Enhancing Survey Microdata with Administrative Records:
A Novel Approach to Microsimulation Dataset Construction

Nikhil Woodruff[*]   Max Ghenis[*]

November 12, 2024

Abstract

We combine the demographic detail of the Current Population Survey (CPS) with the tax precision of the IRS Public Use File (PUF) to create an enhanced microsimulation dataset. Our method uses quantile regression forests to transfer income and tax variables from the PUF to demographically-similar CPS households. We create a synthetic CPS-structured dataset using PUF tax information, stack it alongside the original CPS records, then use dropout-regularized gradient descent to reweight households toward administrative targets from IRS Statistics of Income, Census population estimates, and program participation data. This preserves the CPS's granular demographic and geographic information while leveraging the PUF's tax reporting accuracy. The enhanced dataset provides a foundation for analyzing federal tax policy, state tax systems, and benefit programs. We release both the enhanced dataset and our open-source enhancement procedure to support transparent policy analysis.

1 Introduction

Microsimulation models are essential tools for analyzing the distributional impacts of tax and transfer policies. These models require microdata that accurately represent both the demographic composition of a population and their economic circumstances, particularly their tax situations. However, available data sources typically excel in one dimension while falling short in another.

The Current Population Survey (CPS), conducted by the U.S. Census Bureau, provides rich demographic detail and household relationships but suffers from underreporting of income and lacks tax information. Conversely, the Internal Revenue Service's Public Use File (PUF) offers precise tax data but contains limited demographic information and obscures household

---

[*]PolicyEngine

structure. This tradeoff between demographic detail and tax precision poses a significant challenge for policy analysis.

This paper presents a novel approach to combining these complementary data sources. We develop a methodology that preserves the demographic richness of the CPS while incorporating the tax precision of the PUF, creating an enhanced dataset that serves as the foundation for PolicyEngine's microsimulation capabilities. Our approach differs from previous efforts in three key ways:

First, we employ quantile regression forests to transfer distributions rather than point estimates between datasets, preserving the complex relationships between variables. Second, we maintain household structure throughout the enhancement process, ensuring that family relationships crucial for benefit calculations remain intact. Third, we implement a sophisticated reweighting procedure that simultaneously matches dozens of demographic and economic targets while avoiding overfitting through a dropout-enhanced gradient descent approach.

The resulting dataset demonstrates superior performance in both tax and transfer policy simulation. When compared to administrative totals, our enhanced dataset reduces discrepancies in key tax components by an average of 40% relative to the baseline CPS, while maintaining or improving the accuracy of demographic and program participation variables.

The remainder of this paper is organized as follows: Section 2 reviews related work in survey enhancement and microsimulation data construction. Section 3 describes our data sources and their characteristics. Section 4 presents our methodology in detail. Section 5 validates our results against external benchmarks. Section 6 discusses implications and limitations, and Section 7 concludes.

Our contributions include:

- A novel methodology for combining survey and administrative data while preserving distributional relationships

- An open-source implementation that can be adapted for other jurisdictions and policy models

- A validation framework comparing enhanced estimates against multiple external benchmarks

- A new, publicly available microdata file suitable for US tax and benefit policy analysis

2 Background

Tax microsimulation models are essential tools for analyzing the distributional and revenue impacts of tax policy changes. By simulating individual tax units rather than relying on aggregate statistics, these models can capture the complex interactions between different provisions of the tax code and heterogeneous effects across the population. The core challenges these models face include:

- Combining multiple data sources while preserving statistical validity

- Aging historical data to represent current and future years

- Imputing variables not observed in the source data

- Modeling behavioral responses to policy changes

- Calibrating results to match administrative totals

Each existing model approaches these challenges differently, making tradeoffs between precision, comprehensiveness, and transparency. We build on their methods while introducing new techniques for data synthesis and uncertainty quantification.

2.1   Government Agency Models

The U.S. federal government maintains several microsimulation capabilities through its policy analysis agencies, which form the foundation for official policy analysis and revenue estimation.

