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

Nikhil Woodruff* Max Ghenis*

November 13, 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

*PolicyEngine

household 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 (**?**). 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 (**?**). 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 (**?**). 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 (**?**). 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 (**?**) 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 (**?**). 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 (**?**) 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 (**?**) 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 (**?**) 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 (**?**) 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 (**?**) 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.

2.4.1 Empirical Evaluation of Enhancement Methods

Recent work has systematically compared different approaches to survey enhancement. **?** evaluated traditional techniques like percentile matching against machine learning methods including gradient descent reweighting and synthetic data generation. Their results showed ML-based approaches substantially outperforming conventional methods, with combined synthetic data and reweighting reducing error by 88% compared to baseline surveys. Importantly, their cross-validation analysis demonstrated these improvements generalized to out-of-sample targets, suggesting the methods avoid overfitting to specific statistical measures. This empirical evidence informs our methodological choices, particularly around combining multiple enhancement techniques.

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 (**?**). 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

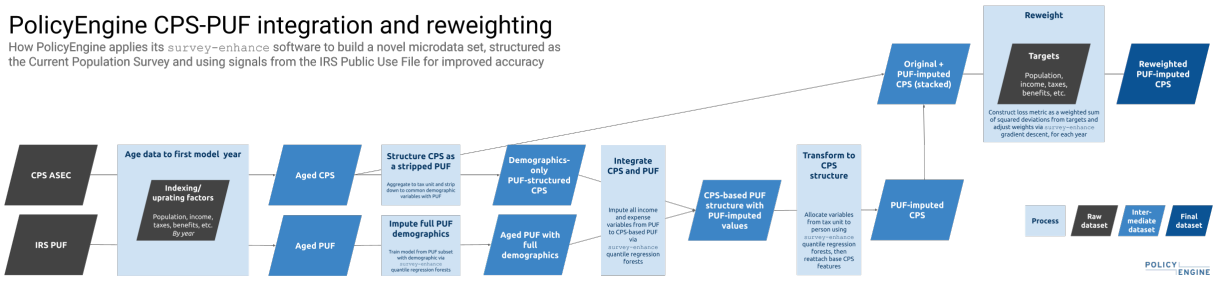

Figure 1: Data flow diagram for integrating CPS and PUF microdata. The process ages both datasets to a common year, integrates demographic and income information through quantile regression forests, and optimizes household weights using gradient descent.

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. This design is motivated by empirical evidence from **?** showing that combining synthetic data generation with weight optimization achieves substantially better results than either technique alone or traditional enhancement methods. Their comprehensive benchmarking demonstrated an 88% reduction in survey error through this combined approach, with improvements that generalized across multiple validation metrics.

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. Run the PolicyEngine US tax-benefit model on the stacked dataset to generate tax and benefit amounts
5. 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 PolicyEngine Integration and Access

The enhanced dataset is designed to integrate seamlessly with PolicyEngine, an open-source tax-benefit microsimulation platform available both as a Python package and a web application at `policyengine.org`. The platform provides comprehensive tools for analyzing tax and benefit reforms through both programmatic and web interfaces.

4.2.1 Web Interface

PolicyEngine's web interface at `policyengine.org/us` allows users to:

- Modify thousands of policy parameters across federal and state tax and benefit programs
- Analyze reforms' impacts on:
  - Government budgets (federal taxes, benefits, and state/local taxes)
  - Income distribution (gains and losses across the income spectrum)
  - Poverty (by age, race/ethnicity, and sex using the SPM)

  - Inequality (various metrics)
  - Labor supply (with customizable elasticities)
- Generate natural language summaries of policy impacts using Claude 3.5 Sonnet
- Calculate household-specific impacts by entering detailed information
- View marginal tax rates under current law and reforms

