Enhancing the CPS with Administrative Tax Data

Machine Learning Meets Microsimulation

Nikhil Woodruff & Max Ghenis

PolicyEngine

National Tax Association Annual Conference
November 14, 2024

The Challenge

- Current Population Survey (CPS)
  - Rich demographics, state ID, household structure
  - Underreports income, especially at top
  - Limited tax information
  - Self-reported program participation
- IRS Public Use File (PUF)
  - Accurate tax records from administrative data
  - No demographics beyond age/sex
  - No state ID
  - Strict confidentiality rules
- Need both for comprehensive policy analysis

- First openly available dataset integrating CPS and PUF
- No confidentiality restrictions
- Preserves demographic detail while matching tax data
- Powers PolicyEngine microsimulation platform
- Enables:
    - Analysis by race, education, disability status
    - Program interactions
    - Transparent, reproducible research

Our Approach

POLICY ENGINE

- Generate synthetic tax variables from PUF using:
    - Quantile regression forests to learn distributions
    - Match to CPS demographic variables

Our Approach

- Generate synthetic tax variables from PUF using:
  - Quantile regression forests to learn distributions
  - Match to CPS demographic variables
- Stack synthetic records with original CPS

Our Approach

- Generate synthetic tax variables from PUF using:
  - Quantile regression forests to learn distributions
  - Match to CPS demographic variables
- Stack synthetic records with original CPS
- Calculate taxes and benefits via microsimulation

Our Approach

- Generate synthetic tax variables from PUF using:
  - Quantile regression forests to learn distributions
  - Match to CPS demographic variables
- Stack synthetic records with original CPS
- Calculate taxes and benefits via microsimulation
- Optimize weights against 570 targets:
  - IRS Statistics of Income income bins
  - Program participation totals
  - Single-year age population counts

Enhanced CPS Pipeline

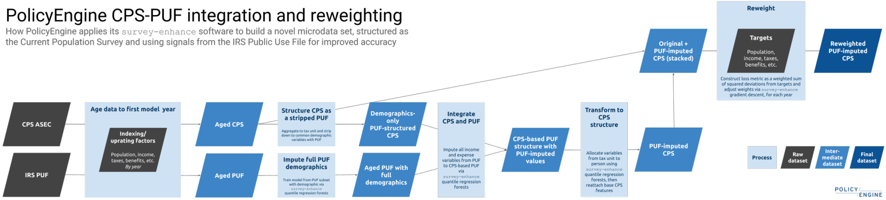


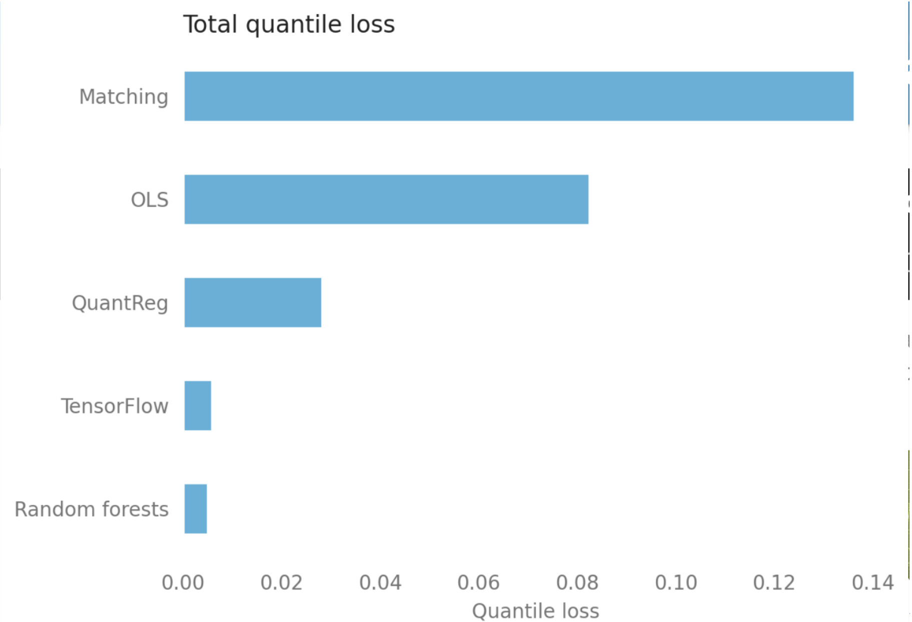
Figure: Overview of dataset enhancement process

Machine Learning Methods

POLICY ENGINE

- 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
  - Outperforms traditional statistical matching

Machine Learning Methods



Weight Optimization

POLICY ENGINE

- Novel dropout-regularized gradient descent
- Optimizes against 570 targets including:
    - IRS Statistics of Income by income bins
    - Program participation totals
    - Single-year age population counts
    - State-level aggregates
- Prevents overfitting to any single target
- Handles sparse subgroups effectively

Weight Optimization Mathematics

POLICY ENGINE

- Minimize relative squared error:

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

where:

  - w are log-transformed household weights
  - M is matrix of household characteristics
  - t are target values from administrative data

- Implementation:
  - PyTorch gradient descent with Adam optimizer
  - 5% dropout rate for regularization
  - 5,000 iterations or convergence

Validation: Example 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: Target Calibration

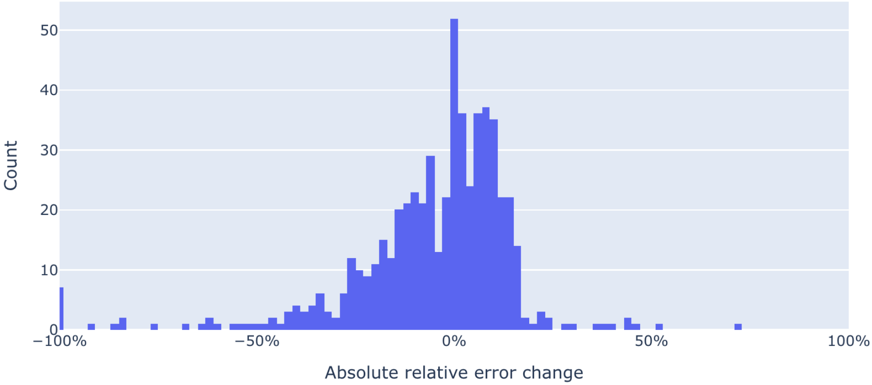


Figure: Error change from ECPS to better of CPS and PUF

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

Validation: Income Distribution

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
- Inequality measured as income after taxes and transfers

Validation: Policy Analysis

- Test case: 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

The PolicyEngine Platform

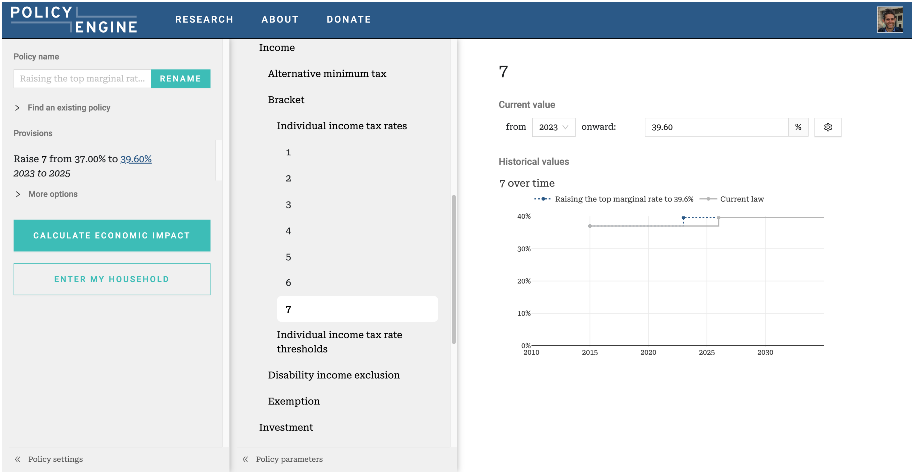


Figure: PolicyEngine's policy editor interface

Interactive Results

POLICY ENGINE

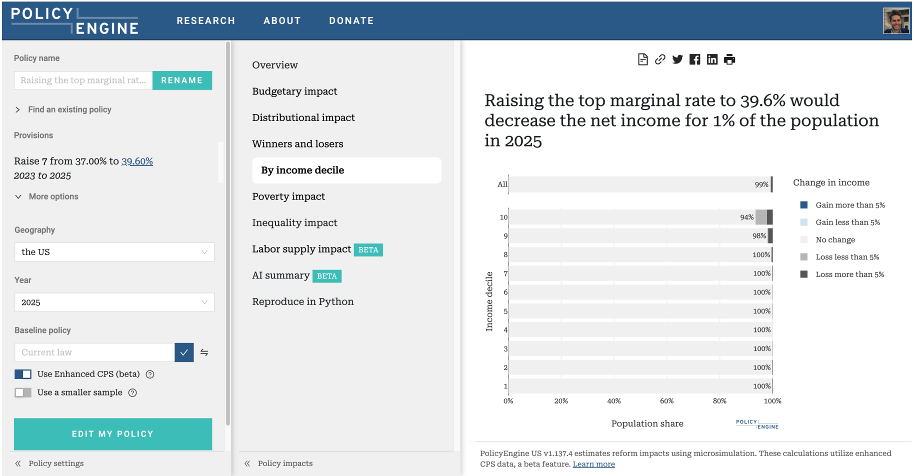


Figure: Example distributional analysis from PolicyEngine

Unique Demographic Analysis

- Direct race/ethnicity analysis without imputation
- Other models must 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.

Open Science

POLICY ENGINE

- 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

Next Steps

POLICY ENGINE

- Geographic extensions:
  - Congressional district weights
  - State-specific calibration
  - County-level synthetic data
- Methodological improvements:
  - Time series validation
  - Uncertainty quantification
  - Alternative ML architectures
- International applications (UK version live)

Thank You

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