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

Nikhil Woodruff[*] Max Ghenis[*]

July 28, 2025

Abstract

We present a methodology for creating enhanced microsimulation datasets by combining the Current Population Survey (CPS) with the IRS Public Use File (PUF). Our two-stage approach uses quantile regression forests to impute 72 tax variables from the PUF onto CPS records, preserving distributional characteristics while maintaining household composition and member relationships. The imputation process alone does not guarantee consistency with official statistics, necessitating a reweighting step to align the combined dataset with known population totals and administrative benchmarks. We apply a reweighting algorithm that calibrates the dataset to over 7,000 targets from six sources: IRS Statistics of Income, Census population projections, Congressional Budget Office program estimates, Treasury expenditure data, Joint Committee on Taxation tax expenditure estimates, and healthcare spending patterns. The reweighting employs dropout-regularized gradient descent optimization to ensure consistency with administrative benchmarks. Validation shows the enhanced dataset reduces error in key tax components by [TO BE CALCULATED]% relative to the baseline CPS. The dataset maintains the CPS's demographic detail and geographic granularity while incorporating tax reporting data from administrative sources. We release the enhanced dataset, source code, and documentation to support policy analysis.

1 Introduction

Microsimulation models require high-quality microdata that accurately represents both demographic characteristics and economic outcomes. The ideal dataset would combine the demographic richness and household structure of surveys with the income precision of administrative tax records. However, publicly available datasets typically excel in one dimension while lacking in the other.

[*] PolicyEngine

The Current Population Survey (CPS) Annual Social and Economic Supplement provides detailed household demographics, family relationships, and program participation data for a representative sample of US households. However, it suffers from well-documented income underreporting, particularly at the top of the distribution. The IRS Public Use File (PUF) contains accurate tax return information but lacks household structure, demographic detail, and state identifiers needed for comprehensive policy analysis.

This paper presents a methodology for creating an Enhanced CPS dataset that combines the strengths of both sources. Through a two-stage enhancement process—imputation followed by reweighting—we create a dataset suitable for analyzing both tax and transfer policies at federal and state levels.

1.1 Related Work

Several approaches have been developed to address the limitations of survey data for microsimulation. Statistical matching techniques have long been used to combine datasets. Radner (1978) pioneered exact matching methods for combining survey and administrative data, while Rodgers (1984) developed statistical matching based on common variables. More recently, D'Orazio et al. (2006) provided a comprehensive framework for modern statistical matching methods.

Economic studies have addressed dataset limitations through various strategies. The Congressional Budget Office combines CPS data with tax return information through statistical matching (CBO, 2022). The Tax Policy Center creates synthetic datasets by statistically matching the CPS to a subset of tax returns (Rohaly et al., 2005). However, these approaches often sacrifice either demographic detail or tax precision, limiting their utility for comprehensive policy analysis.

Reweighting methods to align survey data with administrative totals have been employed by statistical agencies and researchers. The Luxembourg Income Study uses calibration to improve cross-national comparability (Gornick and Jäntti, 2013). The Urban-Brookings Tax Policy Center employs reweighting in their microsimulation model but relies on proprietary data that cannot be shared publicly (Khitatrakun et al., 2016).

Our approach differs from previous efforts in three key ways. First, we employ quantile regression forests to preserve distributional characteristics during imputation, improving upon traditional hot-deck and regression-based methods that may distort variable relationships. We conduct robustness checks comparing QRF performance to gradient boosting and neural network approaches, finding QRF provides the best balance of accuracy and interpretability. Second, we calibrate to over 7,000 targets from multiple administrative sources, far exceeding the scope of previous calibration efforts which typically use fewer than 100 targets. Third, we provide a fully open-source implementation enabling reproducibility and collaborative improvement, addressing the transparency limitations of existing proprietary models.