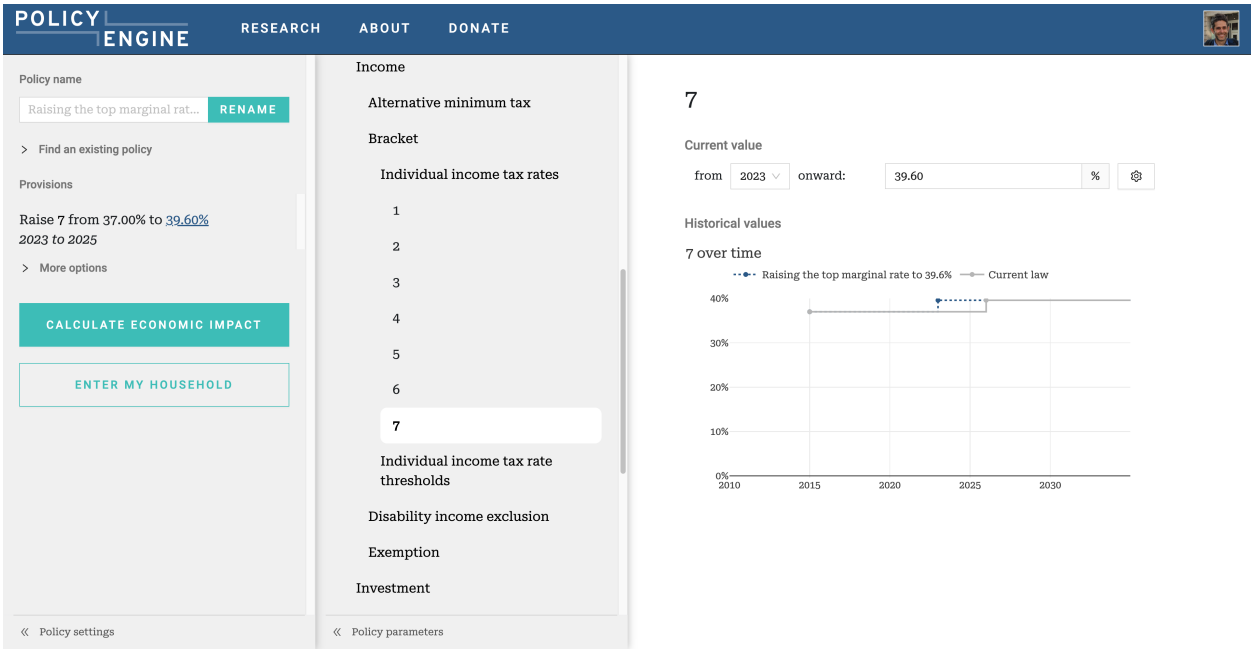

Figure 2: PolicyEngine's policy editor interface, showing modification of the top marginal tax rate.

4.2.2 Python Package

The Python package provides programmatic access with just a few lines of code:

```
from policyengine_us import Microsimulation

# Load enhanced CPS dataset
sim = Microsimulation(dataset="enhanced_cps_2024")

# Analyze a tax reform
reform = {
    "gov.irs.tax_rate.single": {
        2024: [
            {"threshold": 400_000, "rate": 0.396}
        ]
```

