

# Enhancing the CPS with Administrative Tax Data

## Machine Learning Meets Microsimulation

Nikhil Woodruff & Max Ghenis

PolicyEngine

National Tax Association Annual Conference  
November 14, 2024



- Current Population Survey March Supplement (CPS)
  - Rich demographics and program participation
  - Underreports income, especially at top
  - Limited tax information

- Current Population Survey March Supplement (CPS)
  - Rich demographics and program participation
  - Underreports income, especially at top
  - Limited tax information
- IRS Public Use File (PUF)
  - Accurate administrative tax data
  - No demographics or state ID
  - Restricted access

- More Accurate Policy Analysis
  - Taxes and benefits jointly affect household incentives
  - Need accurate data on both to model behavior
  - Many researchers lack access to key datasets

- More Accurate Policy Analysis
  - Taxes and benefits jointly affect household incentives
  - Need accurate data on both to model behavior
  - Many researchers lack access to key datasets
- Better Understanding of Economic Reality
  - CPS misses top incomes
  - PUF can't show demographic patterns
  - Both limit inequality measurement

- Machine learning to combine strengths of CPS and PUF:
  - Learn tax patterns from PUF
  - Preserve CPS demographics and program data
  - Optimize weights to match 570 administrative targets

- Machine learning to combine strengths of CPS and PUF:
  - Learn tax patterns from PUF
  - Preserve CPS demographics and program data
  - Optimize weights to match 570 administrative targets
- Result: First open dataset with:
  - Administrative-quality tax data
  - Rich demographics and program participation
  - Transparent, reproducible methodology

# Two-Stage Approach: ML Imputation + Weight Optimization

## PolicyEngine CPS-PUF integration and reweighting

How PolicyEngine applies its survey-enhance software to build a novel microdata set, structured as the Current Population Survey and using signals from the IRS Public Use File for improved accuracy

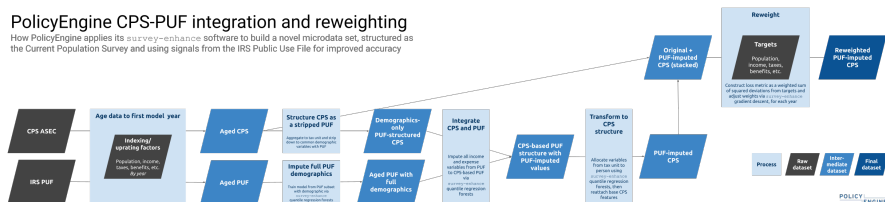
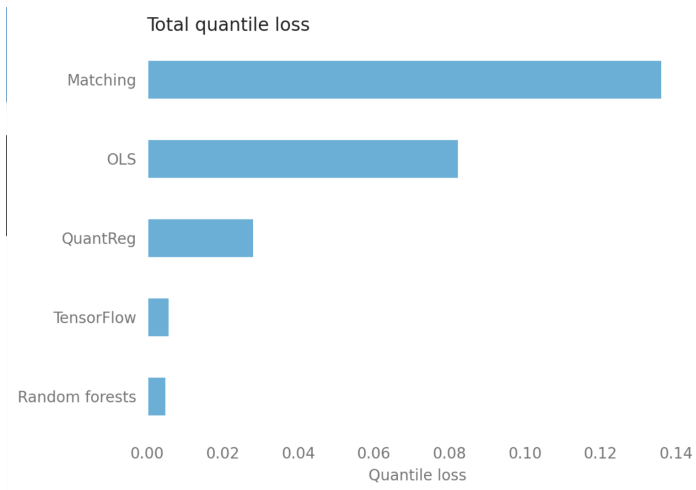


Figure: Overview of dataset enhancement process



- Standard approach: statistical matching or regression
- We use Quantile Regression Forests (QRF) for:
  - Imputing tax variables from PUF
  - Predicting housing costs from ACS
  - Estimating prior year earnings
- Benefits of QRF approach:
  - Captures full conditional distributions
  - Handles non-linear relationships
  - Preserves correlations between variables



**Figure:** Average quantile loss by method, predicting net worth from covariates in SCF

- Standard approach: constrained optimization
- We use dropout-regularized gradient descent
- Optimizes against 570 targets:
  - IRS Statistics of Income by income bins
  - Program participation totals
  - Single-year age population counts
- Mathematics:

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

where  $w$  are weights,  $M$  is characteristics,  $t$  are targets

Table: Examples of calibration targets by source

Source	Example Targets	Count
IRS SOI	AGI by bracket, employment income, capital gains	5,300+
Census	Population by age, state populations	150+
CBO	SNAP benefits, Social Security, income tax	5
JCT	SALT deduction (\$21.2B), charitable (\$65.3B)	4
Healthcare	Medicare Part B premiums by age group	40+

- ECPS is best on qualified dividends and infant population
- PUF better on returns AGI 100-200k
- 567 other targets!

# Validation II: ECPS Outperforms Both Source Datasets

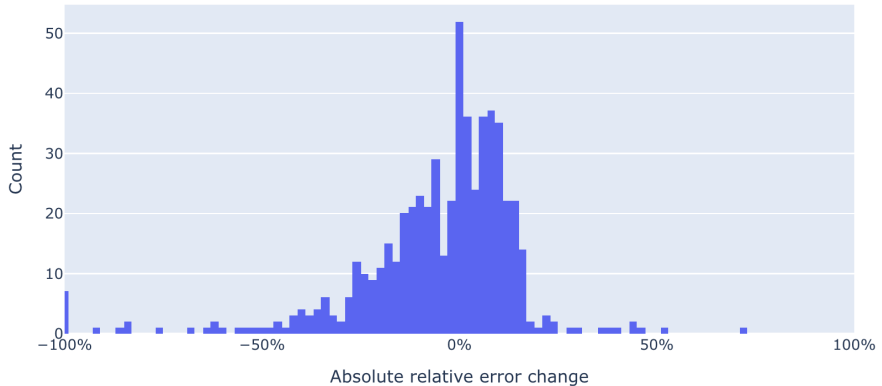


Figure: Error change from ECPS to better of CPS and PUF

- ECPS outperforms CPS on 63% of targets
- ECPS outperforms PUF on 71% of targets

