic variables as well as information on those who do not file a tax return. To address this limitation, we implement a constrained statistical match which augments the PUF with data from the Current Population Survey (CPS).

Unfortunately, no analogous public-use microdata file exists for corporate income tax returns. PWBM instead uses the most recently available IRS summary tables to generate a dataset of synthetic tax returns. The resulting records are heterogeneous across industry,

size, and profit/loss status.

Similarly, there is no publicly available microdata source for federal estate taxes. The basis for PWBM's synthetic estate tax data is the Federal Reserve Board's Survey of Consumer Finances (SCF), augmented with wealth data from the Forbes 400 (who are excluded from the survey by design). We then simulate mortality risk (accounting for differentials across demographic characteristics and income) to arrive at expected estate tax records, making adjustments and imputations to align our estimates with actual asset and deduction values reported in IRS summary tables.

PWBM describes each of these procedures in more detail in the sections that follow. Section 2 focuses on how the individual PUF is processed, Section 3 describes the process for generating the synthetic corporate tax returns, and Section 4 details how we create data for the estate tax calculator. The paper concludes in Section 5 with a brief discussion on how we project the characteristics of future tax-filing populations in accordance with PWBM's demographic microsimulation model.

# 2 Individual income and payroll taxes

## 2.1 Distributing the aggregate returns

In order to protect taxpayer privacy, the IRS takes a number of disclosure-proofing steps when preparing the PUF. These include "blurring" certain income values and subsampling techniques that do not prevent the Tax Module from accurately and representatively calculating tax liabilities.[1] But one technique in particular presents a unique challenge: the aggregation of returns with extreme values. The IRS sorts the microdata by each variable and removes $N$ returns with the highest (and lowest, for variables with negative values) amounts.[2] Values from these returns are aggregated into one of four "aggregate returns",

[1] For a complete account of these procedures, see the most recent PUF documentation.

[2] According to the official documentation, $N$ is "generally" 30, but varies based on the variable. This is a weighted count, not a simple count of records.

depending on AGI group.[3] Running these records through the individual income tax calculator would result in nonsensical liabilities, and simply removing them from the dataset altogether would mean large portions of total income for some variables would be missing.

Fortunately, the IRS provides enough information about the nature of these returns for us to create a set of synthetic records which adequately represent the masked information. For each variable, the IRS reports the total dollar amount and the number of returns reporting nonzero values. These constitute a set of targets that can be used to restore statistical representation.

To create this set of synthetic returns, we first select the top $N$ returns by variable from the existing records in the PUF and copy them. Because some returns will contain extreme values for more than one variable, this initial selection contains duplicates; these duplicates are removed. The result is 134 records representing 1,362 returns, which are then split into groups by AGI according to the same rules as the aggregate returns. These records are used as the basis for distributing the values contained in the aggregate returns.

The first step is to rescale weights for these records to match weighted totals for each aggregate return. Table 1 shows the original weight, the target weight, and the implied rescaling factor.

Table 1: Number of synthetic returns, before and after rescaling

| AGI group | Number of synthetic returns | Number of returns in 'aggregate return' | Rescaling factor |
|---|---|---|---|
| Total | 1361.55 | 1260.77 | 0.93 |
| Negative AGI | 213.48 | 154.38 | 0.72 |
| $0 - $10M AGI | 131.91 | 247.39 | 1.88 |
| $10M - $100M AGI | 792.60 | 372.00 | 0.47 |
| $100M+ AGI | 223.56 | 487.00 | 2.18 |

Table 2 shows totals for selected variables in the synthetic returns and compares them to the actual totals found in the aggregate returns. As expected, dollar amount totals are substantially smaller than what appears on the aggregate returns—these are returns with

[3]The groups are: negative AGI; $0 ≤ AGI < $10M; $10M ≤ AGI < $100M; AGI > $100M+.

definitionally smaller values. Return counts are broadly similar, which suggests that this set of returns shares the characteristics of the aggregate returns' joint distributions.

Table 2: Selected aggregates from the synthetic returns before adjustments

| | Amount | | Number of returns | |
|---|---|---|---|---|
| Variable | Actual | Synthetic returns | Actual | Synthetic returns |
| AGI income | 161.6 | 71.8 | 1106 | 1106 |
| AGI loss | -11.5 | -3.7 | 154 | 154 |
| Wages and salaries | 13.2 | 2.7 | 921 | 918 |
| Qualified dividend income | 22.7 | 12.5 | 1142 | 1062 |
| Sole proprietorship income | 2.3 | 0.7 | 249 | 146 |
| Sole proprietorship loss | -1.1 | -0.5 | 220 | 239 |
| Net capital gain | 88.5 | 47.6 | 903 | 872 |
| Schedule E net income | 30.4 | 9.0 | 668 | 573 |
| Schedule E net loss | -12.8 | -3.8 | 496 | 545 |
| Gross social security benefits | 0.0 | 0.0 | 346 | 335 |
| State and local income/sales tax deduction | 7.1 | 2.4 | 1066 | 1024 |
| Real estate tax deduction | 0.2 | 0.1 | 1031 | 994 |
| Interest paid deduction | 2.1 | 0.7 | 901 | 804 |
| Charitable contribution deduction | 17.3 | 3.9 | 1062 | 996 |

*Note:*
Dollar amounts are in billions.

The next step is to target variable counts as close as is feasible. This is done by solving a constrained optimization problem that adjusts record weights, minimizing total absolute weight adjustment—a procedure that is commonly used to "age" tax microdata into the future. This process is essentially the same as stage 1 of the historical aging method described below in Section 2.2; refer to the discussion there for a full mathematical description of the algorithm. Appendix Table 1 lists the variables targeted and their allowable margin of error.

The final step in this process is to target dollar amount totals found on the aggregate returns. This is done by applying a simple rescaling factor calculated by AGI group. The factor is calculated by dividing the sum of a variable in the aggregate return by the sum in the new records, for each AGI group. In other words, dollar amounts are distributed in proportion to the amount found on each synthetic return. Once completed, the synthetic records are appended to the PUF.