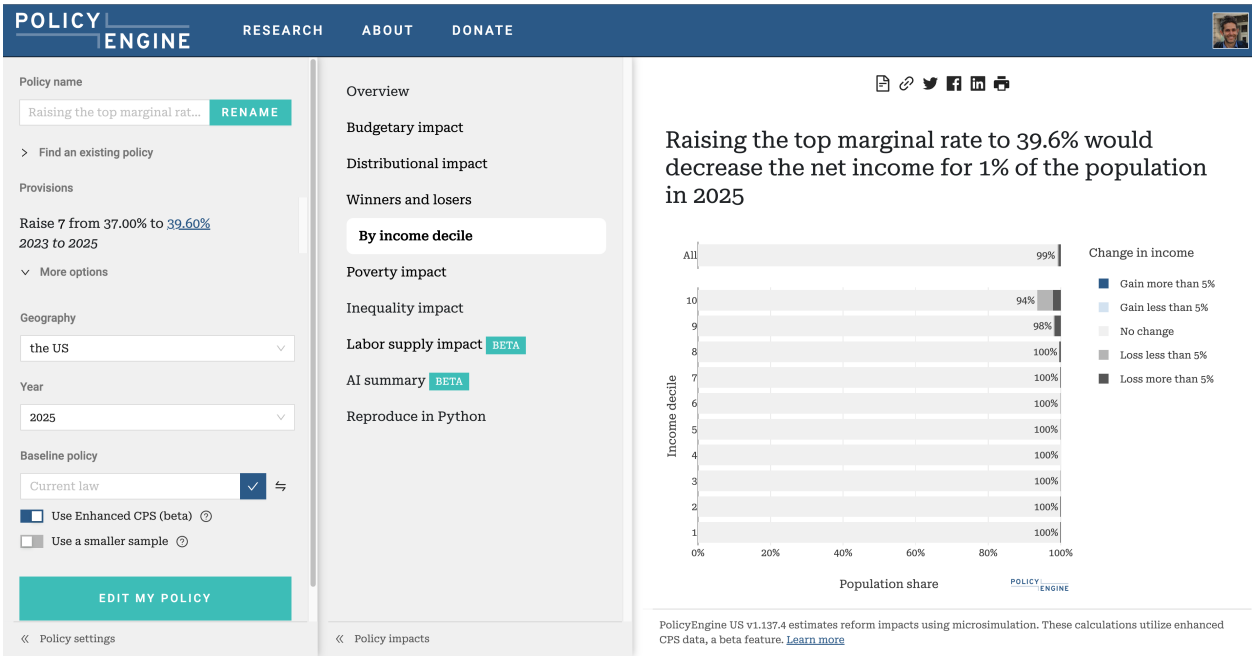

Figure 3: Example distributional analysis from PolicyEngine showing population impacts by income decile.

```
    }
}
reformed = Microsimulation(reform=reform)

# Calculate revenue impact
baseline_revenue = sim.calculate("income_tax").sum()
reform_revenue = reformed.calculate("income_tax").sum()
revenue_impact = reform_revenue - baseline_revenue

# Analyze impacts by group
income_deciles = sim.calculate("income_decile")
for decile in range(1, 11):
    mask = income_deciles == decile
    impact = reformed.calculate(
        "household_net_income",
        mask=mask
    ).mean() - sim.calculate(
        "household_net_income",
        mask=mask
    ).mean()
    print(f"Decile {decile}: ${impact:,.0f}")
```

4.2.3 International Applications

The same enhancement methodology and software infrastructure powers PolicyEngine UK, which incorporates additional data sources including the Living Costs and Food Survey (for consumption) and Wealth and Assets Survey. This demonstrates the approach's adaptability to different national contexts and data environments.

While designed for seamless integration with PolicyEngine, the enhanced dataset is also available as a standalone HDF5 file that can be used with other microsimulation frameworks, promoting broader research applications while maintaining the benefits of integrated analysis tools.

4.3 Demographic Analysis

A key advantage of building from CPS microdata is the ability to analyze policies by demographics not available in tax returns. While organizations using tax return data as their base must develop complex methods to impute race and ethnicity, our approach provides these characteristics directly from the survey data.

The Internal Revenue Service does not collect information on race or ethnicity from tax filers. Other microsimulation models have addressed this limitation through various imputation approaches:

- The Congressional Budget Office statistically matches tax returns to survey records using income and limited demographic characteristics, then validates against linked Census-IRS data (**?**)
- The Tax Policy Center creates multiple copies of each tax unit record, then uses an algorithm to reweight these copies to match aggregate race and ethnicity statistics from survey data (**?**)
- The Institute on Taxation and Economic Policy assigns each tax record probabilities of different racial and ethnic identities based on characteristics like income, marital status, state, and homeownership (**?**)

These approaches require complex statistical methods and face inherent limitations in accuracy, particularly when analyzing subgroups or policy impacts that may vary by demographic characteristics not used in the imputation process.

