Synthetic Individual Income Tax Data

Promises and Challenges

Claire Bowen[1]
Victoria Bryant[2]
Leonard Burman[1]
Surachai Khitatrakun[1]
Robert McClelland[1]
Livia Mucciolo[1]
Madeline Pickens[1]
Aaron Williams[1]

NTA Spring Symposium May 12, 2022

[1] Urban Institute
[2] IRS, Statistics of Income

Disclaimer

Results presented here are preliminary and incomplete. Please do not share with others or cite without authors' permission.

A paper with complete results will be available in early June. Email lburman@urban.org to request a copy.

The findings and conclusions are those of the authors and do not reflect the positions or policies of Internal Revenue Service, the Urban Institute, or its funders.

Taxpayer Privacy and Confidentiality

Any publicly released tax data must protect the confidentiality of individual taxpayers.

Aggregated Tabulations

- Rule of 3
- Rule of 10
- Dominance Rule
- Associated Suppression
- Disclosure by subtraction
- Cross-cell disclosure
- Complimentary disclosure

Public Use File

- Subsampling
  - Reweighting
- Aggregation
- Top Coding
- Blurring
  - Multivariate
  - Univariate
  - Rebalancing
- Random Noise
  - Rounding
- Suppression
- Tax variable calculations

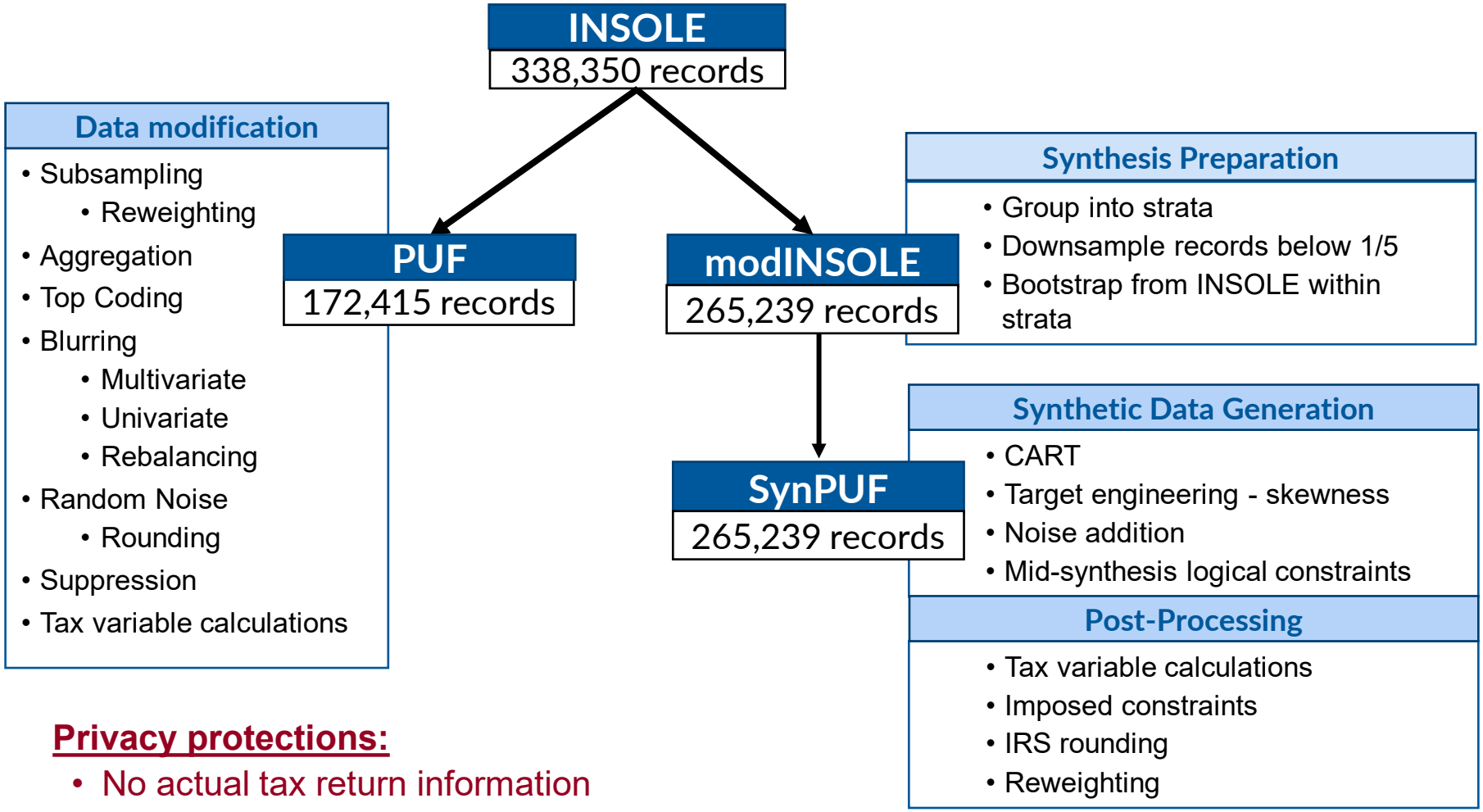
As the scope of information on individuals that is publicly accessible expands, so too must SOI's protection techniques.

Synthetic file objectives

Produce fully synthetic tax data with the same record layouts as IRS administrative data that:

1. Protect the confidentiality of tax return information
2. May be used for statistically valid analysis for certain research purposes
3. May be used as “training data” to develop programs to run on confidential data.