## 2.2 Historical aging

The procedure described in Section 2.1 results in a usable PUF for the base year (2012 at the time of this writing). But microdata from more recent years are necessary for the Tax Module, both to validate tax calculator output historically and to get as-recent-as-possible data to serve as a jump-off point for PWBM's projections. To update the data for these in-between years, PWBM uses a modified version of the two-stage aging procedure initially developed by John O'Hare in the context of projecting tax data into the future (Rohaly, Carasso, and Saleem 2005).[4] First, we solve a linear programming problem that modifies record weights in order to hit record count targets for specified variables (e.g. the number of dependents claimed, or the number of returns reporting nonzero dividends). Second, we rescale dollar values to match actual targets by income group. Below we describe both stages in detail.

Note that our implementation differs from that of O'Hare (2009) in two ways. One, we are updating the PUF to reflect known historical aggregates rather than projecting the PUF into the future. The targets here are not forecasts, but instead actual values. Two, we have reversed the order of the stages: O'Hare (2009) first rescales, then optimizes record weights. We find that flipping the order allows us to more closely match record count targets.

### 2.2.1 Stage 1: Record weight optimization

The historical aging process begins with an earlier-year PUF and a set of later-year IRS summary tables. The summary tables include data for dollar amounts and record counts (i.e. the number of returns with nonzero value) for all major variables in the PUF, broken out by AGI group measured in nominal dollars. Our most recently implemented version uses the 2012 PUF and updates it to 2016 using info from the summary tables.[5]

Before anything else happens, record weights in the PUF are rescaled by the ratio of the

[4] Refer to O'Hare (2009) for a description of the procedure.

[5] The IRS recently released 2017 summary tables; we are in the process of updating our work with this new data.

number of later-year returns $W$ to the number of earlier-year returns, by AGI group $Y$. This ratio $R_W$ is:

$$R_W = \frac{W_Y}{\sum_{i=1}^{N_Y} w_i}$$

where $w_i$ is the weight of record $i$ in AGI group $Y$, and $N_Y$ is the total number of PUF records in AGI group $Y$. For example, there were 12.07 million returns filed with AGI between \$75,000 and \$100,000 according to the 2012 PUF. The 2016 aggregate tables show that number was 12.97 million by 2016. The rescaling factor then is calculated as $12.97/12.07 = 1.08$, which is applied to every PUF record that falls in that AGI group.

At this point, total return count targets by AGI group are matched exactly. But there are many other return count targets that we would like to match (for example, the number of returns reporting a given type of income or deduction, or the number of returns with head of household filing status). To the extent that the growth rate in these return counts differs from the growth rate in the total number of returns (i.e. the rescaling factor $R_W$), these targets will be unmet—and perhaps will be substantially off. Getting these correct means getting the margins of aggregate growth correct, which is critical for proper tax calculation. If, for example, we have too few returns reporting dividends, the aggregate growth in dividends will be distributed in stage 2 disproportionately from the intensive margin, rather than the extensive margin. In other words, average dividend income will be too high.

To solve this issue, we implement an algorithm that, given a set of return count totals to target (within some acceptable margin of error), adjusts record weights in a way that meets those targets. This optimization is constrained in two ways: first, aggregate absolute deviation from initial weights is minimized, and second, no individual record's weight deviates too far from its initial value. One could imagine a different approach to targeting record counts: we could instead randomly zero out values for variables with too-high return counts, or randomly reassign marital status, for example. But rescaling record weights is preferable as it maintains the joint distributions between variables, which can have important implications

for tax calculation.

Formally, the algorithm solves the optimization problem below.[6] Following the notation described by O'Hare (2009), let $z_i$ be the percent change adjustment made to a given record $i$'s $w_i$. The post-adjustment stage 1 weight $w_{i,s=1}$ is then:

$$w_{i,s=1} = w_i(1 + z_i)$$

The adjustment factor $z_i$ is decomposed into positive and negative components, $r_i$ and $s_i$ respectively, such that:

$$r_i = \begin{cases} z_i, & z_i > 0 \\ 0, & z_i \leq 0 \end{cases}$$

$$s_i = \begin{cases} z_i, & z_i < 0 \\ 0, & z_i \geq 0 \end{cases}$$

$$z_i = r_i - s_i$$

The purpose of this decomposition is to construct an absolute value deviation measure:

$$|z_i| = r_i + s_i$$

which forms the basis of the problem's objective function. The goal is to minimize the total absolute deviation from initial weights while meeting a set of aggregate targets, and with no individual adjustment being larger in absolute terms than some threshold $\delta$:

$$\min \sum_{i=1}^{N} |z_i| \quad s.t. \quad Az \geq b_{min}, \quad Az \leq b_{max}, \quad |z_i| < \delta$$

where $z = \left\{z_1, z_2, ..., z_N\right\}$ is the vector of decision variables, $A$ is the coefficient matrix of

[6] For practically implementing this algorithm, PWBM uses the lpSolveAPI R library, an interface to the mixed integer linear programming solver lp_solve.

constraints, and $b_{min}$ are $b_{min}$ are vectors of targets. Specifically, these vectors represent the difference between PUF totals $X_P$ and target values $X$ from the IRS summary tables—adjusted for some variable-specific acceptable error tolerance level $\alpha$:

$$b_{min} = \begin{pmatrix} X_1 * (1 - \alpha_1) \\ X_2 * (1 - \alpha_2) \\ ... \\ X_N * (1 - \alpha_N) \end{pmatrix} - \begin{pmatrix} X_{P,1} \\ X_{P,2} \\ ... \\ X_{P,N} \end{pmatrix}$$

$$b_{max} = \begin{pmatrix} X_1 * (1 + \alpha_1) \\ X_2 * (1 + \alpha_2) \\ ... \\ X_N * (1 + \alpha_N) \end{pmatrix} - \begin{pmatrix} X_{P,1} \\ X_{P,2} \\ ... \\ X_{P,N} \end{pmatrix}$$

For example, at this point the number of returns reporting dividends in the PUF is 30.98 million. The actual value for 2016 is 27.47 million, and we target this variable with $\alpha$ of 3 percent. Thus, the minimum acceptable outcome for the dividend return count is $27.47 * 0.97 = 26.65$ million, and the maximum is $27.47 * 1.03 = 28.29$ million. This variable's element in $b_{min}$ is therefore $26.65 - 30.98 = -4.34$ million and in $b_{max}$ it's $28.29 - 30.98 = -2.69$ million.

