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

Nikhil Woodruff* Max Ghenis*

November 10, 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 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) provides comprehensive demographic and economic information for a nationally representative sample of U.S. households. For tax year 2024, our base dataset contains approximately 150,000 households representing the U.S. civilian non-institutional population.

The CPS's key strengths include:

- Rich demographic detail including age, sex, race, ethnicity, and education
- Complete household relationship matrices
- Program participation indicators

- State and sub-state geographic identifiers
- Monthly employment and labor force status

However, the CPS has known limitations for tax modeling:

- Underreporting of income, particularly at the top of the distribution
- 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, containing approximately 200,000 records. The data are extensively transformed to protect taxpayer privacy while preserving statistical properties. Our analysis uses the 2015 PUF, the most recent available, aged to 2024.

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
- Geographic detail limited to state
- No program participation information
- Privacy protections that mask extreme values

3.3 External Validation Sources

We validate our enhanced dataset against 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
- Employment status

3.3.3 Administrative Program Totals

We incorporate official totals from various agencies:

- 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 timing (calendar year vs. tax year)
- 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

5 Methodology

Our procedure transforms the Current Population Survey (CPS) into an enhanced microsimulation dataset through four key steps:

1. Project both CPS and PUF data to the target year
2. Transfer tax variable distributions from PUF to CPS records
3. Impute program participation
4. Reweight households to match administrative benchmarks

5.1 Data Projection

We project the CPS forward using a combination of economic and demographic factors. For each economic variable y, we apply:

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

where $f(t)$ represents the variable-specific growth index. We derive these indices from:

- CBO economic projections for aggregate income components
- SSA wage index forecasts for employment income
- Census population projections for demographic totals
- Treasury forecasts for tax variables

For the PUF, we first age the 2015 data to 2021 using IRS Statistics of Income data, then project to 2024 using the same indices as the CPS projection.

5.2 Tax Variable Enhancement

We transfer 47 tax variables from the PUF to the CPS using quantile regression forests. For each variable, we:

1. Train a forest on PUF records using age, sex, marital status, and existing income measures as predictors
2. Generate a distribution of predicted values for each CPS record
3. Sample from these distributions using rank preservation within demographic groups

This approach preserves both the marginal distributions of tax variables and their relationships with demographic characteristics.

5.3 Program Participation

We model participation in major benefit programs through a two-stage process:

1. Calculate eligibility using program rules
2. Assign participation probabilities based on:
    - Demographic characteristics
    - Benefit amounts
    - Geographic patterns
    - Historical take-up rates

The final participation patterns emerge from our reweighting procedure's alignment with administrative totals.

5.4 Household Reweighting

We adjust household weights to minimize discrepancies with administrative benchmarks while avoiding overfitting. The optimization problem takes the form:

$$\min_{w} \sum_{j} \left(\frac{\sum_i w_i x_{ij} - t_j}{t_j} \right)^2 + \lambda \sum_i (w_i - w_i^0)^2$$

subject to:

$$w_i \geq 0 \quad \forall i$$

where:

- w_i is the new weight for household i
- w_i^0 is the original CPS weight
- x_{ij} is the value of variable j for household i

- t_j is the administrative target for variable j
- λ controls the strength of regularization

We solve this using gradient descent with dropout, randomly zeroing 5% of household weights during each iteration to improve generalization.

The remainder of the methodology section details each component:

- Section 4.1 describes our quantile regression forest implementation
- Section 4.2 explains the reweighting optimization
- Section 4.3 presents our validation framework

6 Quantile Regression Forests

We use quantile regression forests (QRF) in two distinct ways: direct imputation of missing variables, and generation of synthetic records.

6.1 PUF Integration: Synthetic Record Generation

Unlike our other QRF applications, we use the PUF to generate an entire synthetic CPS-structured dataset:

1. Train QRF models on PUF records with demographic variables
2. Generate a complete set of synthetic CPS-structured records using PUF tax information
3. Stack these synthetic records alongside the original CPS records
4. Allow the reweighting procedure to determine optimal mixing between CPS and PUF-based records

This approach preserves CPS's person-level detail crucial for modeling:

- State tax policies
- Benefit program eligibility
- Age-dependent federal provisions (e.g., Child Tax Credit variations by child age)
- Family structure interactions

6.2 Direct Variable Imputation

For other enhancement needs, we use QRF to directly impute missing variables:

6.2.1 Housing Costs from ACS

We impute rent payments and property taxes using ACS records, with predictors including:

- Household head status
- Age
- Sex
- Tenure type
- Employment income
- Self-employment income
- Social Security income
- Pension income
- State
- Household size

6.2.2 Prior Year Income from CPS ASEC Panel

To support analysis of lookback provisions, we impute prior year earnings using consecutive-year ASEC records, using:

- Employment income
- Self-employment income
- Household weight
- Income imputation flags

6.3 Implementation Details

Our QRF implementation in utils/qrf.py handles:

- Categorical variable encoding
- Consistent feature ordering
- Distribution sampling
- Model persistence

7 Reweighting Procedure

Our reweighting process optimizes household weights to match administrative targets while determining the relative value of original CPS records versus PUF-derived synthetic records.

7.1 Loss Matrix Construction

We construct a matrix of targets including:

7.1.1 IRS Statistics of Income Targets

For each AGI bracket and filing status combination:

- Adjusted gross income totals
- Employment income
- Business income/losses
- Capital gains totals and distributions
- Dividend income (qualified and ordinary)
- Partnership and S-corporation income/losses
- Pension and IRA distributions
- Social Security benefits
- Interest income

7.1.2 Census Population Targets

Single-year age population projections from age 0 to 85+, ensuring demographic representativeness.

7.1.3 Program Totals

Annual administrative totals from:

- IRS: Income tax revenue, EITC claims and amounts by number of children
- Social Security Administration: Benefit payments
- USDA: SNAP participation and benefits
- DOL: Unemployment compensation

7.2 Optimization Approach

We minimize the relative error across all targets using gradient descent with dropout regularization:

1. Initialize with original CPS weights
2. At each iteration:
    - Randomly zero out 5% of weights (dropout)
    - Compute relative errors between weighted sums and targets
    - Update weights using Adam optimizer
3. Continue until convergence or 5,000 iterations

The core optimization uses PyTorch 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
- M is the loss matrix of household characteristics
- t are the administrative targets

7.3 Implementation Details

The Enhanced CPS implementation uses the following parameters:

- Learning rate: 0.1
- Dropout rate: 5
- Optimizer: Adam
- Maximum iterations: 5,000

7.4 Balance Between CPS and PUF Records

The reweighting procedure naturally determines the mix of original CPS and PUF-derived records by:

- Starting with equal initial weights
- Allowing the optimization to up-weight records that better match targets
- Implicitly favoring PUF-derived records for tax variables
- Maintaining CPS records' strength in demographic representation

8 Results

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

8.1 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

8.2 Census Population Targets

From Census projections:

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

8.3 CBO Program Totals

From Congressional Budget Office projections:

- Income tax revenue
- SNAP benefit payments
- Social Security benefit payments
- SSI payments
- Unemployment compensation

8.4 EITC Statistics

From Treasury data:

- Number of returns claiming EITC by number of qualifying children
- Total EITC amounts by number of qualifying children

8.5 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.

9 Discussion

10 Conclusion

References

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/.

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.

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.