Data generation process



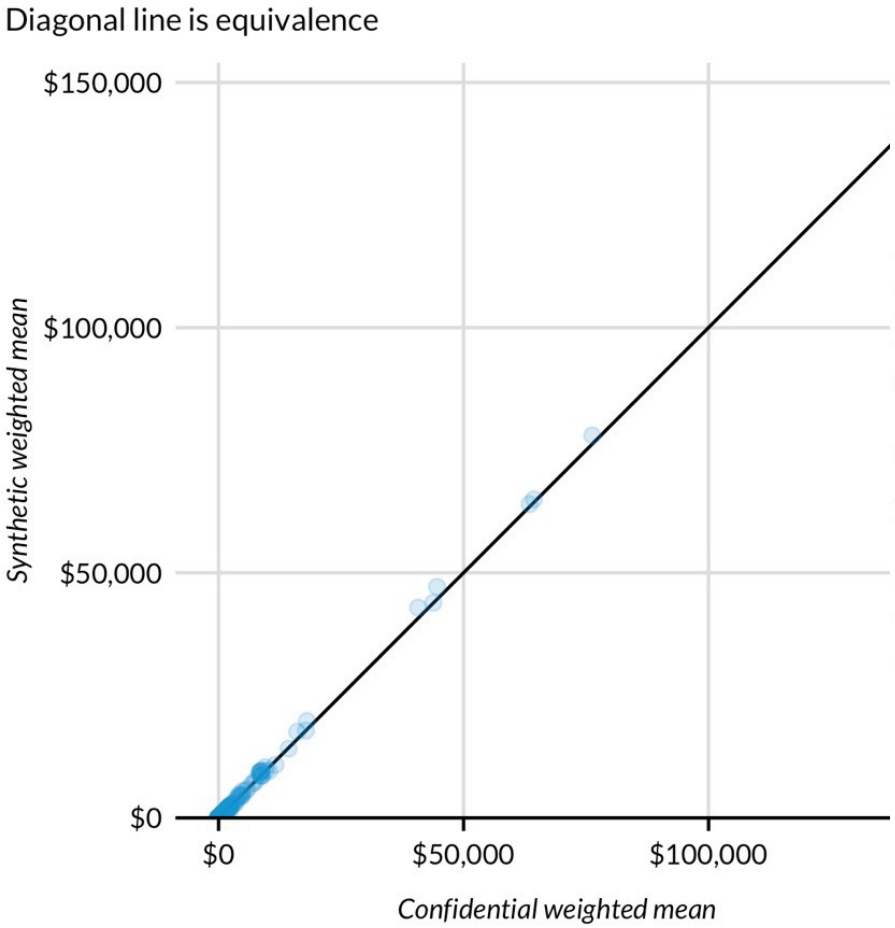
Privacy protections:

- No actual tax return information
- No more than 20% of records are included for any one stratum
- Draw values from a smoothed and unbounded distribution

