Enhancing the CPS with Administrative Tax Data

Machine Learning Meets Microsimulation

Nikhil Woodruff & Max Ghenis

PolicyEngine

National Tax Association Annual Conference
November 14, 2024

POLICY ENGINE

The Challenge

POLICY ENGINE

- Current Population Survey (CPS)
  - Rich demographics, state ID, household structure
  - But underreports income, especially at top
  - Limited tax information

The Challenge

POLICY ENGINE

- Current Population Survey (CPS)
  - Rich demographics, state ID, household structure
  - But underreports income, especially at top
  - Limited tax information
- IRS Public Use File (PUF)
  - Accurate tax records from administrative data
  - But no demographics beyond age/sex
  - No state ID
  - No household structure

Our Solution

POLICY ENGINE

- Use quantile regression forests to:
  - Learn full distributions of tax variables from PUF
  - Generate synthetic records matching CPS demographics

Our Solution

POLICY ENGINE

- Use quantile regression forests to:
  - Learn full distributions of tax variables from PUF
  - Generate synthetic records matching CPS demographics
- Stack synthetic records with original CPS

Our Solution

POLICY ENGINE

- Use quantile regression forests to:
  - Learn full distributions of tax variables from PUF
  - Generate synthetic records matching CPS demographics
- Stack synthetic records with original CPS
- Use gradient descent to optimize weights against:
  - IRS Statistics of Income
  - Census population projections
  - CBO program totals
  - State administrative data

Validation: Inequality Measures

POLICY ENGINE

- Tax unit income inequality in enhanced CPS matches IRS data

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

Table: Income inequality measures, 2023

Validation: Inequality Measures

POLICY ENGINE

- Tax unit income inequality in enhanced CPS matches IRS data

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

Table: Income inequality measures, 2023

- Also matches state-level income inequality
- And program participation by demographics

Validation: Top Rate Reform

POLICY ENGINE

- Biden's 2025 budget: raise top rate to 39.6% above $400k

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

Table: Projected revenue, 2025

Key Innovation: Demographics "For Free"

POLICY ENGINE

- Other models must impute race/ethnicity:
  - CBO: Statistical matching with survey data
  - TPC: Multiple record copies with reweighting
  - ITEP: Probabilistic assignment by characteristics

Key Innovation: Demographics "For Free"

POLICY ENGINE

- Other models must impute race/ethnicity:
  - CBO: Statistical matching with survey data
  - TPC: Multiple record copies with reweighting
  - ITEP: Probabilistic assignment by characteristics
- Our approach:
  - Race/ethnicity direct from CPS
  - Individual-level rather than tax unit only
  - Same applies to disability status, education, etc.

Open Access

POLICY ENGINE

- All code open source
  - github.com/PolicyEngine/policyengine-us-data
  - Reproducible enhancement pipeline
  - Automatic validation dashboard

Open Access

POLICY ENGINE

- All code open source
  - github.com/PolicyEngine/policyengine-us-data
  - Reproducible enhancement pipeline
  - Automatic validation dashboard
- Use through Python package

```
from policyengine_us import Microsimulation
sim = Microsimulation(dataset="enhanced_cps_2024")
```

Open Access

POLICY ENGINE

- All code open source
  - github.com/PolicyEngine/policyengine-us-data
  - Reproducible enhancement pipeline
  - Automatic validation dashboard
- Use through Python package

```
from policyengine_us import Microsimulation
sim = Microsimulation(dataset="enhanced_cps_2024")
```

- Or web interface at policyengine.org
  - Model tax reforms
  - Analyze household impacts
  - Distributional analysis by demographics

Next Steps

POLICY ENGINE

- Enhancing UK Family Resources Survey
  - Add tax returns from Survey of Personal Incomes
  - Integrate wealth from Wealth & Assets Survey
  - Consumption from Living Costs & Food Survey

Next Steps

POLICY ENGINE

- Enhancing UK Family Resources Survey
  - Add tax returns from Survey of Personal Incomes
  - Integrate wealth from Wealth & Assets Survey
  - Consumption from Living Costs & Food Survey
- Expanding US validation
  - State-level tax statistics
  - Program participation by demographics
  - Historical policy changes

Thank You

POLICY ENGINE

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