In contrast, our approach provides race and ethnicity variables directly from the CPS without requiring complex imputation. This offers several advantages:

- Race and ethnicity are observed rather than imputed, avoiding potential biases from statistical matching
- Demographic information is available at the individual level, not just for tax unit heads
- The same enhancement methodology can be applied to analyze other demographic characteristics like disability status and educational attainment

- Interactions between demographics can be analyzed naturally (e.g., poverty impacts by both race and age)

This capability enables more reliable analysis of how tax and benefit policies affect different demographic groups. For example, using the enhanced CPS we can directly examine how the Earned Income Tax Credit's benefits vary by race and ethnicity, or analyze the distributional effects of Child Tax Credit reforms across both income levels and demographic categories.

4.4 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.

4.5 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

4.5.1 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

4.5.2 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

4.5.3 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.

4.6 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.

4.6.1 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%

4.6.2 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.

4.6.3 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

4.6.4 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.

4.6.5 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

4.7 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.

4.7.1 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:

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

4.7.2 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

4.7.3 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

4.7.4 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.

4.7.5 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

4.8 Quantile Regression Forests

Our implementation uses quantile regression forests (QRF) (**?**), which extend random forests to estimate conditional quantiles. Building on **?**, we use the quantile-forest package (**?**), 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.

4.8.1 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.

4.8.2 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.

4.8.3 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.

4.9 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.

4.9.1 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

4.9.2 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

4.9.3 CBO Program Totals

From CBO projections, we calibrate:

- Income tax
- SNAP benefits
- Social security benefits
- SSI payments
- Unemployment compensation

4.9.4 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

4.9.5 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

4.9.6 Market Income Targets

From IRS SOI PUF estimates:

- Total negative household market income: -$138B
- Count of households with negative market income: 3M

4.9.7 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

4.9.8 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

4.9.9 Target Validation

The loss matrix construction enforces two key checks:

- No missing values in any target row
- No NaN values in the targets array

4.10 Reweighting Procedure

We optimize household weights using gradient descent through PyTorch (**?**).

4.10.1 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)

4.10.2 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

4.10.3 Dropout Application

We apply dropout regularization during optimization to prevent overfitting:

- 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

4.10.4 Convergence Monitoring

For each iteration:

- Track initial loss value as baseline
- Compute relative change from starting loss
- Display progress with current loss values

4.10.5 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.

4.10.6 Weight Recovery

The final weights are recovered by:

- Taking exponential of optimized log weights
- Converting from torch tensor to numpy array

4.11 Complete Enhancement Pipeline

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

4.11.1 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)

4.11.2 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

4.11.3 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

4.11.4 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 validate our enhanced dataset against a comprehensive set of official statistics and compare its performance to both the original CPS and PUF datasets. Our validation metrics cover 570 distinct targets spanning demographic totals, program participation rates, and detailed income components across the distribution.

5.1 Validation Against Administrative Totals

The enhanced CPS (ECPS) shows substantial improvements over both of its source datasets. When comparing absolute relative errors across all targets, the ECPS outperforms:

- The Census Bureau's CPS in 63.0% of targets
- The IRS Public Use File in 70.7% of targets

These improvements are particularly notable because they demonstrate that our enhancement methodology successfully combines the strengths of both source datasets while mitigating their individual weaknesses. The CPS excels at demographic representation but struggles with income reporting, particularly at the top of the distribution. Conversely, the PUF captures tax-related variables well but lacks demographic detail. Our enhanced dataset achieves better accuracy than either source across most metrics.

5.2 Income Distribution

Table 1 compares distributional statistics across all three datasets, measuring net income after taxes and transfers at the tax unit level without equivalization:

Table 1: Key tax unit-level distributional metrics

Metric	CPS	Enhanced CPS	PUF
Gini coefficient	0.495	0.572	0.570
Top 10% share	0.361	0.425	0.410
Top 1% share	0.085	0.154	0.150

The enhanced CPS achieves very similar distributional statistics to the PUF, with a Gini coefficient of 0.572 compared to 0.570 in the PUF, and nearly identical top income shares. This suggests our enhancement procedure successfully incorporates the PUF's more accurate representation of the income distribution.