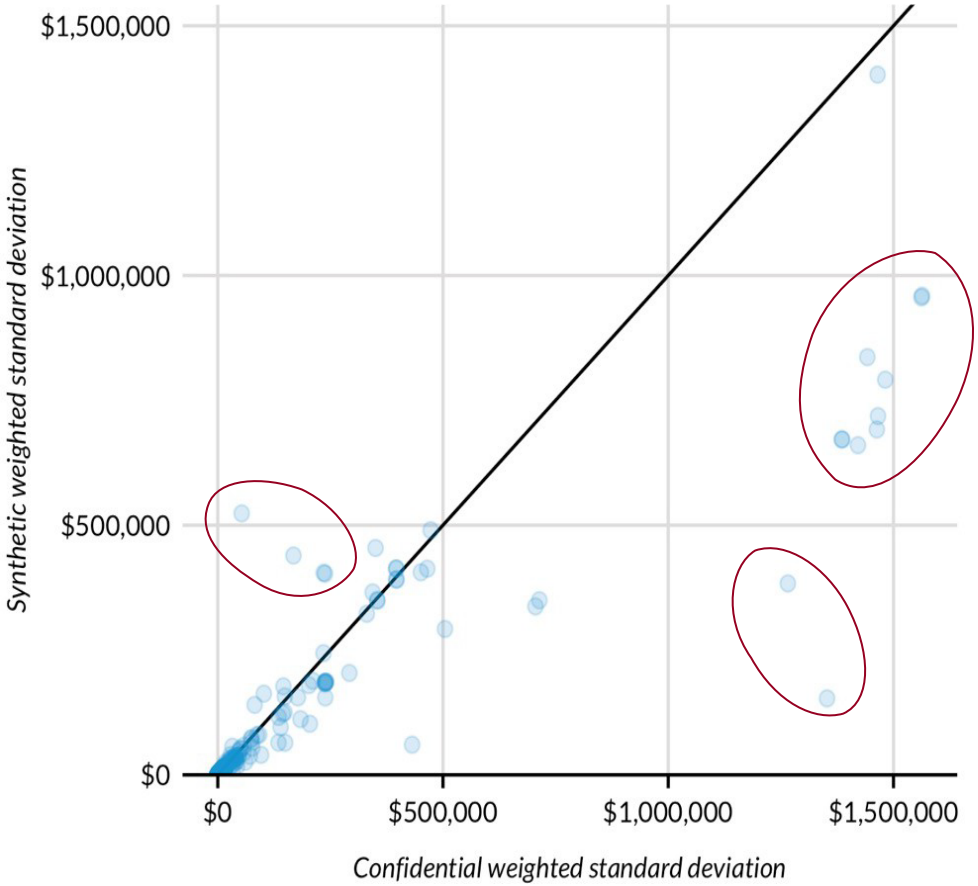
Measures of Quality

Variable properties modINSOLE vs. INSOLE

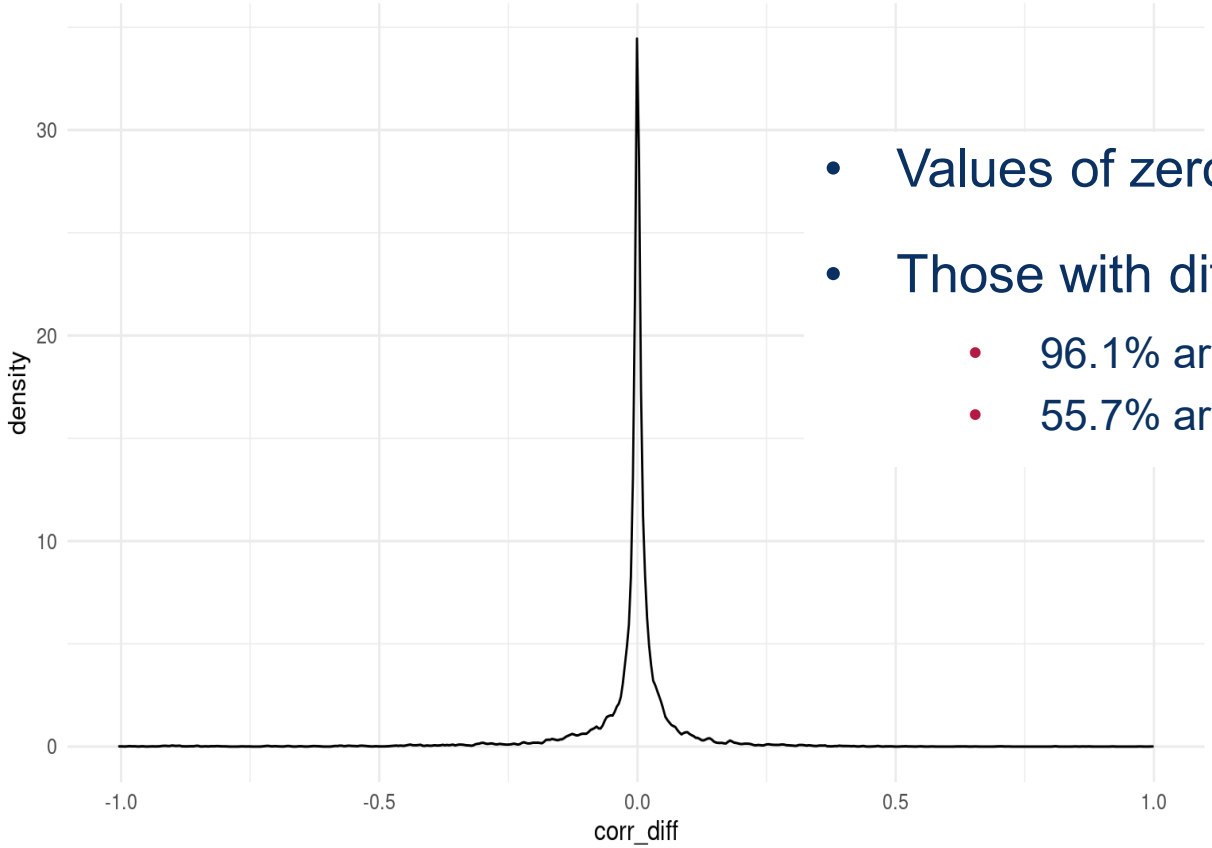
Means



Standard Deviations



Density of pairwise correlation differences

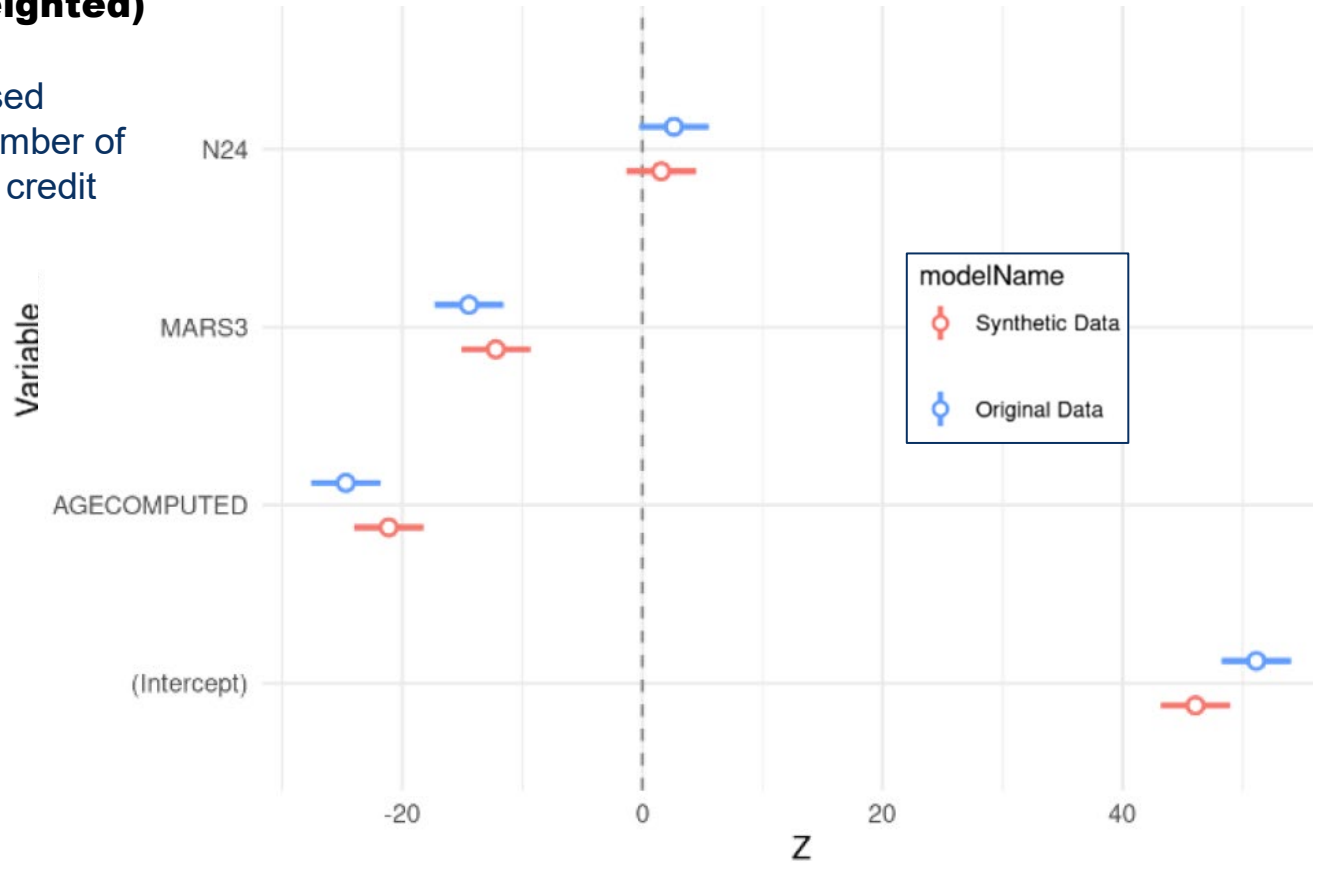


- Values of zero indicate no difference
- Those with differences:
  - 96.1% are less than 0.01
  - 55.7% are less than 0.001

Some multivariate relationships are replicated in the synthetic data

