# A new synthetic data set for tax policy analysis

## The Policy Simulation Library DC meeting

The American Enterprise Institute
Washington, DC, November 26, 2019

Don Boyd, Co-Director
State and Local Government Finance Project, Center for Policy Research
Consultant to the AEI Open Source Policy Center
Download at *github donboyd5*: github.com/donboyd5/slides

Based on work done jointly with Max Ghenis, Consultant to the AEI Open Source Policy Center, and Dan Feenberg, National Bureau of Economic Research

# Motivation

- PSL TaxData prepares & enhances 2 microdata files for Tax-Calculator:
  - PUF-based version enhances IRS Public Use File
  - CPS version enhances Current Population Survey
- PUF version represents taxpaying universe well for many purposes but costs ~$10k, requires legal agreement.
- CPS version is free-unrestricted & includes non-taxable benefits, but does not represent high-income taxpayers.
- CPS-PUF differences can be significant (e.g., 20% tax liability).

# Potential benefits of free-unrestricted tax data

- Much wider use
- Desktop Tax-Calculator analysis by:
  - Students, professors in public policy & economics
  - Data journalists
  - Nonprofit fiscal analysis groups
  - State fiscal analysts/policymakers concerned about federal tax impacts
- Ability to focus TaxData resources on a single data approach
- Spinoff state-specific synthetic data files, an OSPC-incubated project
  - Analysis of state income taxes
  - Analysis of federal tax impact on a state - especially important if SALT arises as an issue again

# Synthetic data

- Build a model of the true data, use it to predict synthetic observations. No “real” data.
- Goal: Preserve characteristics of true data - means, variances, correlations, patterns of missingness, while not requiring suppression.
- Potential uses:
  - Practice on data structured like restricted-access confidential data.
  - Get “pretty good” results prior to using confidential “gold standard” data in a safe (restricted) environment.
  - Holy Grail: Results highly similar to gold-standard data. (CAUTION.)
- Several examples: Synthetic versions of Longitudinal Business Database, SIPP, Scottish Longitudinal Study

# Our goals for synthetic tax data

1. Satisfy IRS/SOI disclosure review
2. Make data available for free, without legal agreements
3. Usable as input to tax-calculator models
4. Useful for some tax policy analyses
5. Can identify good uses and dangerous uses

We currently fully satisfy 1-3 and partially satisfy 4-5.

It will get better from here.

NOTE: Our project differs from and complements the Tax Policy Center synthetic data project. Sooner but less extensive. See appendix slide.

# Steps in full project

1. Create synthetic data file
2. Evaluate file quality
3. Ensure that it meets IRS SOI non-disclosure requirement

   NOTE: the file is not produced or endorsed by IRS/SOI in ANY way. It simply passes their disclosure review.

3. Construct weights for the file
4. Enhance the file via TaxData
5. Test, test, test, test
6. Use cautiously for selected purposes
7. Improve.

# Constructing synthetic files

# Synthesis - sequential regression

$Y_1,\ldots,Y_n$: variables to be synthesized (agi, wages, etc.)

**X**: vector of predictors that will not be synthesized, aka "seeds" (can be null)

1. *Estimate* model for each variable, in sequence, using the true data. RHS is **X** and any already-modeled variables in the "visit sequence":

$$Y_1 = f_1(\boldsymbol{X})$$
$$Y_2 = f_2(\boldsymbol{X}, Y_1)$$
$$\ldots$$
$$Y_n = f_n(\boldsymbol{X}, Y_1, \ldots, Y_{n-1})$$

2. *Predict* synthetic values for each variable. RHS is **X** and any already-synthesized values in the "visit sequence":

$$\hat{Y}_1 = f_1(\boldsymbol{X})$$
$$\hat{Y}_2 = f_2(\boldsymbol{X}, \hat{Y}_1)$$
$$\ldots$$
$$\hat{Y}_n = f_n(\boldsymbol{X}, \hat{Y}_1, \ldots, \hat{Y}_{n-1})$$