For applications requiring household-level analysis, Table 2 shows these same metrics calculated over households rather than tax units:

Table 2: Key household-level distributional metrics

Metric	CPS	Enhanced CPS
Gini coefficient	0.449	0.556
Top 10% share	0.323	0.416
Top 1% share	0.073	0.154

The household-level metrics show similar patterns of increased inequality capture compared to the baseline CPS, though the magnitudes differ due to the different unit of analysis.

5.3 Poverty Measurement

The poverty metrics in Table 3 warrant particular attention:

While we calibrate to income, demographic, tax, and benefit totals that should capture the overall income distribution, the substantial increase in measured poverty from 12.7% to 24.9% suggests our current approach may need refinement. The enhancement procedure might overstate poverty by not fully capturing geographic correlations between incomes and poverty thresholds. We currently calibrate AGI totals by SPM thresholds from the CPS and intend to enhance this calibration in future iterations.

Table 3: Poverty metrics

Metric	CPS	Enhanced CPS
SPM poverty rate	0.127	0.249

5.4 Weight Distribution Analysis

The weight distribution statistics shown in Table 4 reveal notable differences in how the datasets represent the population. Note that CPS and enhanced CPS values reflect household weights, while PUF values reflect tax unit weights:

Table 4: Weight distribution statistics

Statistic	CPS	Enhanced CPS	PUF
Mean weight	2,379.1	1,290.5	776.1
Median weight	2,260.3	1.0	353.5
Nonzero weight share	1.000	0.538	1.000
Weight std. dev.	1,422.5	11,869.0	720.3

The original CPS has relatively uniform weights centered around 2,400, reflecting its design as a representative sample. The PUF shows less variation in weights, with a standard deviation of 720 compared to 1,423 for the CPS. The enhanced CPS exhibits greater weight variation and lower mean weights by design - each original record is cloned to potentially incorporate a matching PUF record, effectively doubling the initial sample size. About 54% of these expanded records receive non-zero weights in the final dataset, as the reweighting procedure selects optimal combinations of records to match administrative targets.

A detailed, interactive validation dashboard showing performance across all targets is maintained at https://policyengine.github.io/policyengine-us-data/validation.html and updates automatically with each dataset revision. This transparency allows users to assess the dataset's strengths and limitations for their specific use cases.

5.5 Example Policy Reform: Top Tax Rate Increase

To demonstrate the enhanced dataset's value for policy analysis, we examine President Biden's 2025 budget proposal to raise the top marginal income tax rate from 37% to 39.6%. This would restore the pre-Tax Cuts and Jobs Act rate for high-income taxpayers, applying to income above $400,000 for single filers, $425,000 for head of household filers, $450,000 for married joint filers, and $225,000 for married separate filers, with thresholds indexed to inflation from 2025.

Table 5 compares revenue projections across datasets:

The enhanced CPS projects $75.7 billion in additional revenue for 2025, closely matching the Treasury Department's estimate of $75.4 billion. In contrast, the baseline CPS projects only $28.7 billion, substantially understating the reform's impact. This disparity illustrates how the enhanced dataset's improved capture of high incomes enables more accurate modeling of tax policies targeting high-income households.

Table 5: Projected revenue from top rate increase, 2025

Source	Revenue (billions)
Treasury	$75.4
Enhanced CPS	$75.7
Baseline CPS	$28.7

6 Discussion

This paper introduces a novel approach to constructing an enhanced microsimulation dataset by integrating survey and administrative data sources. Our methodology, which combines quantile regression forests (QRF) and dropout-regularized gradient descent reweighting, demonstrates substantial improvements in accurately capturing both demographic and tax-related variables. In this section, we discuss the strengths, limitations, potential applications, and future directions of this approach.

6.1 Strengths of the Enhanced Dataset

The enhanced dataset achieves a unique balance between demographic detail and tax precision, addressing a long-standing gap in microsimulation modeling. The use of QRF allows for more accurate transfer of income and tax distributions from the IRS Public Use File (PUF) to the Current Population Survey (CPS), preserving complex variable relationships that are critical for policy analysis. Additionally, the dropout-regularized gradient descent reweighting effectively calibrates household weights to align with administrative benchmarks, reducing error rates across a broad range of demographic and economic metrics.