The Congressional Budget Office's model emphasizes behavioral responses and their macroeconomic effects (Congressional Budget Office, 2018). Their approach uses a two-stage estimation process:

1. Static scoring: calculating mechanical revenue effects assuming no behavioral change

2. Dynamic scoring: incorporating behavioral responses calibrated to empirical literature

CBO's elasticity assumptions have evolved over time in response to new research, particularly regarding the elasticity of taxable income (ETI). Their current approach varies ETI by income level and type of tax change, broadly consistent with the academic consensus surveyed in (Saez et al., 2012). The model also incorporates detailed projections of demographic change and economic growth from CBO's other forecasting models.

The Joint Committee on Taxation employs a similar approach but with particular focus on conventional revenue estimates (Joint Committee on Taxation, 2023). Their model maintains detailed imputations for:

- Business income allocation between tax forms

- Retirement account contributions and distributions

- Asset basis and unrealized capital gains

- International income and foreign tax credits

A distinguishing feature is their treatment of tax expenditure interactions - addressing both mechanical overlap (e.g., between itemized deductions) and behavioral responses (e.g., between savings incentives).

The Treasury's Office of Tax Analysis model features additional detail on corporate tax incidence and international provisions (Office of Tax Analysis, 2012). Their approach emphasizes the relationship between different types of tax instruments through a series of linked models:

- Individual income tax model using matched administrative data

- Corporate microsimulation using tax returns and financial statements

- International tax model incorporating country-by-country reporting

- Estate tax model with SCF-based wealth imputations

This integration allows OTA to analyze proposals affecting multiple parts of the tax system consistently.

2.2 Research Institution Models

2.2.1 Urban Institute Family of Models

The Urban Institute maintains several complementary microsimulation models, each emphasizing different aspects of tax and transfer policy analysis.

The Urban-Brookings Tax Policy Center model (Tax Policy Center, 2022) combines the IRS Public Use File with Current Population Survey data through predictive mean matching, an approach similar to what we employ in Section 4. Their imputation strategy aims to preserve joint distributions across variables using regression-based techniques for:

- Wealth holdings (18 asset and debt categories)

- Education expenses (by level and institution type)

- Consumption patterns (16 expenditure categories)

- Health insurance status (plan type and premiums)

- Retirement accounts (DB/DC split and contribution levels)

TRIM3 emphasizes the time dimension of policy analysis, with sophisticated procedures for converting annual survey data into monthly variables (Urban Institute, 2024b). Key innovations include:

- Allocation of employment spells to specific weeks using BLS benchmarks

- Probabilistic monthly assignment of benefit receipt

- State-specific program rules and eligibility determination

- Integration of administrative data for validation

This monthly allocation approach informs our treatment of time variation in Section 3.

The newer ATTIS model (Urban Institute, 2024a) focuses on interactions between tax and transfer programs. Building on the American Community Survey rather than the CPS provides better geographic detail at the cost of requiring additional tax variable imputations. Their approach to correcting for benefit underreporting in survey data parallels our methods in Section 4.

2.2.2 Other Research Institution Models

The Institute on Taxation and Economic Policy model (Institute on Taxation and Economic Policy, 2024) is unique in its comprehensive treatment of federal, state and local taxes. Key features include:

- Integration of income, sales, and property tax microsimulation

- Detailed state-specific tax calculators

- Consumer expenditure imputations for indirect tax analysis

- Race/ethnicity analysis through statistical matching

The Tax Foundation's Taxes and Growth model (Tax Foundation, 2024) emphasizes macroeconomic feedback effects through a neoclassical growth framework. Their approach includes:

- Production function based on CES technology

- Endogenous labor supply responses

- Investment responses to cost of capital

- International capital flow effects

2.3 Open Source Initiatives

Recent years have seen growing interest in open source approaches that promote transparency and reproducibility in tax policy modeling.