# Sequential synthesis approach (simplified)

**Seed**
Retained as-is in the synthesis

**Calculated seed**
Used for synthesis, later discarded

**Synthesized variables**
Generated from the conditional distribution

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $65,000 | $1,000 |
| 2 | Married | $40,000 | $40,000 | $0 |
| 3 | Married | $200,000 | $130,000 | $40,000 |

# Sequential synthesis approach (simplified)

**Step 1**

Copy the seed variables over

*Optional: sample with replacement*

TRUE

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $65,000 | $1,000 |
| 2 | Married | $40,000 | $40,000 | $0 |
| 3 | Married | $200,000 | $130,000 | $40,000 |

SYNTH

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | | |
| 2 | Married | $40,000 | | |
| 3 | Married | $200,000 | | |

# Sequential synthesis approach (simplified)

**Step 2**

Determine the conditional distribution of the first synthesized variable

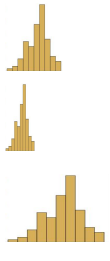
TRUE

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $65,000 | $1,000 |
| 2 | Married | $40,000 | $40,000 | $0 |
| 3 | Married | $200,000 | $130,000 | $40,000 |

SYNTH

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | | |
| 2 | Married | $40,000 | | |
| 3 | Married | $200,000 | | |

Predicted distribution of wages conditional on MARS and AGI



Wages

# Sequential synthesis approach (simplified)

**Step 3**

Impute the first synthesized variable by selecting randomly from the predicted conditional distribution

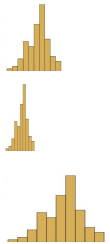
TRUE

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $65,000 | $1,000 |
| 2 | Married | $40,000 | $40,000 | $0 |
| 3 | Married | $200,000 | $130,000 | $40,000 |

SYNTH

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $62,400 | |
| 2 | Married | $40,000 | $40,000 | |
| 3 | Married | $200,000 | $155,100 | |

Predicted distribution of wages conditional on MARS and AGI



Wages

# Sequential synthesis approach (simplified)

**Step 4**

Do steps 2-3 for subsequent variables, conditional also on previously synthesized values

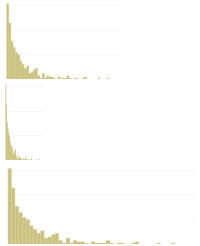
TRUE

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $65,000 | $1,000 |
| 2 | Married | $40,000 | $40,000 | $0 |
| 3 | Married | $200,000 | $130,000 | $40,000 |

SYNTH

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $62,400 | $0 |
| 2 | Married | $40,000 | $40,000 | $0 |
| 3 | Married | $200,000 | $155,100 | $43,900 |

Predicted distribution of dividends conditional on MARS, AGI, and wages



Dividends

# Sequential synthesis approach (simplified)

**Step 5**

Drop the calculated seeds

*Recalculate after synthesizing all variables*

TRUE

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | $70,000 | $65,000 | $1,000 |
| 2 | Married | $40,000 | $40,000 | $0 |
| 3 | Married | $200,000 | $130,000 | $40,000 |

SYNTH

| ID | MARS | AGI | Wages | Dividends |
|---|---|---|---|---|
| 1 | Single | | $62,400 | $0 |
| 2 | Married | | $40,000 | $0 |
| 3 | Married | | $155,100 | $43,900 |

# We use random forests to predict

- Has been shown to perform well
- Works well without a lot of tuning or labor
- Has outperformed (slightly) our CART (classification and regression tree) efforts, especially in “sparse” areas of the data
- We use the synthimpute Python package (written by Max Ghenis)
- 200 trees
- 14.5 hours on an Intel Core i7