Regression Fit Test (weighted)

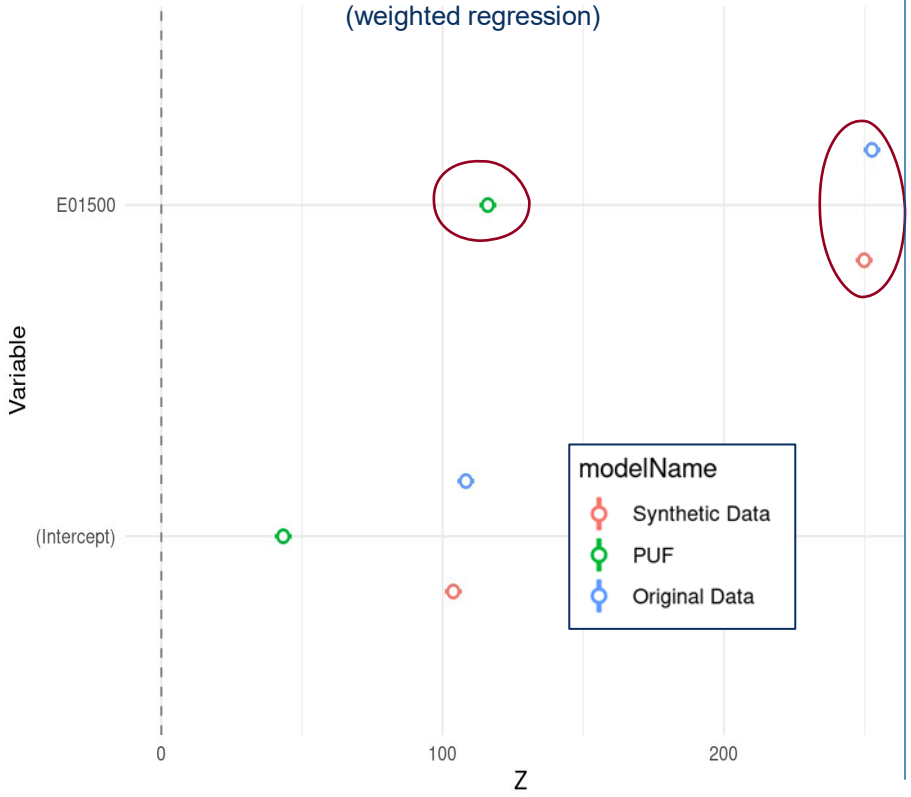
Salaries and Wages regressed against age, filing status, number of children eligible for child tax credit



Multivariate relationships, cont.

Sometimes
Synthetic outperforms PUF

Pensions & annuities in AGI as share of total pensions & annuities received
(weighted regression)



Sometimes
Synthetic trails PUF

Qualified dividends as a share of total dividends
(weighted regression)