The Budget Lab at Yale (Budget Lab, 2024) maintains a fully open source federal tax model distinguished by:

- Modular codebase with clear separation of concerns

- Flexible behavioral response specification

- Comprehensive test suite and documentation

- Version control and continuous integration

Their approach to code organization and testing informs our own development practices.

The Policy Simulation Library's Tax-Data project (Policy Simulation Library, 2024) provides building blocks for tax microsimulation including:

- Data processing and cleaning routines

- Statistical matching algorithms

- Variable imputation methods

- Growth factor calculation

- Validation frameworks

We build directly on several Tax-Data components while introducing new methods for synthesis and uncertainty quantification described in Section 4.

2.4 Key Methodological Challenges

This review of existing models highlights several common methodological challenges that our approach aims to address:

1. **Data Limitations**: Each primary data source (tax returns, surveys) has significant limitations. Tax returns lack demographic detail; surveys underreport income and benefits. While existing models use various matching techniques to combine sources, maintaining consistent joint distributions remains difficult.

2. **Aging and Extrapolation**: Forward projection requires both technical adjustments (e.g., inflation indexing) and assumptions about behavioral and demographic change. Current approaches range from simple factor adjustment to complex forecasting models.

3. **Behavioral Response**: Models must balance tractability with realism in specifying how taxpayers respond to policy changes. Key challenges include heterogeneous elasticities, extensive margin responses, and general equilibrium effects.

4. **Uncertainty Quantification**: Most models provide point estimates without formal measures of uncertainty from parameter estimates, data quality, or specification choices.

Our methodology, detailed in Section 4, introduces novel approaches to these challenges while building on existing techniques that have proven successful. We particularly focus on quantifying and communicating uncertainty throughout the modeling process.

3 Data

3.1 Current Population Survey

The Census Bureau administers the Current Population Survey Annual Social and Economic Supplement (CPS ASEC, or hereafter the CPS) each March. In March 2024, they surveyed 89,473 households representing the U.S. civilian non-institutional population about their activities in the 2023 calendar year.

The CPS's key strengths include:

- Rich demographic detail including age, sex, race, ethnicity, and education

- Complete household relationship matrices

- Program participation indicators

- State identifiers, and partial county identifiers

However, the CPS has known limitations for tax modeling:

- Underreporting of income, particularly at the top of the distribution due to top-coding

- Limited tax-relevant information (e.g., itemized deductions)

- No direct observation of tax units within households

- Imprecise measurement of certain income types (e.g., capital gains)

3.2 IRS Public Use File

The Internal Revenue Service Public Use File (PUF) is a national sample of individual income tax returns, representing the 151.2 million Form 1040, Form 1040A, and Form 1040EZ Federal Individual Income Tax Returns filed for Tax Year 2015. The file contains 119,675 records sampled at varying rates across strata, with 0.07 percent sampling for strata 7 through 13 (Bryant, 2023). The data are extensively transformed to protect taxpayer privacy while preserving statistical properties.

The Public Use Tax Demographic File supplements the PUF with:

- Age ranges for primary taxpayers (different ranges for dependent vs non-dependent filers)

- Dependent age information in six categories (under 5, 5-13, 13-17, 17-19, 19-24, 24+)

- Gender of primary taxpayer

- Earnings splits for joint filers (categorizing primary earner share)

Key disclosure protections include:

- Demographic information limited to returns in strata 7-13

- Suppression of dependent ages for returns with farm income or homebuyer credits

- Minimum population thresholds for dependent age reporting

- Sequential limits on dependent counts by filing status

The PUF's key strengths include:

- Precise income amounts derived from information returns

- Complete tax return information including itemized deductions

- Actual tax unit structure

- Accurate income type classification

The PUF's limitations include:

- Limited demographic information

- No household structure beyond the tax unit

- No geographic information such as state

- No program participation information

- Privacy protections that mask extreme values

- Lag; the latest version as of November 2024 is for the 2015 tax year

3.3 External Validation Sources

We validate our enhanced dataset against 570 targets from several external sources:

3.3.1 IRS Statistics of Income

The Statistics of Income (SOI) Division publishes detailed tabulations of tax return data, including:

- Income amounts by source and adjusted gross income bracket

- Number of returns by filing status

- Itemized deduction amounts and counts

- Tax credits and their distribution

These tabulations serve as key targets in our reweighting procedure and validation metrics.

