

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

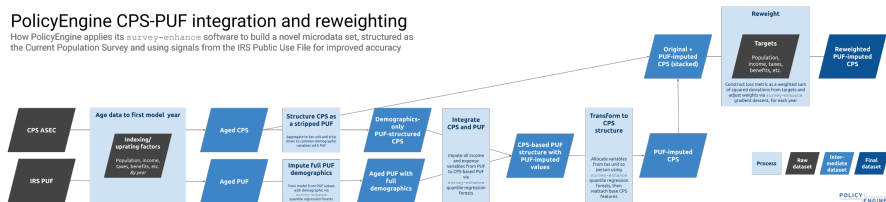


Figure: Overview of dataset enhancement process

- 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

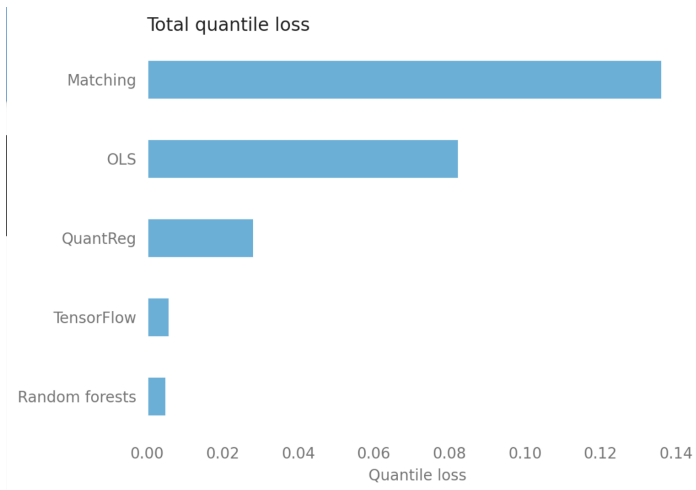


Figure: Average quantile loss by method, predicting net worth from covariates in SCF

- 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

	Admin	CPS	PUF	ECPS
Qual Div	\$314b	\$103b (-67%)	\$263b (-16%)	\$322b (+3%)
Infants	3.6m	2.8m (-23%)	17.0m (+367%)	4.0m (+11%)
AGI 100-200k	24.2m	29.5m (+22%)	24.3m (+0%)	28.3m (+17%)

- 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: 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

- 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: Projected revenue from top rate increase, 2025

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

- Can analyze by demographics, geography, income
- Interactive results at 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
- Growing research applications:
 - Academic studies
 - Think tank analysis
 - Government agency use
 - Community contributions

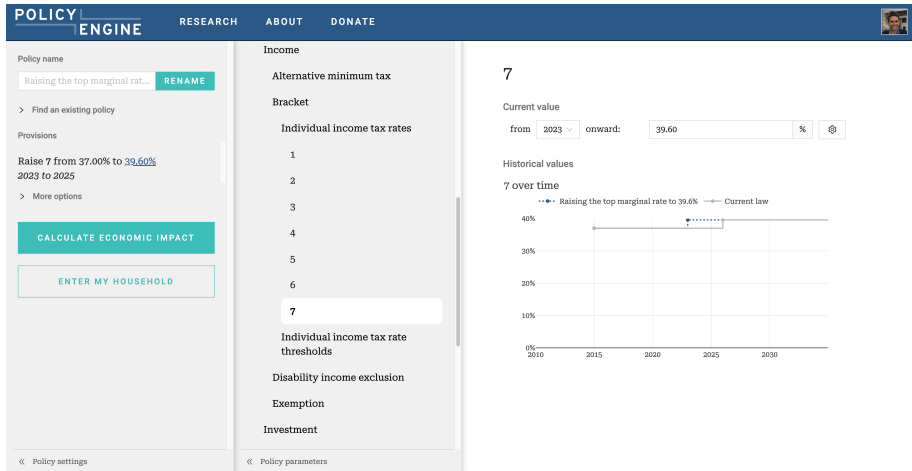


Figure: PolicyEngine's policy editor interface

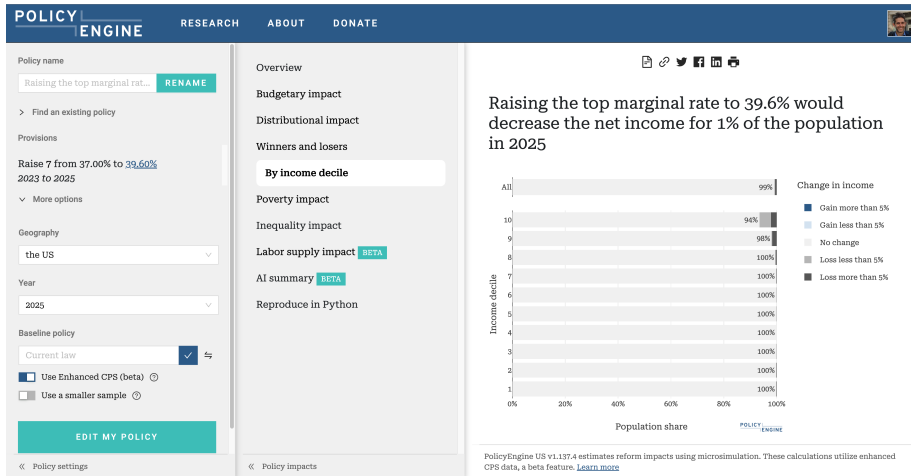


Figure: PolicyEngine's policy impact interface

- 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
- Code: github.com/PolicyEngine/policyengine-us-data
- Web app: policyengine.org
- Contact: max@policyengine.org