1.2 Contributions

This paper makes three main contributions to the economic and public policy literature. Methodologically, we demonstrate how quantile regression forests can effectively impute detailed tax variables while preserving their joint distribution and relationship to demographics. This advances the statistical matching literature by showing how modern machine learning methods can overcome limitations of traditional hot-deck and parametric approaches. The preservation of distributional characteristics is particularly important for tax policy analysis where outcomes often depend on complex interactions between income sources and household characteristics.

Our empirical contribution involves creating and validating a publicly available enhanced dataset that addresses longstanding data limitations in microsimulation modeling. By combining the demographic richness of the CPS with the tax precision of the PUF, we enable analyses that were previously infeasible with public data. The dataset's calibration to over 7,000 administrative targets ensures consistency with official statistics across multiple dimensions simultaneously.

From a practical perspective, we provide open-source tools and comprehensive documentation that enable researchers to apply these methods, modify the approach, or build upon our work. This transparency contrasts with existing proprietary models and supports reproducible research. Government agencies could use our framework to enhance their own microsimulation capabilities, while academic researchers gain access to data suitable for analyzing distributional impacts of tax and transfer policies. The modular design allows incremental improvements as new data sources become available.

The remainder of this paper is organized as follows. Section 2 describes our data sources including the primary datasets and calibration targets. Section 3 details the enhancement methodology including both the imputation and reweighting stages. Section 4 presents validation results comparing performance across datasets. Section 5 discusses limitations, applications, and future directions. Section 6 concludes with implications for 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 **??**. 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 **??**.

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 **??**.

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 ??.

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 **??**, 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 Sources

Our methodology combines two primary data sources with calibration targets from six administrative sources.

3.1 Primary Data Sources

3.1.1 Current Population Survey (CPS)

The Current Population Survey Annual Social and Economic Supplement (ASEC) serves as our base dataset. Conducted jointly by the Census Bureau and Bureau of Labor Statistics, the CPS ASEC surveys approximately 75,000 households annually.

The CPS provides several essential features for microsimulation modeling. It offers a representative sample of US households with detailed demographic information including age, education, race, and employment status. The survey captures family and household relationships through a comprehensive set of relationship codes that allow reconstruction of tax units and benefit units. Geographic identifiers down to the state level enable subnational policy analysis. The survey includes detailed questions about program participation in major transfer programs like SNAP, Medicaid, and housing assistance. Income data is collected by source, distinguishing between wages, self-employment, interest, dividends, and transfers.

However, the CPS faces well-documented limitations that necessitate enhancement. Income underreporting is particularly severe at high income levels, with Rothbaum and Bee (2021) finding that the CPS captures only 50% of top incomes compared to tax records. The survey provides limited tax detail, lacking information on itemized deductions, tax credits, and capital gains realizations that are crucial for revenue estimation. High income values are topcoded to protect confidentiality, further limiting the ability to analyze tax policies affecting high earners. The survey's focus on cash income means it misses important non-cash compensation like employer-provided health insurance premiums.

3.1.2 IRS Public Use File (PUF)

The IRS Statistics of Income Public Use File contains detailed tax return information from a stratified sample of individual income tax returns. The most recent PUF available is from tax year 2015, containing approximately 230,000 returns.

The PUF provides tax-related variables drawn directly from filed tax returns. It provides detailed breakdowns of income by source including wages, interest, dividends, capital gains, business income, and retirement distributions. The file contains complete information on itemized deductions such as mortgage interest, state and local taxes, and charitable contributions. All tax credits claimed by filers are included, from the earned income tax credit to education credits. The stratified sampling design oversamples high-income returns, providing better coverage of the income distribution's upper tail than survey data. Sampling weights allow researchers to produce population-representative estimates.

Despite these strengths, the PUF has significant limitations for comprehensive policy analysis. The file contains minimal demographic information, limited to filing status and exemptions claimed. Geographic identifiers are removed to protect taxpayer privacy, preventing state-level analysis. The population excludes non-filers, who represent approximately 20% of adults and are disproportionately low-income. The substantial time lag means the most recent data is nine years old as of 2024, missing recent economic and demographic changes. Perhaps most critically, the PUF lacks household structure, preventing analysis of how tax policies interact with transfer programs that operate at the household level.