3.3.2 CPS ASEC Public Tables

Census Bureau publications provide demographic and program participation benchmarks, including:

- Age distribution by state

- Household size distribution

- Program participation rates

3.3.3 Administrative Program Totals

We incorporate official totals from various agencies, including but not limited to:

- Social Security Administration beneficiary counts and benefit amounts

- SNAP participation and benefits from USDA

- Earned Income Tax Credit statistics from IRS

- Unemployment Insurance claims and benefits from Department of Labor

3.4 Variable Harmonization

A crucial preparatory step is harmonizing variables across datasets. We develop a detailed crosswalk between CPS and PUF variables, accounting for definitional differences. Key considerations include:

- Income classification (e.g., business vs. wage income)

- Geographic definitions

- Family relationship categories

For some variables, direct correspondence is impossible, requiring imputation strategies described in the methodology section. The complete variable crosswalk is available in our open-source repository.

4 Methodology

4.1 Overview

Our approach enhances the Current Population Survey (CPS) with information from the IRS Public Use File (PUF) through a multi-stage process:

1. Train quantile regression forests on PUF tax records to learn distributions of tax-related variables

2. Generate a synthetic dataset that combines PUF tax precision with CPS-like demographic detail

3. Stack these synthetic records alongside the original CPS records

4. Optimize household weights to match administrative benchmarks while determining the optimal mix of original and synthetic records

This method preserves the CPS's demographic richness and household relationships while incorporating the PUF's precise tax information. Each component is detailed in the following sections.

4.2 Demographic Variable Construction

Following the IRS specifications for the Public Use File, we construct three key demographic variables: dependent ages, primary taxpayer age ranges, and earnings splits between spouses.

4.3 Dependent Ages

For each dependent, we construct age categories following IRS constraints:

- Under 5

- 5 under 13

- 13 under 17

- 17 under 19

- 19 under 24

- 24 or older

The number of dependents is limited by filing status:

- Up to 3 dependents for joint returns and head of household returns

- Up to 2 dependents for single returns

- Up to 1 dependent for married filing separately returns

Dependents are ordered sequentially by type:

1. Children living at home

2. Children living away from home

3. Other dependents

4. Parents

4.3.1 Primary Taxpayer Age

Age ranges are constructed differently for dependent and non-dependent returns:
For non-dependent returns:

- Under 26

- 26 under 35

- 35 under 45

- 45 under 55

- 55 under 65

- 65 or older

For dependent returns:

- Under 18

- 18 under 26

- 26 or older

4.3.2 Earnings Splits

For joint returns, we calculate the primary earner's share of total earnings:

$$\text{Primary Share} = \frac{\text{Primary Wages} + \text{Primary SE Income}}{\text{Total Wages} + \text{Total SE Income}}$$

Where:

- Primary wages and SE income = E30400 - E30500

- Secondary wages and SE income = E30500

This share is categorized into:

- 75 percent or more earned by primary

- Less than 75 percent but more than 25 percent earned by primary

- Less than 25 percent earned by primary

4.3.3 Implementation Details

When decoding age ranges into specific ages, we use random assignment within the range to avoid unrealistic bunching. For example, when the PUF indicates age 80, we randomly assign an age between 80 and 84.

The ordering of dependents is preserved when constructing synthetic tax units to maintain consistency with the original data structure.

4.4 PUF Data Preprocessing

The preprocessing of the IRS Public Use File involves variable renaming, recoding, and construction of derived variables to align with PolicyEngine's analytical framework.