Where feasible, we target these counts by AGI group. We choose which variables to target based on a qualitative assessment of how important each is to correctly calculating tax liabilities, weighing the variable's marginal benefit of inclusion with the additional costs imposed on overall performance of the algorithm. Table 3 lists the variables targeted in stage 1.

We iterate over a series of $\delta$ values, starting at 0.5 and decreasing the value until the program can no longer find a solution.[7] Once an acceptable solution is found, the adjustment

[7]For reference, the model converges with $\delta = 0.4$ using PUF year 2012 and target year 2016; the lowest feasible $\delta$ generally increases when further removed from the PUF base year.

Table 3: Stage 1 target variables

| Variable | Number of AGI groups targeted | Tolerance |
|---|---|---|
| Total returns | All | 0.1% |
| Single returns | All | 0.1% |
| Married returns | All | 0.1% |
| Head of household returns | 0 | 0.1% |
| Dependent exemptions | 3 | 0.1% |
| Taxable interest income | 0 | 3% |
| Dividend income | 0 | 3% |
| Qualified dividend income | 0 | 3% |
| Positive sole proprietorship income | 0 | 3% |
| Net capital gains | 0 | 0.5% |
| Net capital losses | 0 | 10% |
| Gross social security income | 0 | 1% |
| Medical expense deduction | 1 | 1% |
| State and local tax deduction | 0 | 1% |
| Mortgage interest deduction | 5 | 1% |
| Charitable contribution deduction | 0 | 1% |
| Positive net S corp and partnership income | 0 | 3% |
| Positive passive S corp income | 0 | 10% |
| Positive passive partnership income | 0 | 10% |
| Positive active partnership income | 0 | 10% |
| American Opportunity Credit expenses | 0 | 5% |
| Lifetime Learning Credit expenses | 0 | 5% |

*Note:*
All income and deduction variables refer to the number of returns reporting those variables. The total number of AGI groups is 17. For variables targeted by AGI group, the tolerance value reported is for the overall total target; tolerance levels may be different for individual AGI group targets.

factors $z$ are applied to the set of weights in the PUF, and stage 1 is complete.

### 2.2.2 Stage 2: Rescaling

After stage 1, return counts have been matched as best as possible for critical variables. The next step is to account for per-return income growth by rescaling the value of each dollar amount variable such that aggregate PUF totals match the targets. The rescaling factor for variable $X$, denoted $R_X$, is calculated as:

$$R_X = \frac{X_Y}{\sum_{i=1}^{N_Y} x_i w_{i,s=1}}$$

where $X_Y$ is the target dollar amount from the IRS summary tables for AGI group $Y$, and $x_i$ is the value of variable $X$ for record $i$. Note that this is equivalent to "distributing"

the aggregate target income/deduction value in proportion to each record's base year share. To continue the example from above, filers with AGI between $20,000 and $30,000 report wage income of $843.28 billion in aggregate in the post-stage 1 PUF. In 2016, that figure was $852.52 billion. $R_X$ in this instance is thus $852.52/843.28 = 1.0196$.

By definition, dollar amount aggregates in the PUF now match the targets exactly. In practice, there are a few variables for which the data in the IRS summary tables is either incomplete (e.g. not available by income group) or unavailable entirely. In these cases, calculations are done with a single factor for all AGI groups or are rescaled with a similar variable's factor, respectively.

## 2.3 Statistical matching

At this point, the PUF has been updated to the most recent year possible and its aggregate returns have been distributed. The final data processing step for individual income tax data is to statistically match data from the Current Population Survey (CPS) to the historically-aged PUF. This practice is standard among tax microsimulation models; see Perese (2017), Rohaly, Carasso, and Saleem (2005), and O'Hare (2010) for examples. Doing so allows PWBM to impute important variables missing from the PUF (such as age, gender, and race) that the Tax Module requires when projecting the population of tax units into the future. The match also imputes records for tax units that do not file a Form 1040, which are required to properly calculate payroll tax liability and to evaluate certain policy proposals that would bring nonfilers into the tax-filing population.

In the most general sense, statistical matching involves taking variables from one dataset (the "donor" file) and transferring them to another dataset (the "host" file) in a way that attempts to preserve the statistical characteristics of each file. In our case, the PUF is the host file and the CPS is the donor file. Records from the donor file are matched to records from the host file according to some definition of *distance* that is minimized. The goal is that each record in the final matched dataset contains information from both datasets in an

internally consistent way. A statistical match is said to be "constrained" when all records from the donor file are required to appear on the host file; it is "unconstrained" otherwise. We implement a constrained match: all records from each dataset appear in the final matched file.

Our matching procedure begins by creating tax units our of households and determining tax filers in the CPS. We then subdivide each dataset into partitions based on demographic attributes available in both datasets. Next, within each demographic group, we estimate a regression for total income based on common variables. Fitted values are generated for income in each dataset. Finally, we sort records in each dataset by predicted income (again within demographic cell), then align the records using a weight-splitting procedure. Nonfiling tax units from the CPS are appended at the end. Each of these steps is described in detail below.