3.2 Additional Data Sources for Imputation

Beyond the PUF, we incorporate data from three additional surveys to impute specific variables missing from the CPS:

3.2.1 Survey of Income and Program Participation (SIPP)

The SIPP provides detailed income and program participation data. We use SIPP to impute:

- **Tip income**: Using a Quantile Regression Forest model trained on SIPP data, we impute tip income based on employment income, age, and household composition

3.2.2 Survey of Consumer Finances (SCF)

The SCF provides comprehensive wealth and debt information. We use SCF to impute:

- **Auto loan balances**: Matched based on household demographics and income
- **Interest on auto loans**: Calculated from imputed balances
- **Net worth components**: Various wealth measures not available in CPS

The SCF imputation uses their reference person definition (male in mixed-sex couples or older person in same-sex couples) to ensure proper matching.

3.2.3 American Community Survey (ACS)

The ACS provides detailed housing and geographic data. We use ACS to impute:

- **Property taxes**: For homeowners, imputed based on state, household income, and demographics
- **Rent values**: For specific tenure types where CPS data is incomplete
- **Housing characteristics**: Additional housing-related variables

These imputations use Quantile Regression Forests to preserve distributional characteristics while accounting for household heterogeneity.

3.3 Calibration Data Sources

We calibrate the enhanced dataset to over 7,000 targets from six authoritative sources:

3.3.1 IRS Statistics of Income (SOI)

Annual tabulations from tax returns provide income distributions by:

- Adjusted Gross Income (AGI) bracket
- Filing status
- Income type

We use SOI Table 1.4 which cross-tabulates income components by AGI ranges, creating over 5,300 distinct targets.

3.3.2 Census Population Projections

National and state-level demographic targets from:

- Single-year-of-age populations (ages 0-85)
- State total populations
- State populations under age 5

3.3.3 Congressional Budget Office (CBO)

Program participation and revenue projections:

- SNAP (Supplemental Nutrition Assistance Program)
- Social Security benefits
- Supplemental Security Income (SSI)
- Unemployment compensation
- Individual income tax revenue

3.3.4 Joint Committee on Taxation (JCT)

Tax expenditure estimates for major deductions:

- State and local tax deduction: $21.2 billion
- Charitable contribution deduction: $65.3 billion
- Mortgage interest deduction: $24.8 billion
- Medical expense deduction: $11.4 billion

3.3.5 Treasury Department

Additional program totals:

- Earned Income Tax Credit by number of children
- Total EITC expenditure

3.3.6 Healthcare Spending Data

Age-stratified medical expenditures:

- Health insurance premiums (excluding Medicare Part B)
- Medicare Part B premiums
- Other medical expenses
- Over-the-counter health expenses

3.4 Addressing the Temporal Gap

The nine-year gap between the 2015 PUF and 2024 CPS presents a methodological challenge. Economic conditions, tax law, and demographic patterns have changed significantly since 2015. We address this temporal inconsistency through several approaches. Dollar amounts in the PUF are uprated using income-specific growth factors from IRS Statistics of Income publications, ensuring that income levels reflect current economic conditions. The calibration process forces the combined dataset to match contemporary administrative totals, partially compensating for demographic shifts. However, structural changes in the economy, such as the growth of gig work or shifts in retirement patterns, may not be fully captured. Users should consider this limitation when analyzing policies sensitive to recent economic trends.