4.4.1 Medical Expense Categories

Total medical expenses are decomposed into specific categories using fixed ratios derived from external data:

- Health insurance premiums without Medicare Part B: 45.3%

- Other medical expenses: 32.5%

- Medicare Part B premiums: 13.7%

- Over-the-counter health expenses: 8.5%

4.4.2 Variable Construction

Key derived variables include:

Qualified Business Income (QBI)   Calculated as the maximum of zero and the sum of:

- Schedule E income (E00900)

- Partnership and S-corporation income (E26270)

- Farm income (E02100)

- Rental income (E27200)

W2 wages from qualified business are then computed as 16% of QBI.

Filing Status   Mapped from MARS codes:

- 1 → SINGLE

- 2 → JOINT

- 3 → SEPARATE

- 4 → HEAD_OF_HOUSEHOLD

Records with MARS = 0 (aggregate records) are excluded.

4.4.3   Income Component Separation

Several income sources are separated into positive and negative components:

- Business income split into net profits (positive) and losses (negative)

- Capital gains split into gross gains and losses

- Partnership and S-corporation income split into income and losses

- Rental income split into net income and losses

4.4.4   Variable Renaming

The following PUF variables are renamed to align with PolicyEngine conventions:

Direct Renames

- E03500 → alimony_expense

- E00800 → alimony_income

- E20500 → casualty_loss

- E32800 → cdcc_relevant_expenses

- E19800 → charitable_cash_donations

- E20100 → charitable_non_cash_donations

- E03240 → domestic_production_ald

- E03400 → early_withdrawal_penalty

- E03220 → educator_expense

- E00200 → employment_income

- E26390 - E26400 → estate_income

- T27800 → farm_income

- E27200 → farm_rent_income

- E03290 → health_savings_account_ald

- E19200 → interest_deduction

- P23250 → long_term_capital_gains

- E24518 → long_term_capital_gains_on_collectibles

- E20400 → misc_deduction

- E00600 - E00650 → non_qualified_dividend_income

- E00650 → qualified_dividend_income

- E03230 → qualified_tuition_expenses

- E18500 → real_estate_taxes

Weight Adjustment   S006 weights are divided by 100 to convert to population units.

4.4.5   Data Cleaning

The preprocessing includes:

- Removal of aggregate records (MARS = 0)

- Missing value imputation with zeros

- Construction of unique household identifiers from RECID

- Assignment of household weights from S006

- Extraction of exemption counts from XTOT

4.5 Data Aging and Indexing

The process of projecting historical microdata involves both demographic aging and economic indexing based on US government forecasts. Our aging process occurs in two stages: first to reach our baseline year (2024), and then to project the calibrated dataset forward.

4.5.1 Growth Factor Construction

For each variable in the tax-benefit system with a specified growth parameter, we compute change factors from the base year through 2034:

$$\text{Index Factor}_t = \frac{\text{Index}_t}{\text{Index}_{\text{base}}}$$

4.5.2 Population Adjustment

Most economic variables are adjusted for changes in total population:

$$\text{Per Capita Factor}_t = \frac{\text{Index Factor}_t}{\text{Population Growth}_t}$$

Exceptions include:

- Weight variables maintain raw growth

- Population itself uses Census projections directly

4.5.3 Data Sources

Projection factors come from:

- Congressional Budget Office economic projections

- Census Bureau population estimates

- Social Security Administration wage index forecasts

- Treasury tax parameter indexing

4.5.4 Initial Aging Implementation

For any variable y, the projected value to reach our baseline year is computed as:

$$y_{2024} = y_{2023} \cdot \frac{f(2024)}{f(2023)}$$

where f(t) represents the index factor for time t.