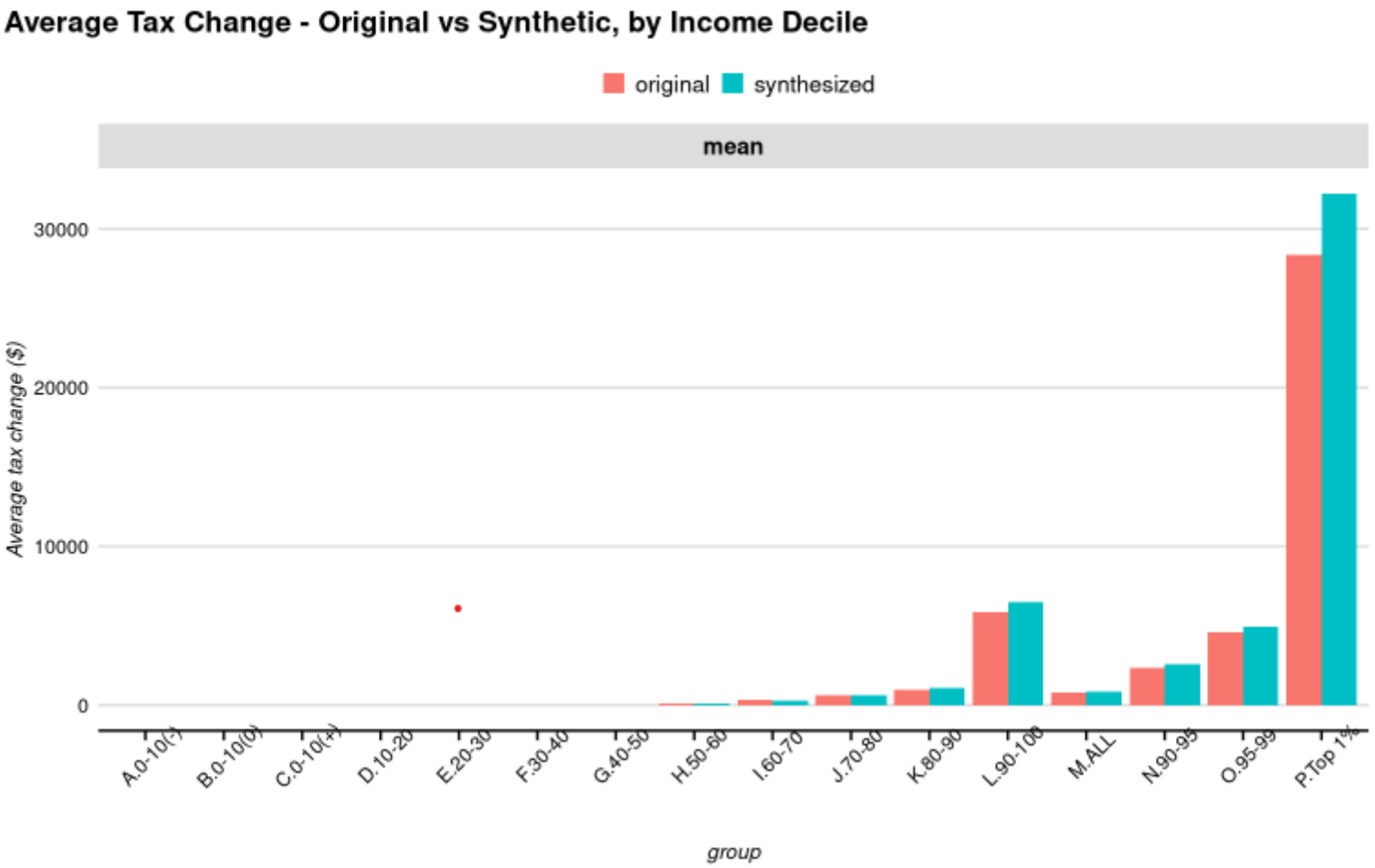
Simulated policy changes

Microsimulation of Tax Model Estimates of Possible Policy Changes in 2012

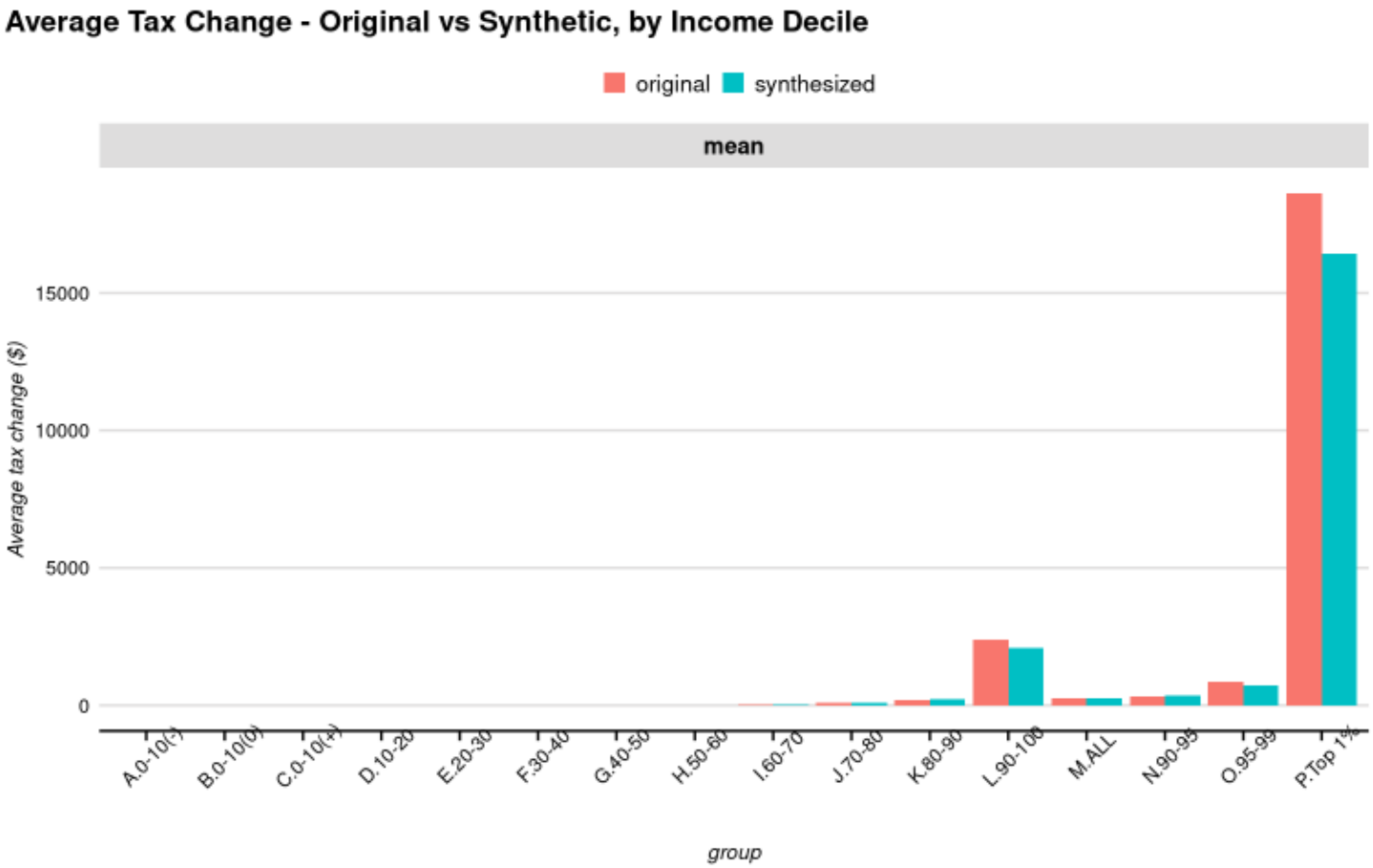
- Returning to 2000 law (i.e. ending all the Bush tax cuts, which had been extended through 2012)
- Implement 2013 law in 2012 (rollback of high-income Bush tax cuts)