### 2.3.1 Step 1: Tax unit creation and filer determination

The CPS is a household-level survey that reports intra-household relationships between individuals. We begin the matching process by creating tax units—defined as the group of people who, if required, would file a tax return together—in the CPS. This procedure harmonizes the unit of observation between datasets. Our algorithm takes into account rules about household relationships, income requirements, dependency tests, and so on. It also assigns filing status (single, married filing jointly, head of household, or married filing separately) on the basis of which status would save the tax unit the most in taxes, subject to IRS filing restrictions. For 2016, we estimate a total of 176 million nondependent tax units.[8]

By definition, the PUF only includes information on tax units who file an individual income tax return. The next step, then, is to determine which tax units in the CPS will file a 1040—only filers in the CPS should be matched with records from the PUF. Tax units are generally required to file tax returns if they meet certain income criteria. Some tax

[8]Note that the number of tax units is fundamentally unascertainable as we do not observe nonfilers.

units falling below these thresholds, however, opt to file a tax return even if not required to, usually in order to claim a refundable credit. Therefore, our filer determination procedure involves two portions. The first is a deterministic rules-based algorithm mimicking the logic involved when actual taxpayers determine whether to file.[9] The second is a probabilistic module that models elective filing among CPS records not flagged as filers in the first step. This portion is aimed primarily at selecting low-income wage earners with children, a group that generally qualifies for refundable credits such as the Earned Income Tax Credit and the Additional Child Tax Credit. The elective filing probabilities are calibrated such that the number of CPS filers by demographic group more closely matches that of the PUF; see Section 2.3.2 below for more details.

Of our estimated 176 nondependent tax units in 2016, our filing algorithm selects 148 million to be filers, with 9.5 million more dependents opting to file a tax return.[10]

### 2.3.2 Step 2: Subdivision

The next step is to subdivide each dataset into demographic cells. These groups form the basis of the statistical match: only records in corresponding cells are merged across datasets. Partitioning the data in this way prevents records with dissimilar attributes from being matched.

We subdivide each dataset into 14 cells based on a combination of demographic characteristics: filing status, number of dependents, and age. Table 4 shows the number of tax units (sum of record weights) in each bucket. Note that in the final alignment step of the matching process, we rescale record weights in the CPS such that total returns by demographic group are equal across datasets. Section 2.3.4 details this adjustment.

Why not subdivide the datasets into even finer demographic cells? Doing so would allow for records to be matched along more dimensions—i.e. reduce the *distance* between any two

[9] See page 7 of the 2016 Form 1040 instructions for details on these rules.

[10] In step 4, these numbers change slightly as the alignment process adjusts CPS weights such that the actual number of filers is targeted precisely.

Table 4: Number of tax units and filers by demographic cell

| Demographic cell | CPS tax units | CPS filers | SOI filers | CPS/SOI ratio |
|---|---|---|---|---|
| Single, < 65, 0 dep. | 62.7 | 52.6 | 47.3 | 1.11 |
| Single, < 65, 1 dep. | 4.1 | 1.4 | 2.9 | 0.47 |
| Single, < 65, 2+ dep. | 5.7 | 1.6 | 1.1 | 1.48 |
| Single, 65+ | 18.7 | 14.8 | 11.4 | 1.29 |
| Married, < 65, 0 dep. | 27.2 | 26.1 | 14.5 | 1.79 |
| Married, < 65, 1 dep. | 6.7 | 6.7 | 9.3 | 0.72 |
| Married, < 65, 2 dep. | 7.9 | 7.9 | 10.8 | 0.74 |
| Married, < 65, 3+ dep. | 5.5 | 5.4 | 7.3 | 0.74 |
| Married, 65+ | 16.4 | 14.7 | 15.5 | 0.95 |
| Head of household, < 65, 1 dep. | 8.2 | 8.2 | 11.0 | 0.75 |
| Head of household, < 65, 2 dep. | 4.3 | 4.3 | 6.5 | 0.66 |
| Head of household, < 65, 3+ dep. | 2.6 | 2.5 | 3.2 | 0.78 |
| Head of household, 65+ | 1.7 | 1.2 | 0.9 | 1.32 |
| Dependents | 91.7 | 9.5 | 8.5 | 1.12 |

*Note:*
Counts are in millions.

records in the match. But there is a fundamental tradeoff: as the number of cells increases, so does the likelihood that larger weight adjustments will be required for records in the donor file. Thus the exercise becomes a judgmental optimization problem of choosing the number of cells that minimizes distance (broadly defined) between matched records without distorting the donor records too harshly. We feel that 14 cells is appropriate in our case.

#### 2.3.3 Step 3: Estimation

The demographic cells produced from step 2 above are used to condition on categorical traits when matching records. We need an additional criterion to measure distance between records after controlling for demographics. A natural approach is to use income, a variable common to both datasets. But there are conceptual differences between income as defined in either dataset, not to mention substantial measurement error in the CPS.[11]

To address this issue, we estimate a regression model of total income as a function of commonly available independent variables. The regression is estimated on the PUF, and is estimated separately for each demographic group resulting in 14 separate sets of parameter estimates. Then, we obtain fitted values for total income in both the PUF and the CPS and

[11]Business income in particular is heavily underreported.

use them to determine which records in one dataset are closest to records in the other.

More formally, for each cell $i$, we estimate an OLS regression of the form:

$$Y = \alpha_i + \beta_i X + \epsilon_i$$

where $Y$ is income, $\alpha$ is an intercept, and $X$ is a vector of independent variables: wage earnings, interest income, dividend income, sole proprietorship net income, combined partnership and S corporation net income, pension and annuity income, and unemployment compensation.

### 2.3.4 Step 4: Alignment

Next, we rescale CPS weights by a factor $R_i$ for demographic cell $i$:

$$R_i = \frac{N_{i,PUF}}{N_{i,CPS}}$$

which is applied to every CPS record's sample weight. Doing so ensures that when we align each cell, there is an equal number of filers to match between datasets.