# Other approaches merit consideration

- Our methods are used widely and work well with large data files
- Econometric approaches, quantile regression, and related methods can be useful, often in two stages (e.g., TPC, Census):
  - Predict whether a variable is nonzero
  - Predict its value or distribution
- Generative Adversarial Networks (GANs): deep learning methods in which a data-generating neural net communicates with a data-discriminating neural net; potential value re: differential privacy
- Methods require choices about hyperparameters, functional form
- These are on our radar screen and are longer-term possibilities

# Disclosure review

# Disclosure review

- Synthesized records are not sampled data; but it is possible to produce by chance a synthesized record that exactly matches a PUF record.
- More likely for a trivially simple record that occurs multiple times in the data set (e.g., all income and deduction variables are zero), than for a complex record that is unique.

  Rule: No synthesized record may match any unique PUF record, exactly, on every synthesized variable.

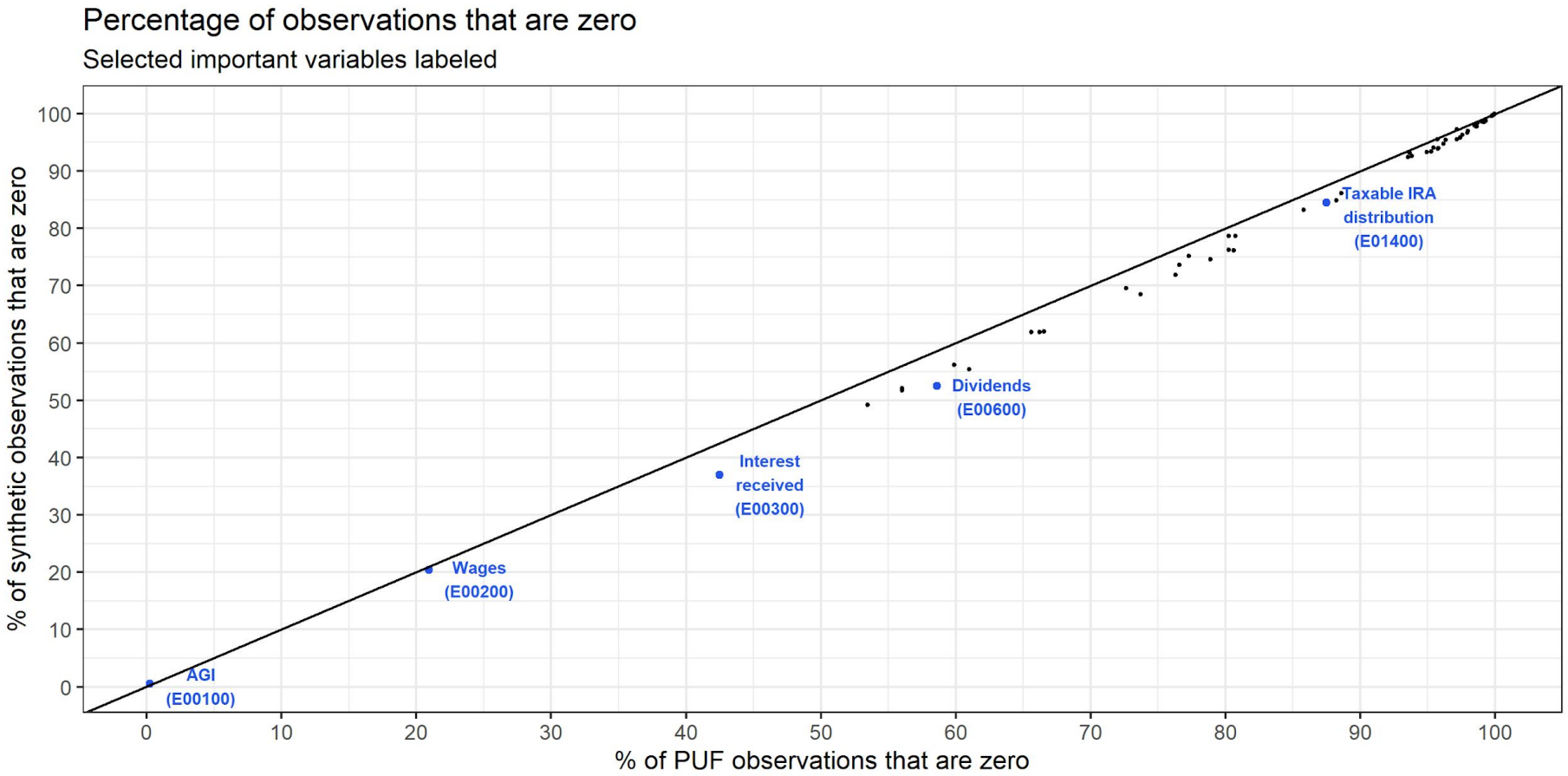
- We ensured that our data meet this rule, and had this certified.