Note: following figures based on synthesized PUF rather than modINSOLE

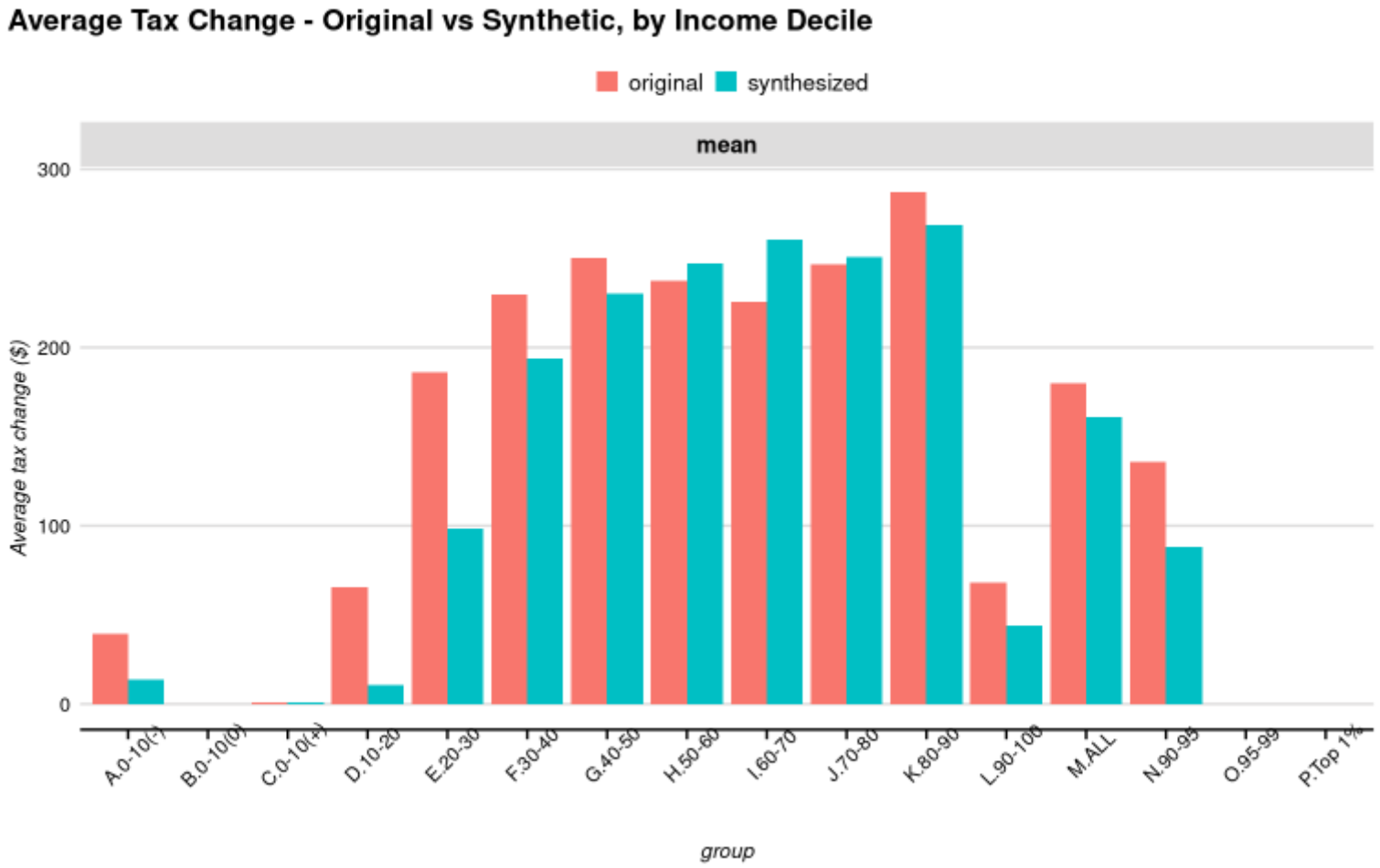
Rates and brackets to 2000 law levels



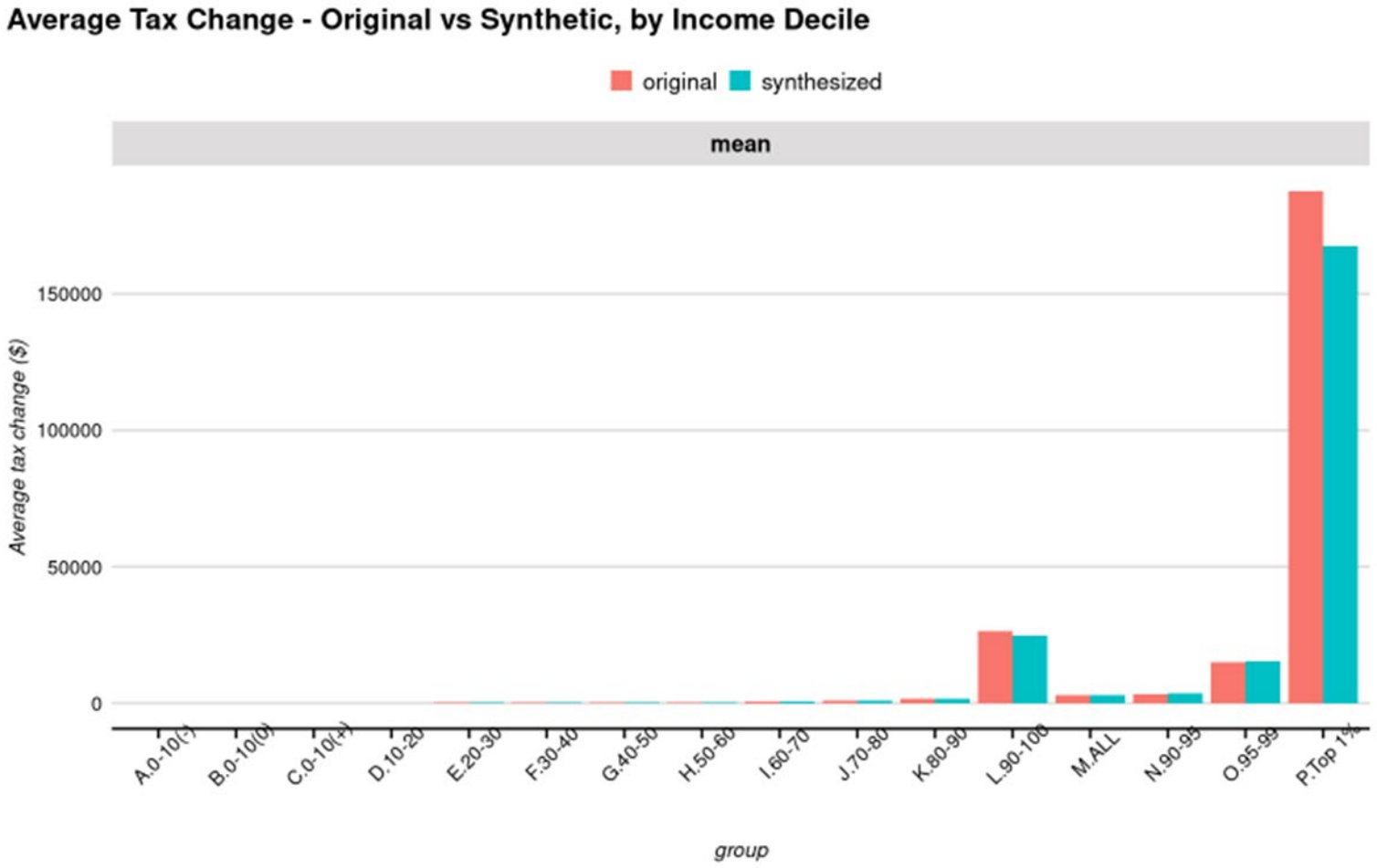