3.5 Data Preparation

3.5.1 CPS Processing

We use the CPS ASEC from survey year 2024 (covering calendar year 2023 income). The Census Bureau provides:

- Person-level records with demographics
- Hierarchical identifiers linking persons to families and households
- Initial survey weights

3.5.2 PUF Processing

The 2015 PUF requires several adjustments:

- Dollar amounts uprated using SOI growth factors by income type
- Records filtered to remove those with insufficient data
- Weights normalized to represent the filing population

3.5.3 Target Preparation

Administrative targets are collected for the appropriate year:

- Most targets use 2024 projections
- Historical data uprated using official growth rates
- State-level targets adjusted for population changes

3.5.4 Variable Harmonization

Combining datasets requires careful harmonization of variable definitions and concepts. Tax units in the PUF must be mapped to CPS households, accounting for multi-generational households and unmarried partners. Income concepts differ between sources, with the PUF using tax definitions while the CPS follows survey conventions. For example, the PUF reports taxable Social Security benefits while the CPS reports total benefits received. We harmonize these differences by using PolicyEngine's tax calculator to compute tax concepts from CPS variables before imputation.

Time periods also require harmonization. The CPS collects income for the previous calendar year while asking about current-year program participation. The PUF reports tax year data with some income received in different calendar years. We align all amounts to a common tax year basis using payment timing assumptions documented in our code repository.

The combination of these data sources enables us to create a dataset that maintains the CPS's demographic richness while achieving tax reporting accuracy comparable to administrative data.

4 Methodology

The Enhanced CPS dataset is created through a two-stage process: imputation followed by reweighting. This approach leverages the strengths of both data sources while mitigating their individual limitations. The imputation stage uses Quantile Regression Forests to transfer 72 tax variables from the PUF onto CPS records, creating what we call the Extended CPS. The reweighting stage then optimizes household weights to match over 7,000 administrative targets, producing the final Enhanced CPS with weights calibrated to official statistics. A visual overview of this process is provided in Appendix Figure A1.

4.1 Stage 1: Variable Imputation

We impute missing variables from multiple data sources using Quantile Regression Forests (QRF). This includes both tax variables from the PUF and additional variables from SIPP, SCF, and ACS.

4.1.1 Quantile Regression Forests

We use Quantile Regression Forests (QRF), an extension of random forests that estimates conditional quantiles rather than conditional means. This approach better preserves distributional characteristics compared to standard imputation methods. QRF works by building an ensemble of decision trees on the training data, but unlike standard random forests, it stores all observations in leaf nodes rather than just their means. This enables estimation of any quantile of the conditional distribution at prediction time, allowing us to sample from the full conditional distribution rather than relying on point estimates.

4.1.2 Implementation

We use the `quantile-forest` package, which provides efficient scikit-learn compatible QRF implementation. The specific implementation details are provided in Appendix A.1.

4.1.3 Predictor Variables

The imputation uses seven variables available in both datasets. These include age of the person, a gender indicator, tax unit filing status (whether joint or separate), and the number of dependents in the tax unit. We also use tax unit role indicators specifying whether each person is the head, spouse, or dependent within their tax unit. These predictors capture key determinants of tax variables while being reliably measured in both datasets. The limited set of predictors ensures common support between the datasets while capturing the primary sources of variation in tax outcomes.

4.1.4 Imputed Variables

We impute 72 tax-related variables spanning six categories: employment and business income (6 variables), retirement and Social Security (4 variables), investment income (6 variables), deductions (12 variables), tax credits and adjustments (20 variables), and other income and special items (24 variables). The complete list of imputed variables is provided in Appendix Table A1. These variables cover the major components needed for tax simulation while maintaining reasonable imputation quality given the available predictors.

4.1.5 Additional Imputations

Beyond the 72 PUF tax variables, we impute additional variables from three other data sources. From the Survey of Income and Program Participation (SIPP), we impute tip income using employment income, age, and household composition as predictors. The Survey of Consumer Finances (SCF) provides data for imputing auto loan balances, interest payments, and net worth components. For SCF matching, we use their reference person definition to ensure proper household alignment. From the American Community Survey (ACS), we impute property taxes for homeowners, rent values for specific tenure types, and additional housing characteristics. These supplementary imputations fill gaps in the CPS that are important for comprehensive policy analysis but not available in tax data.