We then sort records within each demographic cell by the predicted value of total income. The two datasets are then *aligned* and their records are matched—the first-ranked tax unit in the PUF is matched with the first-ranked tax unit for CPS, the second-rank tax units are matched with one another, and so on. If each record had a sample weight of one, this process would be straightforward; different-sized weights complicate the process. Imagine two records are chosen to be matched, but the PUF record's weight is larger than the CPS record's weight:

$$w_{PUF} = w_{CPS} + \alpha, \quad \alpha > 0$$

The PUF record is then split into two records:

$$w_{PUF} = w_{PUF\,1} + w_{PUF\,2}$$

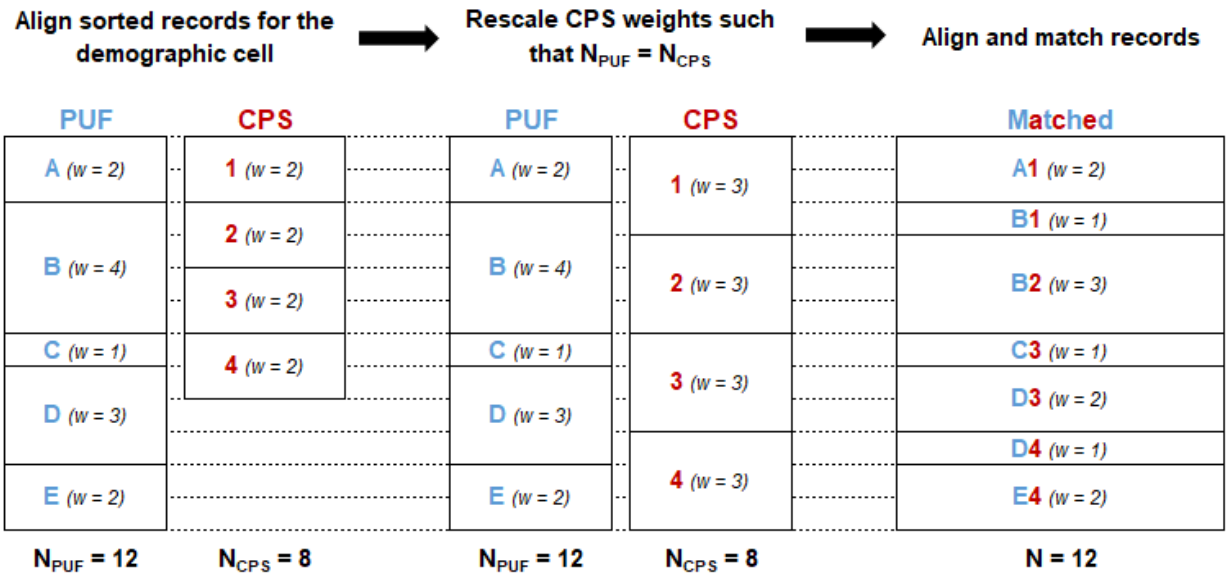
$$w_{PUF\,1} = w_{CPS}$$

$$w_{PUF\,2} = \alpha$$

and the equally-sized records are matched. The process continues with the next-ranked record, on and on until all records are exhausted.

A visual illustration of this weight-splitting process is instructive. Figure 1 recreates a helpful graphic presented in Perese (2017).

Figure 1. Illustration of alignment and matching process




Finally, we append the matched dataset with the nonfilers from the CPS. These records are necessary when calculating payroll taxes and for modeling proposals that would bring current-law nonfilers into the tax-filing population. This concludes the statistical matching process.

# 3 Corporate income taxes

A description of the corporate tax data will be included in a forthcoming version of this document.

# 4 Estate taxes

While there is no public-use microdata source for estate taxes, the IRS publishes summary tables containing aggregate dollar values and counts for important asset and deduction variables.[12] Following a procedure similar to that of the Tax Policy Center (Burman, Lim, and Rohaly 2008), we combine data from these tables with survey microdata on wealth and estimates of mortality to create synthetic estate tax microdata. The aggregates from the summary tables serve as targets for our microdata.

## 4.1 Base data

The starting point for our synthetic dataset is the SCF, a detailed survey of American households' asset and liability holdings. The survey is conducted by the Federal Reserve Board every three years. At the time of this writing the most recent survey was from 2016, which we eventually match with the 2017 filing-year estate tax summary tables. Critically, the SCF is designed to accurately measure the top of the wealth distribution by oversampling households with high net worth. This is especially important for modeling the estate tax, which is limited to the richest American families. By design, the SCF excludes the 400 richest Americans whose identities are published every year by *Forbes* magazine (the "Forbes 400"). We augment the SCF by appending the Forbes 400, taking their stated wealth as given and assigning a sample weight of one.[13] From this point forward, "SCF" refers to the raw SCF

[12] Given the very small number of estate tax returns filed, it would be impractical for the IRS to provide estate tax microdata in a way that meets privacy standards.

[13] In practice, a small overlap between the Forbes 400 and SCF has been found, implying the need for an adjustment to SCF weights when appending the Forbes 400 (see Batty et al. (2019)). We have not incorporated such an adjustment at this point.

appended with the Forbes 400.

While the SCF has detailed information on asset composition, we currently only take two financial variables from the survey: gross asset value ("gross estate" in the context of estate taxes) and the value of all outstanding debts. As noted by Burman, Lim, and Rohaly (2008), there are substantial discrepancies between the mortality-adjusted SCF and the IRS summary tables for certain asset categories. Targeting each component of gross estate can create problems for the adjustment procedure described below in Section 4.3.[14] Additionally, asset composition is not particularly important for modeling estate tax reforms, as proposals typically involve changing the exemption level and rate structure rather than narrowing the tax base (for example by exempting closely-held businesses).

Estate taxes are levied at the individual level, but the SCF's unit of observation is the "primary economic unit", which is roughly equivalent to a family. We therefore split records of married couples into two records, assigning 50 percent of total net worth to each person.

## 4.2 Mortality imputations

The next step is to multiply each SCF record's weight by the probability of death. We begin by assigning mortality rates conditional on age and gender, estimated from PWBMsim. Stopping here, however, would produce an expected estate tax population that overweights richer married couples because of mortality gradients by income and marital status. We therefore make two adjustments to mortality rates.