# Evaluating the synthetic file (before weighting)

# Examining synthesis results (unweighted)

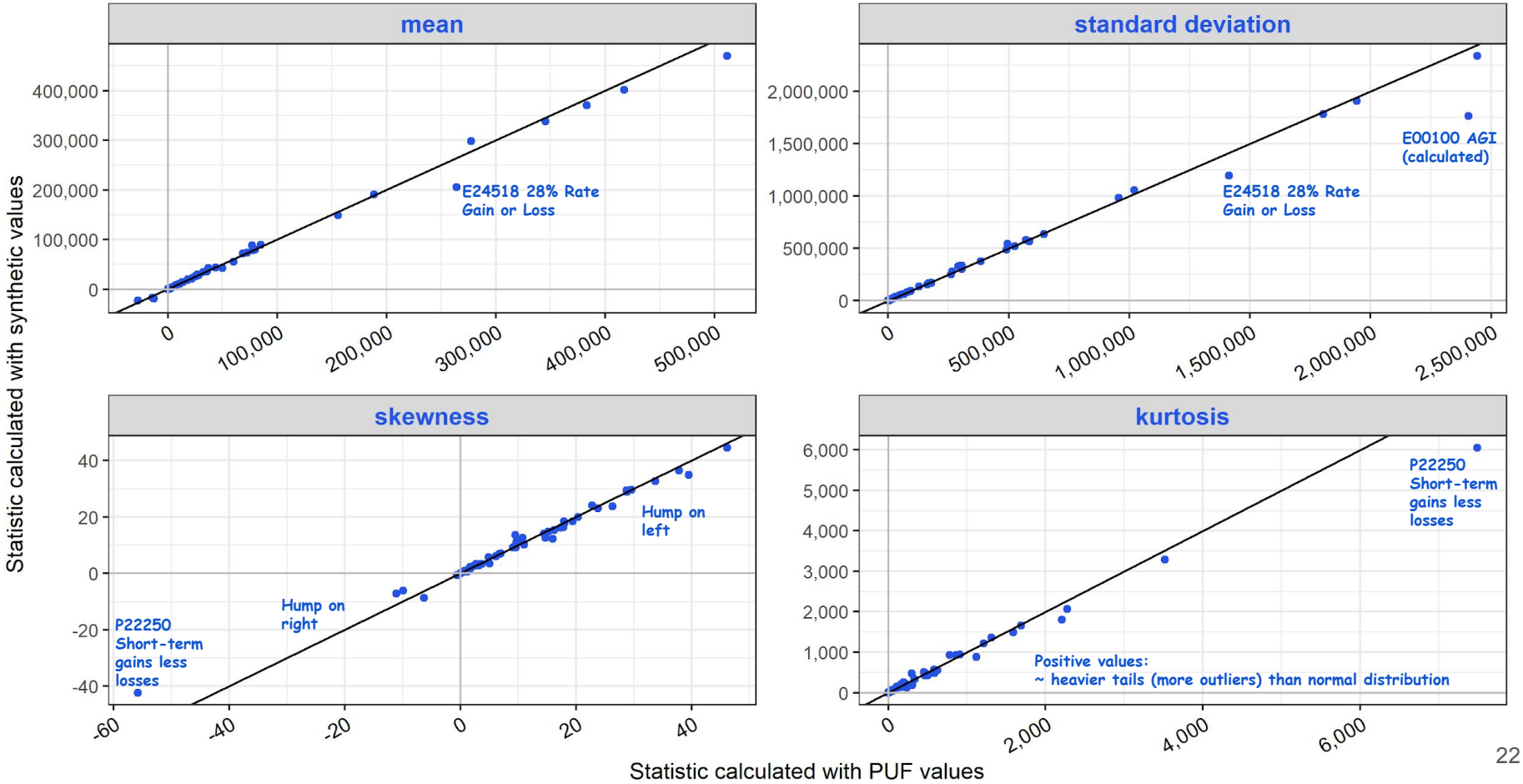
- Individual variables
  - Many zero-values, highly skewed distributions
  - Examine percentage of values that are zero
  - Statistics that describe the variable and its distribution
  - Extent to which PUF and synthetic distributions overlap
- Relationships among variables
  - Correlations
  - Plots of continuous vs. categorical (e.g., by marital status); not shown