Capital gains rates and brackets to 2000 law levels



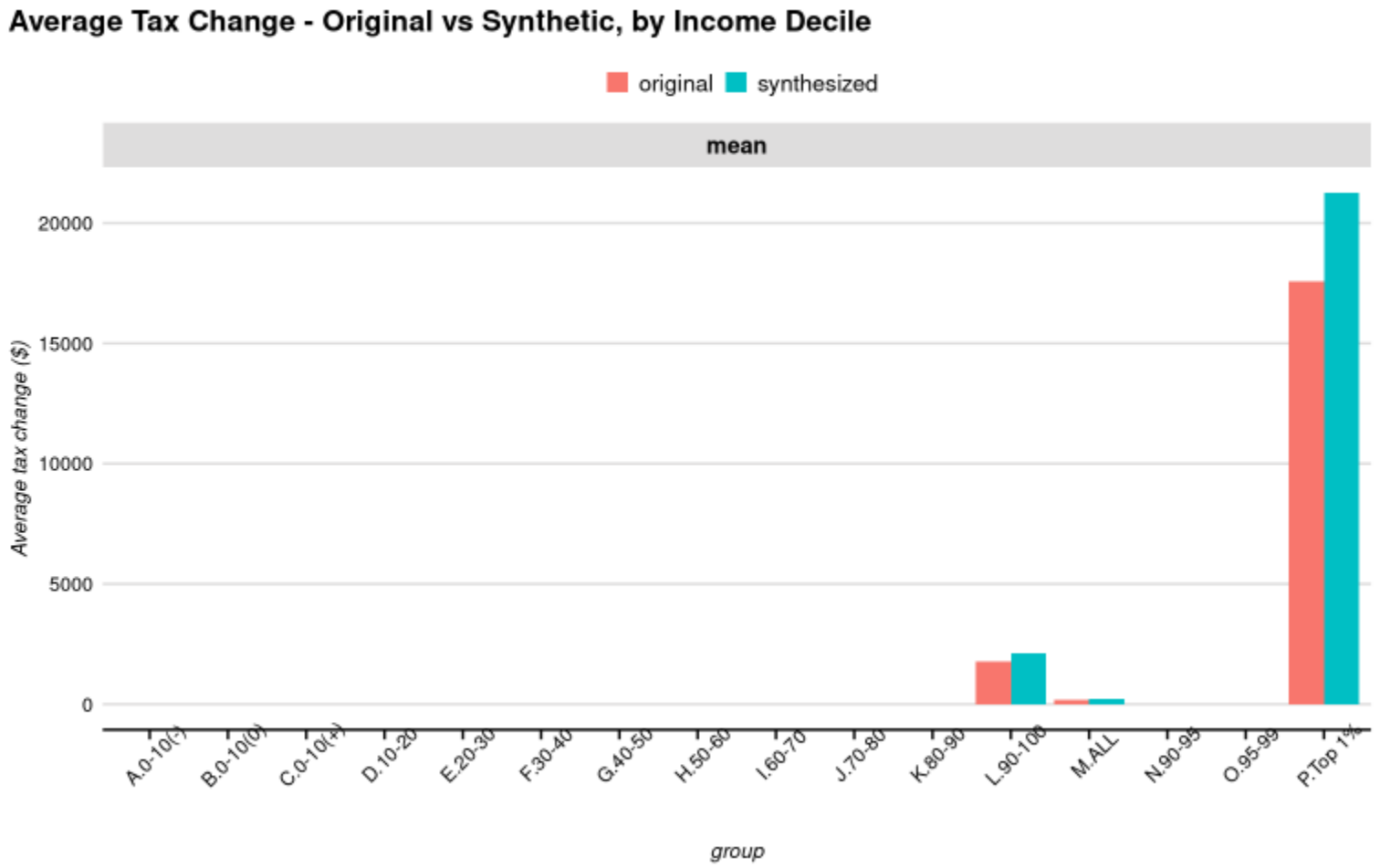
Restore CTC amount and refundability rules to 2000 levels