4.1.6 Sampling Process

Rather than using point estimates, we sample from the conditional distribution to preserve realistic variation in the imputed variables. We first train QRF models on each source dataset, then for each CPS record, we estimate the conditional distribution of each variable given the predictors. We sample from this distribution using a random quantile drawn from a uniform distribution. To ensure consistency across related variables, we use the same random quantile for variables that should be correlated, such as different types of capital gains. This approach preserves realistic correlations between imputed variables while maintaining the marginal distributions observed in the source data.

4.2 Stage 2: Reweighting

4.2.1 Problem Formulation

The reweighting stage adjusts household weights to ensure the enhanced dataset matches known administrative totals. Given a loss matrix M R$^{nm} containingnhouseholds'contributionstomtargets,$ $transformedweightswtominimizemeansquaredrelativeerror.TheobjectivefunctionisL(w) =$ $(1/m)\_i((exp(w)^T M\_i{-}t\_i)/t\_i)^2, whereexp(w)representstheexponentiatedweightsappliedtohouseholds.T$

4.2.2 Optimization

We use PyTorch for gradient-based optimization with the Adam optimizer. The implementation uses log-transformed weights to ensure positivity constraints are satisfied throughout the optimization process. The detailed optimization code is provided in Appendix A.2.

4.2.3 Dropout Regularization

To prevent overfitting to calibration targets, we apply dropout during optimization. We randomly mask 5% of weights each iteration and replace masked weights with the mean of unmasked weights. This percentage was selected through sensitivity analysis on validation performance, testing rates from 0% to 10%. The dropout helps ensure that no single household receives excessive weight in matching targets, improving the stability of policy simulations.

4.2.4 Calibration Targets

The loss matrix includes over 7,000 targets from six sources. IRS Statistics of Income provides the largest share with over 5,300 targets covering income by AGI bracket and filing status, counts of returns by category, and aggregate income totals. Census data contributes over 200 targets including population by single year of age, state populations, and demographic distributions. Program totals from CBO projections and Treasury EITC statistics add approximately 10 targets. Tax expenditure estimates from JCT cover four major deductions. Healthcare spending patterns stratified by age contribute over 40 targets. The remaining 1,500+ targets come from various sources including state-level program participation and income distributions by geography. The complete list of calibration targets is provided in our online documentation.

4.2.5 Convergence

The optimization typically converges within 3,000 iterations. We run for 5,000 iterations to ensure stability. Convergence is monitored through the loss value trajectory, weight stability across iterations, and target achievement rates. The optimization is considered converged when the relative change in loss falls below 0.001% for 100 consecutive iterations.

4.3 Validation

4.3.1 Cross-Validation

We validate the methodology through three approaches. First, we employ 5-fold cross-validation on calibration targets, holding out subsets of targets to assess out-of-sample performance. Second, we test stability across multiple random seeds to ensure results are not sensitive to initialization. Third, we validate the imputation quality through out-of-sample prediction on held-out records from the source datasets.

4.3.2 Quality Checks

Throughout the enhancement process, we implement several quality checks to ensure data integrity. We verify that all weights remain positive after optimization, as negative weights would violate the interpretation of survey weights as population representations. Weight magnitudes are checked to ensure no single household receives excessive influence on aggregate statistics. We preserve demographic relationships by verifying that household members maintain consistent relationships after reweighting. Finally, we ensure household structures remain intact, with all members of a household receiving the same weight adjustment factor.

4.4 Implementation

The complete implementation is available at: https://github.com/PolicyEngine/policyengine-us-data

Key files:

- `policyengine_us_data/datasets/cps/extended_cps.py` - Imputation stage
- `policyengine_us_data/datasets/cps/enhanced_cps.py` - Reweighting stage
- `policyengine_us_data/utils/loss.py` - Loss matrix construction

The modular design allows researchers to modify or extend individual components while maintaining the overall framework.

