

Enhancing the CPS with Administrative Tax Data

Machine Learning Meets Microsimulation

Nikhil Woodruff & Max Ghenis

PolicyEngine

Society of Government Economists

April 4, 2025



- 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

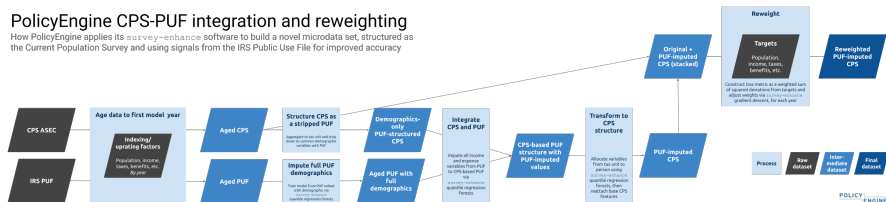


Figure: Overview of dataset enhancement process

- 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

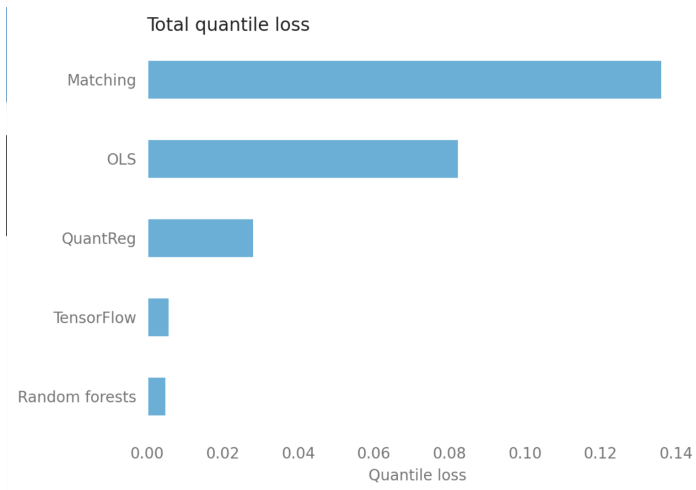


Figure: Average quantile loss by method, predicting net worth from covariates in SCF

- Standard approach: constrained optimization
- We use dropout-regularized gradient descent
- Optimizes against 570 targets:
 - IRS Statistics of Income by income bins
 - Tax expenditure reports
 - 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

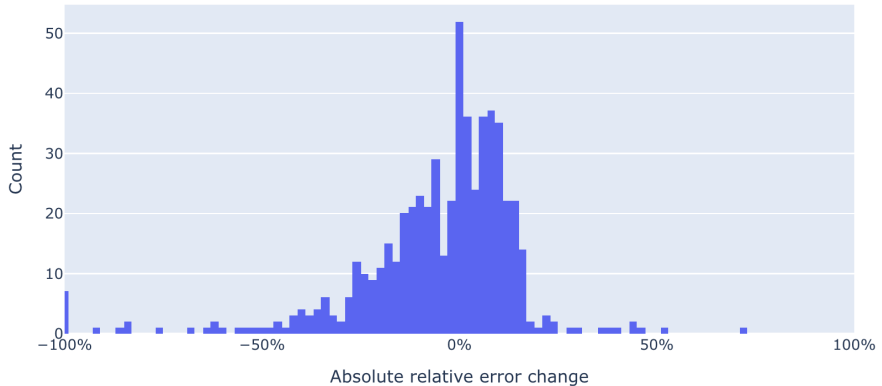


Figure: Error change from ECPS to better of CPS and PUF

- 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

- 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.

- Standard approach: Optimize single weight per household
- For UK local analysis, we optimize a matrix of weights:
 - One weight per household per constituency
 - Allows different households to have different importance in different areas
 - Includes constituency-level targets in gradient descent

$$W = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,C} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ w_{H,1} & w_{H,2} & \cdots & w_{H,C} \end{pmatrix}$$

where $w_{h,c}$ is the weight of household h in constituency c

- Unlike UK local model, US requires different policy rules by state
- Our approach:
 - Perform calibration separately for each of 51 states (including DC)
 - Propagate national targets to individual states
 - Apply L0 penalty to prune household-state rows for computational efficiency
 - Results in 51 separate weight matrices rather than a single large matrix
 - Reassemble into a single matrix for analysis

$$W_s = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,D_s} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,D_s} \\ \vdots & \vdots & \ddots & \vdots \\ w_{H,1} & w_{H,2} & \cdots & w_{H,D_s} \end{pmatrix}$$