# We (slightly) under-synthesize zero-values



# Statistics calculated using nonzero values, with 45-degree line

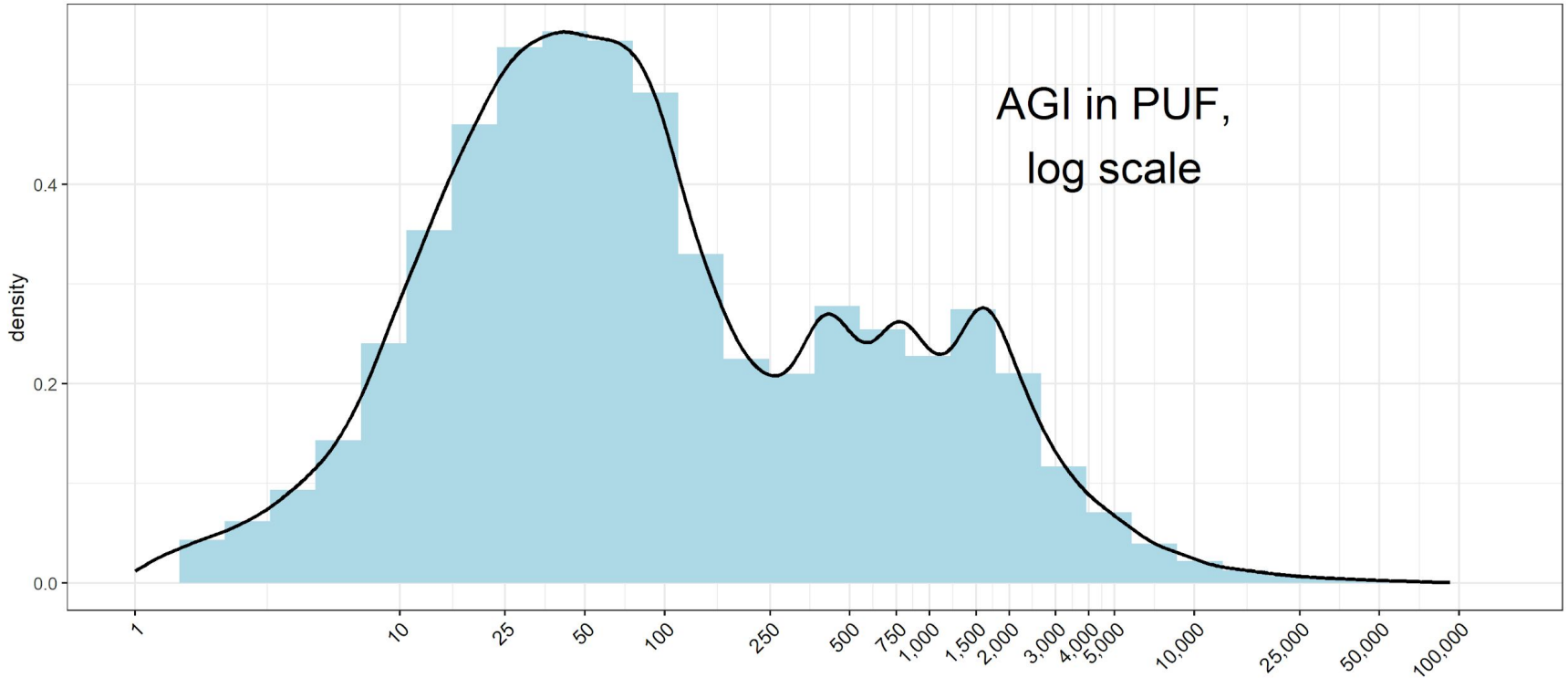
Each point gives the PUF and synthetic-file statistic for a continuous variable