5 Results

We validate our enhanced dataset against official statistics and compare its performance to both the original CPS and PUF datasets. Our validation framework covers over 7,000 distinct targets spanning demographic totals, program participation rates, and income components across the distribution.

5.1 Validation Against Administrative Totals

The Enhanced CPS is validated against all 7,000+ administrative targets used in the calibration process. While the dataset is explicitly calibrated to these targets, the large number and diversity of targets makes achieving good fit across all dimensions a significant challenge.

Detailed validation results are available in our interactive dashboard at https://policyengine.github.io/po
us-data/validation.html.

5.2 Target Category Performance

The enhanced dataset is calibrated to various categories of targets:

IRS Income Components: AGI-stratified income targets including employment income, capital gains, partnership and S-corp income, and dividend income across filing statuses and income ranges.

Program Participation: CBO projections for SNAP benefits, Social Security, unemployment compensation, and income tax revenue.

Demographic Targets: Population by single year of age, state populations, and healthcare spending by age group.

Tax Expenditures: JCT estimates for SALT deduction ($21.2B), charitable deduction ($65.3B), mortgage interest ($24.8B), and medical expense deduction ($11.4B).

5.3 Income Distribution

Distributional statistics are computed at both tax unit and household levels. Tax unit metrics allow comparison with the PUF, while household metrics are relevant for many policy applications.

5.3.1 Tax Unit Level Metrics

The Enhanced CPS achieves distributional statistics between those of the CPS and PUF. The imputation of tax variables from the PUF increases measured inequality compared to the baseline CPS.

Key metrics include:

- Gini coefficient
- Top 10% income share
- Top 1% income share

5.3.2 Household Level Metrics

For applications requiring household-level analysis, we also compute metrics over households rather than tax units. The PUF cannot provide household-level statistics as it lacks household structure.

5.4 Poverty Measurement

Poverty metrics require careful interpretation. The interaction between imputed tax variables and poverty measurement is complex, and results may differ from official statistics. Users analyzing poverty should:

- Compare results across different weight specifications
- Consider the impact of tax variable imputation
- Reference official poverty statistics for validation

5.5 Weight Distribution

The weight distribution reflects the enhancement methodology:

- Original CPS weights are relatively uniform
- Enhanced CPS weights show greater variation due to the calibration process
- Some records receive zero weight as the optimization selects representative combinations

5.6 Policy Application: Top Tax Rate Reform

To demonstrate practical applications, we analyze a reform raising the top marginal tax rate from 37% to 39.6%. This reform affects high-income taxpayers and tests the dataset's ability to model policies targeting the top of the income distribution.

The Enhanced CPS incorporates detailed income data from the PUF, enabling analysis of high-income tax reforms that would be difficult with the CPS alone.

5.7 Validation Dashboard

Our comprehensive validation dashboard provides:

- Performance metrics for all 7,000+ targets
- Comparison across datasets
- Filtering by target category and source
- Regular updates with each data release

Visit https://policyengine.github.io/policyengine-us-data/validation.html to explore detailed results.

5.8 Summary

The Enhanced CPS successfully combines the demographic detail of the CPS with tax precision approaching that of administrative data. While specific validation metrics depend on the target category, the dataset provides a suitable foundation for analyzing both tax and transfer policies.

Users should:

- Consult the validation dashboard for metrics relevant to their analysis
- Consider the dataset's strengths and limitations for their specific use case
- Compare results with official statistics where available

6 Discussion

This section examines the strengths, limitations, and potential applications of the Enhanced CPS dataset, along with directions for future development.

6.1 Strengths

6.1.1 Comprehensive Coverage

The Enhanced CPS uniquely combines:

- Demographic detail from the CPS including state identifiers
- Tax precision from IRS administrative data
- Calibration to contemporary official statistics
- Open-source availability for research use

This combination enables analyses that would be difficult or impossible with existing public datasets alone.