The first adjustment is based on the interaction between income and age. We draw on estimated mortality rates from Chetty et al. (2016), which can be conditioned on age and income percentile, among other variables.[15] We apply a rescaling factor $R_{y,a}$ representing the relative likelihood of death by income percentile, conditional on age:

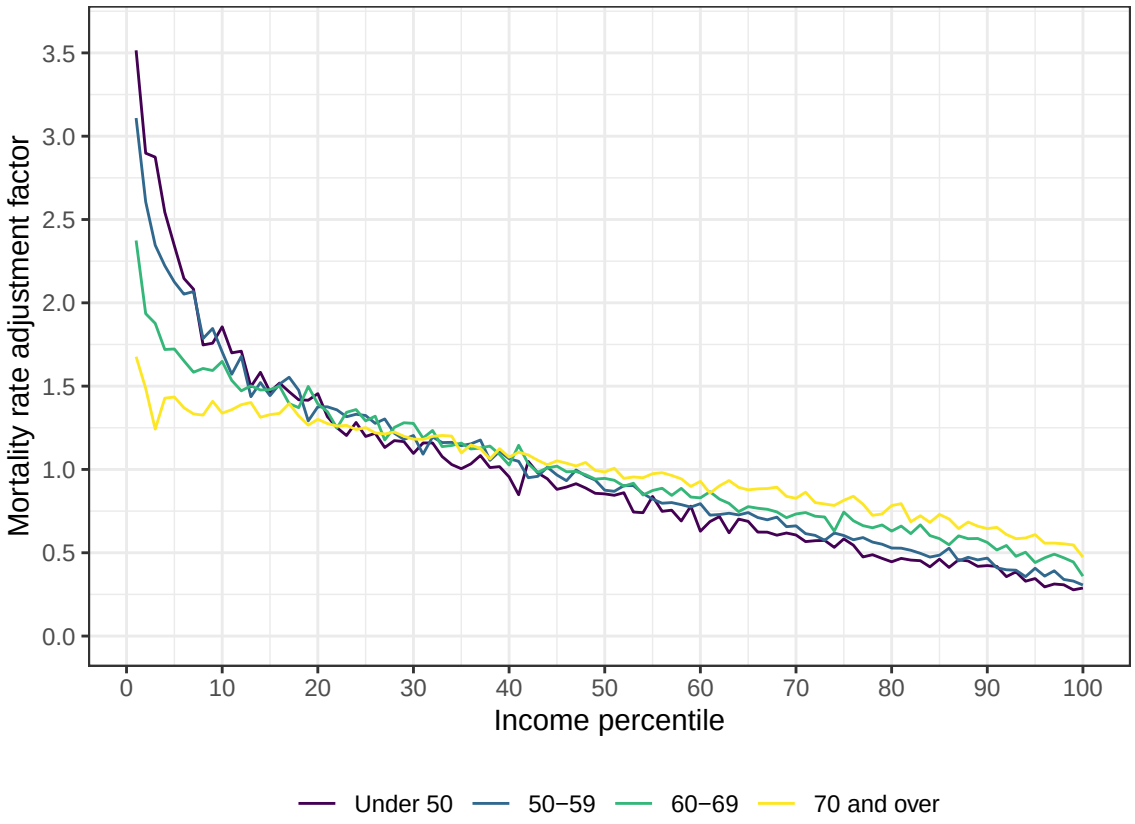
[14] Making large rescaling adjustments to components of gross estate can move records between gross estate groups, which serve as the groups by which the rescaling factors themselves are defined.

[15] See Table 15 from the online data appendix.

$$R_{y,a} = \frac{m_{y,a}}{m_a}$$

where $m_{y,a}$ is the mortality rate for income percentile group $y$ and age group $a$.[16] Values of $R_{y,a}$ fall as income increases. Figure 2 shows that this gradient is steeper for the youngest age group which reflects diminishing marginal returns to costly health interventions as age increases.

Figure 2. Mortality rate adjustments by income and age




Next, we apply a second set of rescaling factors to account for the mortality advantage for married people. Much like the income mortality gradient, these factors vary as a function of age. We draw on summary statistics from Table 3 of Johnson et al. (2000) to calculate rescaling factors along two dimensions, gender and age.[17] These factors are calculated such

[16] $y = \{1, 2, 3...100\}$, $a = \{< 50, 50-59, 60-69, >= 70\}$

[17] We take simple averages of the three unmarried groups(widowed, divorced/separated, never married), and average across race according to racial shares.

that overall mortality is unchanged. Table 5 presents the factors.

Table 5: Mortality adjustments by marital status

| Group | Married | Unmarried |
|---|---|---|
| Males, under 65 | 0.94 | 1.26 |
| Males, 65+ | 0.97 | 1.10 |
| Females, under 65 | 0.88 | 1.27 |
| Females, 65+ | 0.95 | 1.06 |

After this step, we multiply each SCF record's weight by its mortality rate. The dataset now reflects the population of expected estates in 2016. Going forward, we refer to this dataset as the estate tax microdata.

## 4.3 Two-stage adjustment procedure

The estate tax microdata at this point do not match the aggregates from the IRS summary tables. We predict nearly four times as many estate tax returns with gross estate above $5 million. This discrepancy, also found by other researchers (Burman, Lim, and Rohaly 2008), could reflect a number of factors: soon-to-be-decedents might engage in tax planning to minimize/eliminate estate tax liability (through either legal or illegal means); our estimates of mortality may contain error; and/or the SCF itself contains error, either through sampling issues or through respondents misreporting asset values. Regardless of the sources of differences, we need to adjust the estate tax microdata—both record weights and variable values—to match the aggregates as close as is feasible. To do so, we employ a version of two-stage procedure described in Section 2.2. Rather than using it to age microdata forward to a future year, here the procedure is used to correct for the sources of error detailed above.