4.5.5 Forward Projection

After constructing and calibrating the enhanced 2024 dataset, we project it to future years using the same indexing framework. This maintains the dataset's enhanced distributional properties while reflecting:

- Economic growth forecasts for monetary variables

- Statutory adjustments to program parameters

- Population projections applied to household weights

4.6 Quantile Regression Forests

Our implementation uses quantile regression forests (QRF) (Meinshausen and Ridgeway, 2006), which extend random forests to estimate conditional quantiles. We use the quantile-forest package (Zillow Group, 2024), a scikit-learn compatible implementation that provides efficient, Cython-optimized estimation of arbitrary quantiles at prediction time without retraining.

QRF works by generating an ensemble of regression trees, where each tree recursively partitions the feature space. Unlike standard random forests that only store mean values in leaf nodes, QRF maintains the full empirical distribution of training observations in each leaf. To estimate conditional quantiles, the model identifies relevant leaf nodes for new observations, aggregates the weighted empirical distributions across all trees, and computes the desired quantiles from the combined distribution.

The key advantages over traditional quantile regression include QRF's ability to capture non-linear relationships without explicit specification, model heteroscedastic variance across the feature space, estimate any quantile without retraining, and maintain the computational efficiency of random forests.

4.6.1 PUF Integration: Synthetic Record Generation

Unlike our other QRF applications, we use the PUF to generate an entire synthetic CPS-structured dataset. This process begins by training QRF models on PUF records with demographic variables. We then generate a complete set of synthetic CPS-structured records using PUF tax information, which are stacked alongside the original CPS records. The reweighting procedure ultimately determines the optimal mixing between CPS and PUF-based records.

This approach preserves CPS's person-level detail crucial for modeling various aspects of the tax system. These include state tax policies, benefit program eligibility, age-dependent federal provisions (such as Child Tax Credit variations by child age), and family structure interactions.

4.6.2 Direct Variable Imputation

For other enhancement needs, we use QRF to directly impute missing variables. When imputing housing costs from ACS records, we incorporate a comprehensive set of predictors

including household head status, age, sex, tenure type, various income sources (employment, self-employment, Social Security, and pension), state, and household size.

To support analysis of lookback provisions, we impute prior year earnings using consecutive-year ASEC records. This imputation relies on current employment and self-employment income, household weights, and income imputation flags from the CPS ASEC panel.

4.6.3   Implementation Details

Our QRF implementation, housed in utils/qrf.py, provides a robust framework for model development and deployment. The implementation handles categorical variable encoding and ensures consistent feature ordering across training and prediction. It also manages distribution sampling and model persistence, enabling efficient reuse of trained models.

4.7   Loss Matrix Construction

The loss matrix measures deviation from 570 administrative targets sourced from IRS Statistics of Income (SOI), Census population estimates, CBO projections, and other administrative data.

4.7.1   IRS Statistics of Income Targets

For each combination of AGI bracket and filing status, we create targets for:

- Adjusted gross income

- Count of returns

- Employment income

- Business net profits

- Capital gains (gross)

- Ordinary dividends

- Partnership and S-corporation income

- Qualified dividends

- Taxable interest income

- Total pension income

- Total social security

For aggregate-level targets only, we track:

- Business net losses

- Capital gains distributions

- Capital gains losses

- Estate income and losses

- Exempt interest

- IRA distributions

- Partnership and S-corporation losses

- Rent and royalty net income and losses

- Taxable pension income

- Taxable social security

- Unemployment compensation

4.7.2 Census Population Targets

From Census population projections (np2023_d5_mid.csv), we include:

- Single-year age population counts from age 0 to 85

- Filtered to total population (SEX = 0, RACE_HISP = 0)

- Projected to the target year

4.7.3 CBO Program Totals

From CBO projections, we calibrate:

- Income tax

- SNAP benefits

- Social security benefits

- SSI payments

- Unemployment compensation

4.7.4 EITC Statistics

From Treasury EITC data (eitc.csv), we target:

- EITC recipient counts by number of qualifying children

- Total EITC amounts by number of qualifying children

  The EITC values are uprated by:

- EITC spending growth for amounts

- Population growth for recipient counts

17

4.7.5 CPS-Derived Statistics

We calibrate to hardcoded totals for:

- Health insurance premiums without Medicare Part B: $385B

- Other medical expenses: $278B

- Medicare Part B premiums: $112B

- Over-the-counter health expenses: $72B

- SPM unit thresholds sum: $3,945B

- Child support expense: $33B

- Child support received: $33B

- SPM unit capped work childcare expenses: $348B

- SPM unit capped housing subsidy: $35B

- TANF: $9B

- Alimony income: $13B

- Alimony expense: $13B

- Real estate taxes: $400B

- Rent: $735B

4.7.6 Market Income Targets

From IRS SOI PUF estimates:

- Total negative household market income: -$138B

- Count of households with negative market income: 3M

4.7.7 Healthcare Spending by Age

Using healthcare_spending.csv, we target healthcare expenditures by:

- 10-year age groups

- Four expense categories:

    - Health insurance premiums without Medicare Part B
    - Over-the-counter health expenses
    - Other medical expenses
    - Medicare Part B premiums

4.7.8 AGI by SPM Threshold

From spm_threshold_agi.csv, we target:

- Adjusted gross income totals by SPM threshold decile

- Count of households in each SPM threshold decile

4.7.9 Target Validation

The loss matrix construction enforces two key checks:

- No missing values in any target row

- No NaN values in the targets array

4.8 Reweighting Procedure

We optimize household weights using gradient descent through PyTorch (Paszke et al., 2019).

4.8.1 Problem Formulation

Given a loss matrix M of household characteristics and a target vector t, we optimize the log-transformed weights w to minimize:

$$L(w) = \text{mean}\left(\left(\frac{w^T M + 1}{t + 1} - 1\right)^2\right)$$

where:

- w are the log-transformed weights (requires grad=True)

- M is the loss matrix in tensor form (float32)

- t are the targets in tensor form (float32)

4.8.2 Optimization Implementation

The procedure follows these steps:

1. Initialize with log-transformed original weights

2. Create a PyTorch session with retries for robustness

3. Use Adam optimizer with learning rate 0.1

4. Apply dropout (5% rate) during optimization

5. Run for 5,000 iterations or until convergence

4.8.3 Dropout Application

The dropout process:

- Randomly masks p% of weights each iteration (p = 5)

- Replaces masked weights with mean of unmasked weights

- Returns original weights if dropout rate is 0

4.8.4 Convergence Monitoring

For each iteration:

- Track initial loss value as baseline

- Compute relative change from starting loss

- Display progress with current loss values

4.8.5 Error Handling

The implementation includes checks for:

- NaN values in weights

- NaN values in loss matrix

- NaN values in loss computation

- NaN values in relative error calculation

If any check fails, the procedure raises a ValueError with diagnostic information.

4.8.6 Weight Recovery

The final weights are recovered by:

- Taking exponential of optimized log weights

- Converting from torch tensor to numpy array

4.9 Complete Enhancement Pipeline

The full dataset enhancement process proceeds through the following sequential steps, each documented in the preceding sections:

4.9.1 Initial Data Loading

1. Load CPS ASEC public use file

2. Load IRS public use file (if before 2021)

3. Load external validation sources (SOI, Census data)

4.9.2 PUF Processing (Pre-2021)

1. Apply variable renaming and recoding

2. Construct derived variables

3. Split income components

4. Remove aggregate records

5. Generate synthetic demographic variables

4.9.3 CPS Enhancement

1. Project both datasets to target year using uprating factors

2. Train quantile regression forests on PUF records

3. Generate synthetic CPS-structured records

4. Stack synthetic records with original CPS

4.9.4 Final Calibration

1. Construct loss matrix from administrative targets

2. Initialize weights for original and synthetic records

3. Optimize weights using gradient descent with dropout

4. Validate results against external benchmarks

Each step is implemented in separate modules:

- PUF processing in datasets/puf/puf.py

- Uprating in utils/uprating.py

- QRF models in utils/qrf.py

- Loss matrix in utils/loss.py

- Weight optimization in datasets/cps/enhanced_cps.py

5 Results

We evaluate our enhanced dataset against administrative targets by constructing a loss matrix (defined in utils/loss.py) measuring relative deviations from:

5.1 IRS Statistics of Income Targets

By AGI bracket and filing status, we track:

- Adjusted gross income totals

- Return counts

- Wages, salaries, and tips

- Business net profits and losses (separately)

- Capital gains (gross amounts and distributions)

- Ordinary dividends

- Partnership and S-corporation income and losses

- Qualified dividends

- Taxable interest income

- Pension income

- Social Security benefits

- Estate income and losses

- Tax-exempt interest

- IRA distributions

- Rent and royalty income and losses

- Taxable pension income

- Taxable Social Security

- Unemployment compensation

5.2 Census Population Targets

From Census projections:

- Population counts for each single year of age from 0 to 85

5.3 CBO Program Totals

From Congressional Budget Office projections:

- Income tax revenue

- SNAP benefit payments

- Social Security benefit payments

- SSI payments

- Unemployment compensation

5.4 EITC Statistics

From Treasury data:

- Number of returns claiming EITC by number of qualifying children

- Total EITC amounts by number of qualifying children

5.5 Other Targets

From various government sources:

- Healthcare spending by age group and type

- Child support payments

- Housing costs and subsidies

- Market income losses

The reweighting procedure minimizes the relative squared error between weighted sums of these variables and their administrative targets.

6 Discussion

7 Conclusion

References

Victoria Bryant. General description booklet for the 2015 public use tax file demographic file. Technical documentation, Statistics of Income Division, Internal Revenue Service, February 2023. URL https://drive.google.com/file/d/1WoTU7OGEjYMOOKHsHvTTH0NwCc-kN5cE/view.

Budget Lab. Tax microsimulation at the budget lab, 2024. URL https://budgetlab.yale.edu/research/tax-microsimulation-budget-lab.

Congressional Budget Office. An overview of cbo's microsimulation tax model. Technical report, Congressional Budget Office, 2018. URL https://www.cbo.gov/publication/54096.

Institute on Taxation and Economic Policy. Itep tax model overview, 2024. URL https://itep.org/itep-tax-model/.

Joint Committee on Taxation. Overview of jct revenue estimating methods. Technical Report JCX-48-23, Joint Committee on Taxation, 2023. URL https://www.jct.gov/publications/2023/jcx-48-23/.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Office of Tax Analysis. Revenue estimating models at the u.s. treasury department. Technical Report Technical Paper 12, U.S. Department of the Treasury, 2012. URL https://home.treasury.gov/system/files/131/TP-12.pdf.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Policy Simulation Library. Tax-data documentation, 2024. URL https://github.com/PSLmodels/taxdata.

Emmanuel Saez, Joel Slemrod, and Seth H Giertz. The elasticity of taxable income with respect to marginal tax rates: A critical review. *Journal of Economic Literature*, 50(1): 3–50, 2012.

Tax Foundation. Overview of the tax foundation's taxes and growth model, 2024. URL https://taxfoundation.org/research/all/federal/overview-tax-foundations-taxes-growth-model/.

Tax Policy Center. Brief description of the tax model, 2022. URL https://www.taxpolicycenter.org/resources/brief-description-tax-model. Updated March 2022.

Urban Institute. Attis microsimulation model, 2024a. URL https://www.urban.org/research-methods/attis-microsimulation-model.

Urban Institute. Trim3 project documentation: Transfer income model, version 3, 2024b. URL https://boreas.urban.org/documentation/input/Concepts%20and%20Procedures/Modifications%20to%20the%20Underlying%20Surveys.php.

Zillow Group. quantile-forest: Scikit-learn compatible quantile regression forests. https://zillow.github.io/quantile-forest/, 2024.