6.1.2 Methodological Contributions

The use of Quantile Regression Forests for imputation represents an advance over traditional matching methods:

- Preserves full conditional distributions
- Captures non-linear relationships
- Maintains realistic variable correlations
- Allows uncertainty quantification

The large-scale calibration to 7,000+ targets ensures consistency with administrative benchmarks across multiple dimensions simultaneously.

6.1.3 Practical Advantages

For policy analysis, the dataset offers:

- State-level geographic detail enabling subnational analysis
- Household structure for distributional studies
- Tax detail for revenue estimation
- Program participation for benefit analysis
- Recent data calibrated to current totals

6.2 Limitations

6.2.1 Temporal Inconsistency

The most significant limitation is the temporal gap between data sources:

- 2015 PUF data imputed onto 2024 CPS
- Nine-year gap in underlying populations
- Demographic shifts not fully captured
- Tax law changes since 2015

While dollar amounts are uprated and calibration partially addresses this, fundamental demographic changes may not be reflected.

6.2.2 Imputation Assumptions

The QRF imputation assumes:

- Relationships between demographics and tax variables remain stable
- Seven predictors sufficiently capture variation
- PUF represents the tax-filing population well
- Missing data patterns are ignorable

These assumptions may not hold perfectly, particularly for subpopulations underrepresented in the PUF.

6.2.3 Calibration Trade-offs

With 7,000+ targets, perfect fit to all benchmarks is impossible. The optimization must balance:

- Competing objectives across target types
- Relative importance of different statistics
- Stability of resulting weights
- Preservation of household relationships

Users should consult validation metrics for targets most relevant to their analysis.

6.3 Applications

6.3.1 Tax Policy Analysis

The dataset excels at analyzing federal tax reforms:

- Accurate income distribution at high incomes
- Detailed deduction and credit information
- State identifiers for SALT analysis
- Household structure for family-based policies

6.3.2 State and Local Analysis

Unlike the PUF, the Enhanced CPS enables state-level studies:

- State income tax modeling
- Geographic variation in federal policies
- State-specific program interactions
- Regional economic impacts

6.3.3 Integrated Policy Analysis

The combination of tax and transfer data supports:

- Universal basic income proposals
- Earned income tax credit expansions
- Childcare and family benefit reforms
- Healthcare subsidy design

6.3.4 Microsimulation Model Development

As the foundation for PolicyEngine US, the dataset demonstrates how enhanced microdata improves model capabilities:

- More accurate baseline distributions
- Better behavioral response modeling
- Improved validation against benchmarks
- Enhanced credibility of results

6.4 Comparison with Alternatives

6.4.1 Versus Synthetic Data

Unlike fully synthetic datasets, our approach:

- Preserves actual survey responses where possible
- Imputes only missing tax variables
- Maintains household relationships
- Provides transparent methodology

6.4.2 Versus Administrative Data

While not replacing restricted administrative data, the Enhanced CPS offers:

- Public availability
- Household structure
- Geographic detail
- Integration with survey content
- No access restrictions

6.4.3 Versus Other Matching Approaches

Compared to traditional statistical matching:

- QRF better preserves distributions
- Large-scale calibration ensures consistency
- Open-source implementation enables replication
- Modular design allows improvements

6.5 Future Directions

6.5.1 Methodological Enhancements

Potential improvements include:

- Incorporating additional predictors for imputation
- Using more recent administrative data when available
- Developing time-series consistency methods
- Adding uncertainty quantification

6.5.2 Additional Data Integration

Future versions could incorporate:

- State tax return data
- Program administrative records
- Consumer expenditure information
- Health insurance claims data

6.5.3 Model Development

The framework could be extended to:

- Dynamic microsimulation over time
- Behavioral response estimation
- Geographic mobility modeling
- Life-cycle analysis

6.5.4 International Applications

The methodology could be adapted for other countries:

- Similar data availability challenges
- Need for tax-benefit integration
- Open-source implementation
- Cross-national comparisons

6.6 Conclusion for Researchers

The Enhanced CPS provides a valuable resource for policy analysis, though users should:

- Understand the limitations, particularly temporal inconsistency
- Validate results against external benchmarks
- Consider sensitivity to methodological choices
- Contribute improvements to the open-source project

The dataset represents a pragmatic solution to data limitations, enabling analyses that advance our understanding of tax and transfer policy impacts while we await improved data access.