Before running the two-stage procedure, we rescale all record weights to match the total number of returns within groups of gross estate and marital status.[18] This set of rescaling factors is akin to $R_W$ defined in Section 2.2.1, but with gross estate and marital status

[18]The gross estate groups are: < $5M; $5M-$10M; $10M-$20M; $20M-$50M; > $50M.

replacing the AGI grouping. Two complications arise here. First, marital status is not actually available in the summary tables; gross estate group is the sole breakdown. We impute the married/non-married split based on data from Appendix Table 3 of Burman, Lim, and Rohaly (2008), who were given supplemental breakdowns by the IRS. Second, because decedents are generally not required to file an estate tax return if their gross estate falls below the exemption amount ($5.45 million in 2016), we have no means of observing actual aggregates for the vast majority of decedents.[19] Therefore we only calculate rescaling factors for returns with gross estate above $5 million. For other returns, we run the same mortality imputations on 2001 SCF data and compare the resulting population to 2002 filing year summary tables—a year when the exemption was $1M. The implied rescaling factor for 2001 decedents with gross estate between $1 million and $5 million is roughly 0.5, so we halve all weights with gross estate of less than $5 million in the 2016 data.

Rather than begin with the count-targeting linear programming algorithm as we do for the individual PUF, we reverse the order of the two stages here: first we rescale values, then we target counts. The reason for this change is the larger distance between predicted and target values.[20] Rescaling gross estate values can move records into a different gross estate group—the grouping on which the count targets are defined. Therefore, we first rescale gross estate (and debts and mortgages), redefining gross estate group based on new values. Then we run the linear programming algorithm, targeting the variables shown in Table 6. This ordering ensures that we hit count targets within actual gross estate groupings.

The algorithm solves at a $\delta$ of 0.66. While number of returns, gross estate, and marital status are easily targeted within small margins of error, debts and mortgages is substantially off. We make one final rescaling adjustment to match dollar amount values by gross estate group exactly. In practice, this means that any discrepancies driven by the extensive margin are corrected with an intensive margin adjustment.

[19] While this is not actually a problem for calculating liabilities under current law, we need to have good estimates of the nonfiling population for modeling proposals that lower the exemption.

[20] As noted before, this larger discrepancy makes sense in the context of this exercise: we're correcting for error here, not aging microdata year to year.

Table 6: Targeted variables, before and after stage 2 adjustments

| Variable | Gross estate | Tolerance | Actual | Microdata, before | Microdata, after |
|---|---|---|---|---|---|
| Number of returns | $5M-$10M | 1% | 7374.0 | 7272.1 | 7300.3 |
| | $10M-$20M | 1% | 2705.0 | 2752.9 | 2694.8 |
| | $20M-$50M | 1% | 1063.0 | 1020.3 | 1064.5 |
| | $50M+ | 1% | 423.0 | 382.9 | 418.8 |
| Number of married returns | $5M-$10M | 5% | 3318.3 | 3203.9 | 3232.1 |
| | $10M-$20M | 5% | 1298.4 | 1358.8 | 1363.3 |
| | $20M-$50M | 5% | 542.1 | 483.7 | 556.0 |
| | $50M+ | 5% | 236.9 | 212.4 | 246.6 |
| Gross estate | $5M-$10M | 1% | 51.8 | 51.5 | 51.7 |
| | $10M-$20M | 1% | 36.1 | 37.3 | 36.4 |
| | $20M-$50M | 1% | 31.7 | 32.0 | 32.0 |
| | $50M+ | 1% | 68.7 | 66.7 | 69.3 |
| Debts and mortgages | $5M-$10M | 40% | 1.7 | 1.7 | 1.7 |
| | $10M-$20M | 40% | 1.5 | 1.5 | 1.8 |
| | $20M-$50M | 40% | 1.4 | 2.5 | 2.0 |
| | $50M+ | 40% | 4.3 | 3.2 | 3.3 |
| Debts and mortgages, number of returns | $5M-$10M | 40% | 4882.0 | 3031.6 | 3031.6 |
| | $10M-$20M | 40% | 2014.0 | 834.8 | 1208.4 |
| | $20M-$50M | 40% | 852.0 | 356.0 | 511.2 |
| | $50M+ | 40% | 345.0 | 199.5 | 227.5 |

*Note:*
Dollar amounts are in billions.

## 4.4 Deduction imputations

The final step in preparing the synthetic estate tax microdata is to impute other variables that are not included in the SCF (because they are mostly contingent on the death of decedent), but are necessary for estate tax liability calculation. These variables are funeral expenses, executor's commissions, attorney's fees, other expenses and losses, charitable deductions, state death tax liability, adjusted lifetime taxable gifts. The IRS summary tables report dollar amounts and return counts for each of these variables broken down by gross estate group.

Following Burman, Lim, and Rohaly (2008), we estimate imputation functions for each variable by gross estate group. The probability $p$ of an estate tax return recording a nonzero value for a variable $x$ is estimated as the number of returns in a gross estate group reporting the variable divided by the total number of returns $N$ in that group:

$$p = \frac{N_{x>0}}{N}$$

For each variable, we probabilistically assign reporting status to record $i$ using its estimated $p$ by drawing a value $z_i$ from the uniform distribution. If $z_i < p$, we assign a value $x_i$ to record $i$ as a function of gross estate $g_i$. We use a simple linear function $s$ defined as the average value of $x$ divided by average gross estate (by gross estate group):

$$s = \frac{X/N_{x>0}}{G/N}$$

where $X$ is the sum of variable $x$ and $G$ is total gross estate. $s$ is then applied to record $i$'s gross estate $g_i$ to get the value $x_i$:

$$x_i = g_i s$$

In practice we have to "unpack" sample weights by duplicating records and correspondingly decreasing sample weights to achieve a satisfactory imputation. The SCF simply has too large of sample weights to match targets for narrow gross estate tax groups when probabilistically imputing variables.

There are three more variables with slightly different imputation procedures: lifetime gift tax paid, the deceased spousal unused exemption, and bequests to a surviving spouse. The former two are exactly the same as described above, but are conditioned on other variables (adjusted taxable lifetime gifts and being unmarried, respectively). This prevents internally inconsistent records, for example a married decedent taking the deceased spousal unused exemption. Bequests to surviving spouses are assigned to married records only, with 90 percent of records taking a deduction equal to the full value of the estate (Burman, Lim, and Rohaly 2008). The other 10 percent are imputed as per the procedure described above.