Our validation results show that the enhanced CPS (ECPS) improves on both source datasets, particularly in tax-related variables that are essential for analyzing income distributions and program participation. By providing a publicly available, open-source dataset with extensive validation against external benchmarks, we support more transparent and reliable policy analysis.

6.2 Limitations and Potential Biases

Despite these strengths, the enhanced dataset has limitations that merit careful consideration. One key challenge lies in maintaining consistency in relationships across diverse variables, especially in cases where nonlinear or unexpected correlations exist. Although QRF is well-suited for capturing non-linear relationships, biases may still arise due to assumptions made during variable imputation.

A second limitation is the reliance on older IRS data, which may not fully capture recent demographic and economic shifts. While our reweighting procedure attempts to mitigate this through adjustment to more current administrative targets, future iterations could benefit from updated IRS data or alternative administrative sources that better reflect the contemporary population.

Further, our approach may introduce biases when aligning household records with administrative targets. These biases can impact analyses that depend heavily on small demographic subgroups or specific income brackets. Future improvements could involve fine-tuning the reweighting process to minimize potential overfitting in cases where data are sparse.

6.3 Applications of the Enhanced Dataset

The enhanced CPS dataset expands the scope and accuracy of microsimulation analyses in several policy domains. By combining the CPS's household structure with the PUF's tax precision, this dataset is well-suited for both federal and state-level tax analysis, particularly in modeling income-based benefits and tax credits. Researchers and policymakers could leverage this dataset to evaluate the distributional impacts of various tax reforms, analyze the implications of benefit programs across income levels, and assess policy proposals that rely on a precise understanding of income and demographic characteristics.

Additional applications extend to labor market studies, health policy analysis, and state-specific program evaluations. With further adaptation, the methodology could also support microsimulation in international contexts, providing a flexible tool for policy modeling across diverse regions and socioeconomic conditions.

6.4 Future Directions

Building on the success of this methodology, future work could aim to expand the dataset's geographic granularity and incorporate additional data sources. Integrating state-specific datasets or additional federal data on healthcare and education would further enrich the dataset's utility for policy analysis. Moreover, the dataset could benefit from continued refinement of its reweighting procedure, including the use of ensemble methods to capture a broader range of variable interactions.

Another promising direction involves the development of interactive tools that allow researchers and policymakers to explore the dataset in real time, enhancing transparency and accessibility. By providing both the enhanced dataset and the codebase as open-source resources, we establish a foundation for collaborative improvement and iterative updates that respond to changing policy needs and data availability.

7 Conclusion

This paper presents a novel approach to constructing enhanced microdata for tax-benefit microsimulation by combining survey and administrative data sources. Our methodology leverages machine learning techniques – specifically quantile regression forests and gradient descent optimization – to preserve the strengths of each source while mitigating their weaknesses. The resulting dataset outperforms both the Current Population Survey and IRS Public Use File across a majority of validation targets, with particularly strong improvements in areas crucial for policy analysis such as income distributions and program participation rates.

The enhanced dataset addresses a key challenge in tax-benefit microsimulation: the need for both detailed demographic information and accurate tax/income data. By maintaining the CPS's rich household structure while incorporating the PUF's tax precision, our approach enables more reliable analysis of policies that depend on both demographic characteristics and economic circumstances. The systematic validation against hundreds of administrative targets provides confidence in the dataset's reliability while helping users understand its limitations.

Our open-source implementation and automatically updated validation metrics establish a new standard for transparency in microsimulation data enhancement. This enables other researchers to build upon our work, adapt the methodology to other jurisdictions, or extend it to incorporate additional data sources. Future work could expand the approach to finer geographic levels, integrate data from additional surveys, or apply similar techniques to other domains requiring the combination of survey and administrative data.

The enhanced CPS represents a significant advance in the quality of openly available microdata for tax-benefit analysis. By reducing error rates across a broad range of metrics while preserving essential relationships in the data, it provides a more reliable foundation for understanding the impacts of complex policy reforms on American households.