7 Conclusion

We have presented a methodology for creating enhanced microsimulation datasets that combine the strengths of survey and administrative data sources. The Enhanced CPS dataset demonstrates that careful application of modern statistical methods can substantially improve the data available for policy analysis.

7.1 Summary of Contributions

Our work makes several key contributions:

Methodological Innovation: The use of Quantile Regression Forests for imputation preserves distributional characteristics while maintaining computational efficiency. The large-scale calibration to 7,000+ targets pushes the boundaries of survey data enhancement.

Practical Tools: We provide open-source implementations that enable researchers to apply, modify, and extend these methods. The modular design facilitates experimentation with alternative approaches.

Validated Dataset: The Enhanced CPS itself serves as a public good for the research community, enabling studies that would otherwise require restricted data access.

Reproducible Research: All code, data, and documentation are publicly available, supporting reproducibility and collaborative improvement.

7.2 Key Findings

The validation results demonstrate that combining survey and administrative data through principled statistical methods can achieve:

- Improved income distribution representation
- Better alignment with program participation totals
- Maintained demographic and geographic detail

- Suitable accuracy for policy simulation

While no dataset perfectly represents the full population, the Enhanced CPS provides a pragmatic balance of accuracy, detail, and accessibility.

7.3 Implications for Policy Analysis

The availability of enhanced microdata has immediate implications:

Improved Revenue Estimates: More accurate representation of high incomes enables better analysis of progressive tax reforms.

Integrated Analysis: Researchers can analyze tax and transfer policies jointly rather than in isolation.

State-Level Studies: Geographic identifiers enable subnational policy analysis not possible with administrative tax data alone.

Distributional Analysis: Household structure allows examination of policy impacts across family types and income levels.

7.4 Broader Implications

Beyond the specific dataset, this work demonstrates:

Value of Data Integration: Combining multiple data sources can overcome individual limitations.

Open Science Benefits: Making methods and data publicly available accelerates research progress.

Practical Solutions: Perfect data may never exist, but pragmatic enhancements can substantially improve analysis capabilities.

Collaborative Potential: Open-source approaches enable community contributions and continuous improvement.

7.5 Limitations and Future Work

We acknowledge important limitations:

- Temporal inconsistency between data sources
- Imputation model assumptions
- Calibration trade-offs
- Validation challenges

Future work should address these through:

- More recent administrative data
- Enhanced imputation methods
- Additional validation exercises
- Uncertainty quantification

7.6 Call to Action

We encourage researchers to:

Use the Dataset: Apply the Enhanced CPS to policy questions where combined demographic and tax detail adds value.

Validate Results: Compare findings with other data sources and contribute validation results.

Improve Methods: The open-source nature invites methodological enhancements.

Share Experiences: Document use cases, limitations discovered, and suggested improvements.

7.7 Final Thoughts

The Enhanced CPS represents one approach to a fundamental challenge in microsimulation: the need for comprehensive, accurate microdata. While not perfect, it demonstrates that substantial improvements are possible through careful methodology and open collaboration.

As data availability evolves and methods advance, we hope this work contributes to a future where policy analysis rests on increasingly solid empirical foundations. The ultimate goal remains better informed policy decisions that improve social welfare.

The enhanced dataset, complete documentation, and all source code are available at https://github.com/PolicyEngine/policyengine-us-data.