# 5 Projecting tax microdata forward

Each of the processes described above produce tax return microdata for the most recently available year of data. But to calculate tax liabilities into the future, the Tax Module requires information on the characteristics of tax units well into the future. This section briefly describes how each dataset is projected forward during model runs in accordance with PWBM's demographic and economics forecasts.

PWBM's demographic microsimulation model, PWBMsim, forms the basis for how we project the characteristics of tax units into the future. PWBMsim simulates a representative population of individuals and households, with attributes such as gender, marital status, race, residency, immigration status, educational attainment, among others. It models key economic characteristics and decisions such as labor productivity, labor force participation, consumption-savings choices, and more. The population evolves based on historically-calibrated transition probabilities, which are then forecasted into the future.

For individual income taxes, the first step for projecting future populations of tax units is to simulate demographic change. Trends in fertility, mortality, marriage, household formation, and other demographic indicators have significant implications for tax revenues; it's important that the projected tax return microdata reflect these dynamics. For each year $t$ in a PWBMsim model run, the number of tax units $N$ is aggregated by demographic group $g$. This time series of demographic group population totals is used to construct a series of demographic aging factors, which is simply the year-over-year growth rate:

$$\frac{N_{g,t+1}}{N_{g,t}} - 1$$

This factor is applied to record weights during model runs. Age and marital status are currently the only characteristics used to define $g$ for tax units. We are exploring expanding these groupings to directly include more demographic attributes. For estate taxes, we follow a very similar approach: the demographic aging factor is simply PWBMsim's projection of

the growth rate in the number of deaths.

The second step for modeling the evolution of tax returns involves simulating income growth. As with demographic change, the Tax Module models growth in tax variables in accordance with PWBMsim forecasts. Projections of macroeconomic variables are mapped to conceptually similar tax variables in the Tax Module. For example, W2 wages grow with overall wages, and dividend income is paired with corporate profits. Each tax variable is adjusted by a rescaling factor $R$ each year. The factor starts with one plus the per-capita growth in its respective PWBMsim economic aggregate $X_P$:

$$R = \frac{(X_{P,t+1}/N_{t+1})}{(X_{P,t}/N_t)}$$

Then, following Perese (2016), we make an adjustment to reflect differential growth rates by income group. To get next year's $(t+1)$ value of $x$ for record $i$, the value at time $t$ is adjusted by the above rescaling factor $R$ times a second distributional factor $D_Y$, which varies by income group $Y$:

$$x_{i,t+1} = x_t * R * D_Y$$

where $Y$ is one of three AGI groups: sub-90th percentile, 90th-99th percentile, and the top 1 percent. Unlike demographics, which are fixed across policy scenarios, projections of reported income in the Tax Module are endogenous to tax policy. Changes in absolute tax rates or relative tax rates and temporary provisions produce micro-dynamic responses wherein taxpayers shift income to pay a lower effective tax rate.[21]

Note that this projection process is applied *during* each model run of the Tax Module. This process contrasts with the approach that some other tax microsimulation models—those from the Tax Policy Center and Open Source Policy Center among others—take. These models project data in a separate module using the two-stage linear programming algorithm

[21] Conceptually, these responses represent accounting and timing tricks rather than changes in real economic activity such as labor supply.

described in Section 2.2, with exogenously defined population and income projections used as targets. The output is a series of tax return datasets for future years that are sequentially read into a tax calculator. Our in-model aging procedure is similar to CBO's approach; see Perese (2016) for a discussion on the relative merits of each.

# 6 References


Batty, Michael, Jesse Bricker, Joseph Briggs, Elizabeth Holmquist, Susan Mcintosh, Kevin Moore, Eric Nielsen, et al. 2019. "Introducing the Distributional Financial Accounts of the United States." *Finance and Economics Discussion Series* 2019 (017). https://doi.org/10.17016/feds.2019.017.

Burman, Leonard, Katherine Lim, and Jeffrey Rohaly. 2008. "Back from the Grave: Revenue and Distributional Effects of Reforming the Federal Estate Tax." Tax Policy Center. https://www.taxpolicycenter.org/sites/default/files/alfresco/publication-pdfs/411777-Back-from-the-Grave.PDF.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. "The Association Between Income and Life Expectancy in the United States, 2001-2014." *Jama* 315 (16): 1750. https://doi.org/10.1001/jama.2016.4226.

Johnson, Norman J, Eric Backlund, Paul D Sorlie, and Catherine A Loveless. 2000. "Marital Status and Mortality." *Annals of Epidemiology* 10 (4): 224–38. https://doi.org/10.1016/s1047-2797(99)00052-6.

O'Hare, John. 2009. "Extrapolation Methodology." https://github.com/PSLmodels/taxdata/blob/1f5f317e37b41a233efd75fb94f436f923b3e2d1/doc/Stage1_Stage2_Methodology.pdf.

———. 2010. "Statiscal Matching Documentation." https://github.com/PSLmodels/taxdata/blob/1f5f317e37b41a233efd75fb94f436f923b3e2d1/puf_data/StatMatch/doc/MatchingDocumentationRevised.pdf.

Perese, Kevin. 2016. "Projection and Alignment Methods for Static Microsimulation Models." Congressional Budget Office; Association for Public Policy Analysis and Management 2016 Pre-Conference Workshop. https://www.cbo.gov/sites/default/files/114th-congress-2015-2016/presentation/52147-presentation.pdf.

———. 2017. "Statistically Matching Administrative Tax Data with Household Survey

Data." Congressional Budget Office; Workshop at the Washington Center for Equitable Growth. https://www.cbo.gov/system/files/115th-congress-2017-2018/presentation/52914-presentation.pdf.

Rohaly, Jeffrey, Adam Carasso, and Mohammed Adeel Saleem. 2005. "The Urban-Brookings Tax Policy Center Microsimulation Model: Documentation and Methodology for Version 0304." Tax Policy Center. https://www.urban.org/sites/default/files/alfresco/publication-pdfs/411136-The-Urban-Brookings-Tax-Policy-Center-Microsimulation-Model.PDF.