where $w_{h,d}$ is the weight of household h in district d for state s

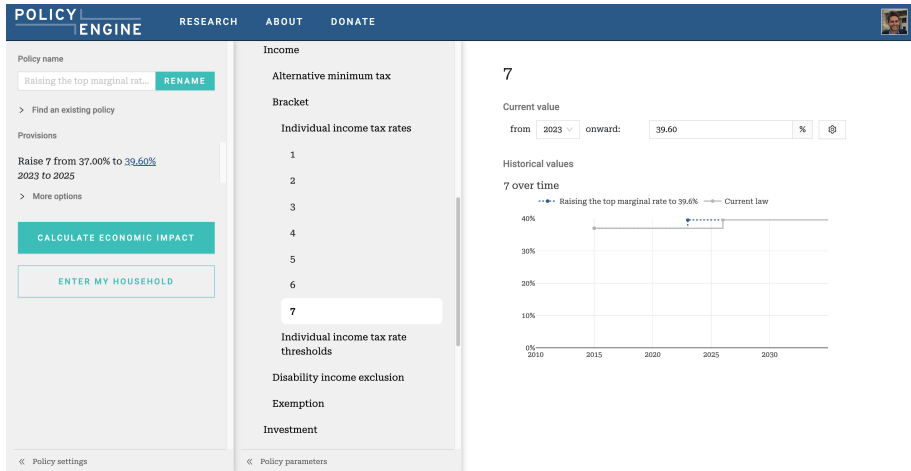


Figure: PolicyEngine's policy editor interface

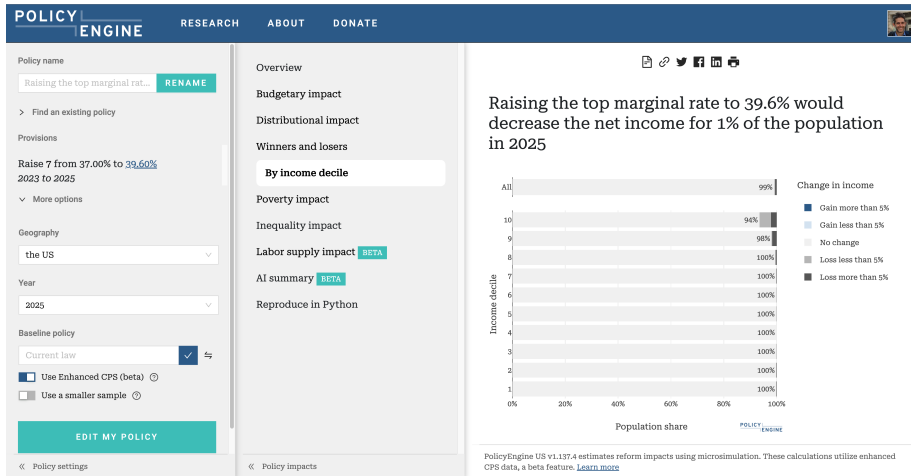


Figure: PolicyEngine's policy impact interface

PolicyEngine: UK Parliamentary Constituency Choropleth

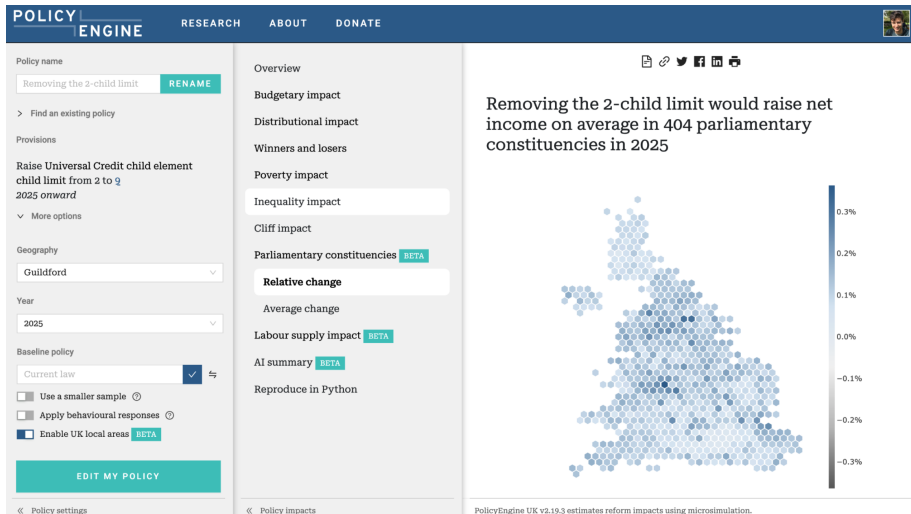


Figure: PolicyEngine UK showing impact by parliamentary constituency

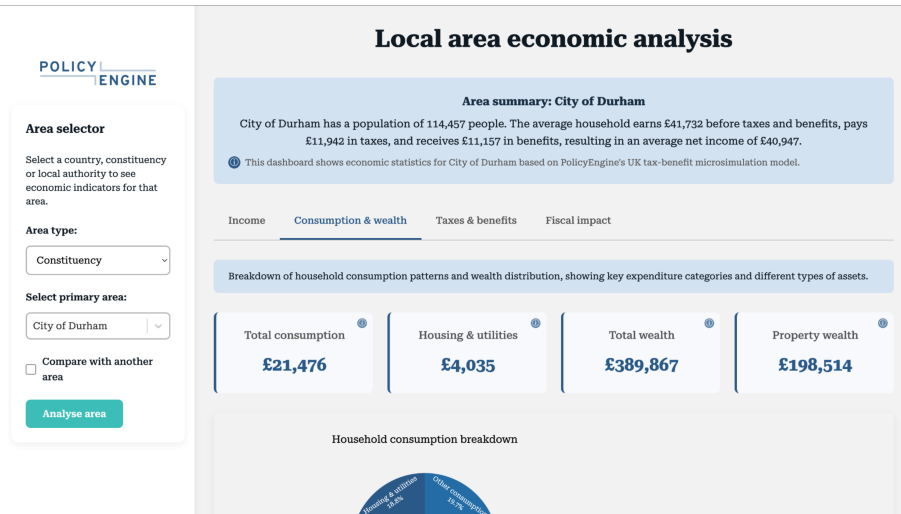


Figure: PolicyEngine UK local area validation dashboard

- 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

- Geographic and data extensions:
 - Calibrate to states and Congressional districts
 - Integrate SCF and CE
- Making contributions more modular:
 - Creating a `microimpute` package (using quantile regression forests)
 - Developing a `microreweight` package (using gradient descent)
 - These packages can be used across different microdata files
 - Planning separate papers benchmarking these new methods against traditional approaches
- Prediction-oriented validation:
 - Backtest
 - Benchmark ML architectures

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