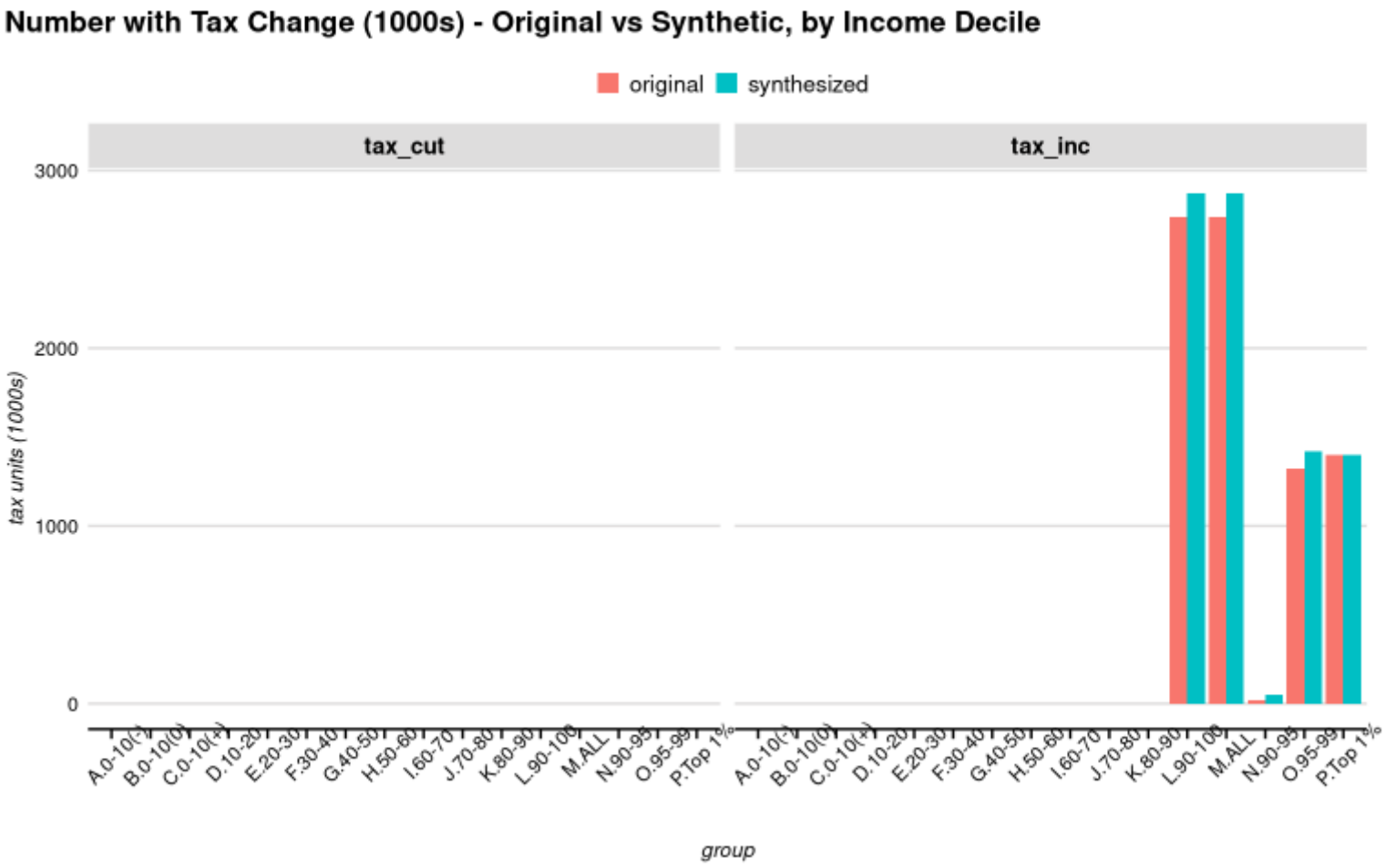
Restore all 2000 law parameters in 2012



Roll back high-end Bush tax cuts (2013 law) in 2012



2013 law – number with tax change, by income



Accomplishments and Future Work

- Synthetic Supplemental PUF of nonfiler* information
- Alpha version of synthetic PUF of individual income tax returns:
  - Separately synthesize positive and negative values for variables that can take negative value
  - Reweight SynPUF to match published SOI aggregates by income and filing status
  - Develop and apply more sophisticated machine learning methods
  - Optimize synthesis so it produces accurate microsimulation analyses for a range of policies
  - Apply this method to later years
- Prototype validation server focusing on user-friendly interface, making users comfortable with a new way of doing research

* Nonfilers are defined as individuals who did not file a federal tax return, had no obligation to file, and were not claimed as a dependent in 2012 but had income reported to the IRS on at least one information return.

Caveats

- Synthesis and validation server technologies are still in relatively early stages of development.
    - We don’t yet know how to develop a large, complex synthetic dataset or validation server that can handle complex statistical queries in a formally private framework (e.g. differential privacy)
    - Relaxations to differential privacy can handle arbitrarily complex statistical queries, but we don’t know how to measure and monitor a “privacy budget” to measure cumulative privacy loss in those models
- Our research team and others are working on these problems