Table: Tax unit-level distributional metrics

Metric	CPS	Enhanced CPS	PUF
Gini coefficient	[TBC]	[TBC]	[TBC]
Top 10% share	[TBC]	[TBC]	[TBC]
Top 1% share	[TBC]	[TBC]	[TBC]

- CPS inequality measures 12-45% lower than PUF
- ECPS inequality within 4% of PUF
- Unlike PUF, ECPS includes nonfilers
- Inequality measured as income after taxes and transfers

- Example: Biden's proposed top rate increase
- Would raise rate from 37% to 39.6% above \$400k

**Table:** Revenue projections from top rate increase (37% to 39.6%)

Dataset	Revenue Impact (\$B)	Affected Tax Units (M)	Avg Tax
CPS	[TBC]	[TBC]	[TBC]
Enhanced CPS	[TBC]	[TBC]	[TBC]
PUF	[TBC]	[TBC]	[TBC]

- Can analyze by demographics, geography, income
- Interactive results at [policyengine.org](https://policyengine.org)

- Direct race/ethnicity analysis without imputation
- Other models use complex methods:
  - CBO: Statistical matching with Census data
  - Tax Policy Center: Multiple copies with reweighting
  - ITEP: Probability assignment based on characteristics
- Our approach:
  - Uses observed demographics from CPS
  - Individual-level rather than tax unit only
  - Enables analysis of intersectional effects
  - Extends to disability, education, etc.



- Full codebase on GitHub
- Automatic validation dashboard
- Python package for programmatic access
- Web interface at [policyengine.org](https://policyengine.org)
- Growing research applications:
  - Academic studies
  - Think tank analysis
  - Government agency use
  - Community contributions

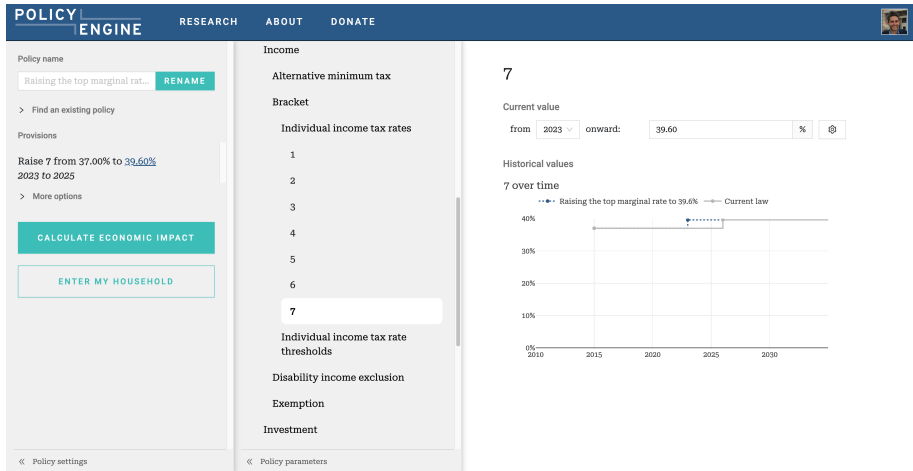


Figure: PolicyEngine's policy editor interface

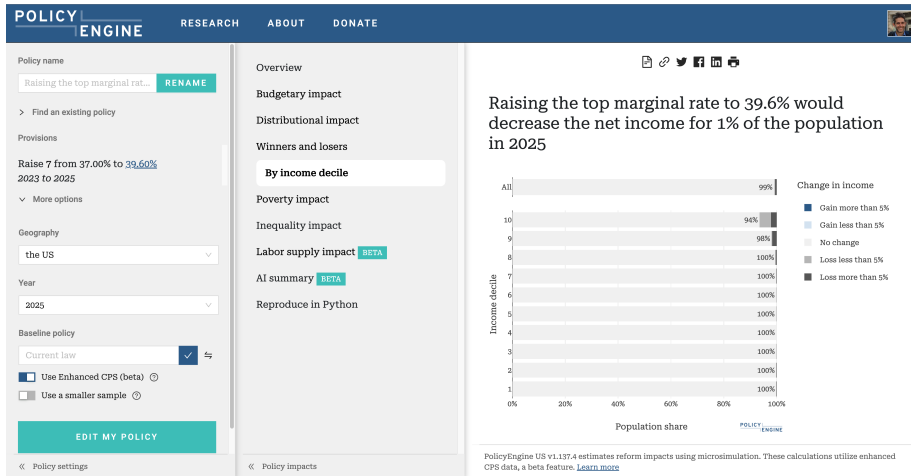


Figure: PolicyEngine's policy impact interface

- Geographic extensions:
  - Congressional district weights
  - State-specific calibration
  - County-level synthetic data
- Prediction-oriented validation:
  - Compare to tax expenditure reports
  - Backtest
  - Benchmark ML architectures
- International applications (UK version live)

- Paper: [github.com/PolicyEngine/policyengine-us-data/paper](https://github.com/PolicyEngine/policyengine-us-data/paper)
- Code: [github.com/PolicyEngine/policyengine-us-data](https://github.com/PolicyEngine/policyengine-us-data)
- Web app: [policyengine.org](https://policyengine.org)
- Contact: [max@policyengine.org](mailto:max@policyengine.org)