# Kernel density plot is like a smoothed histogram. Useful for comparing distributions.

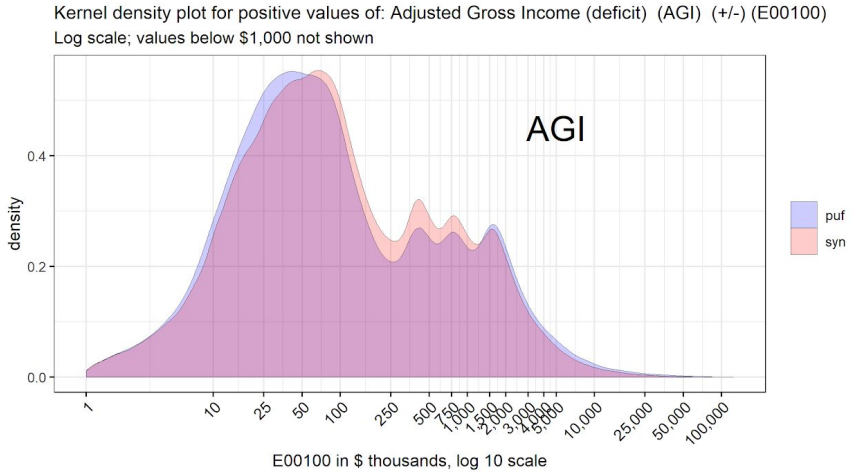
Histogram and kernel density plot for positive values of AGI in the PUF

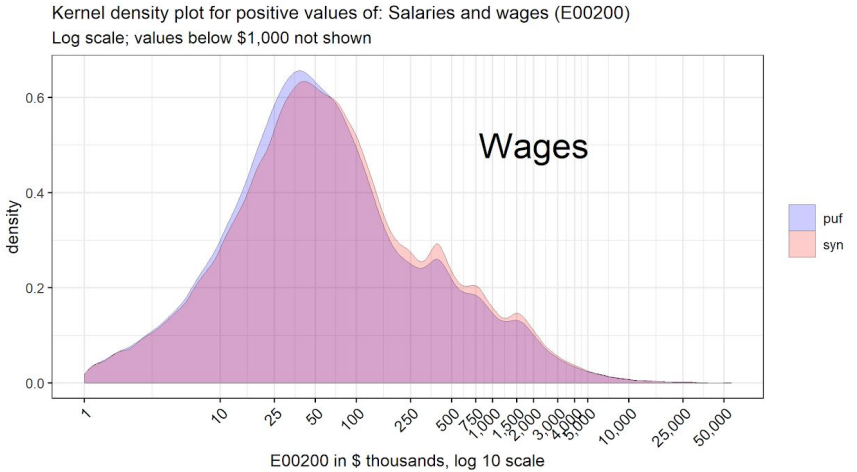
Log scale; values below $1,000 not shown

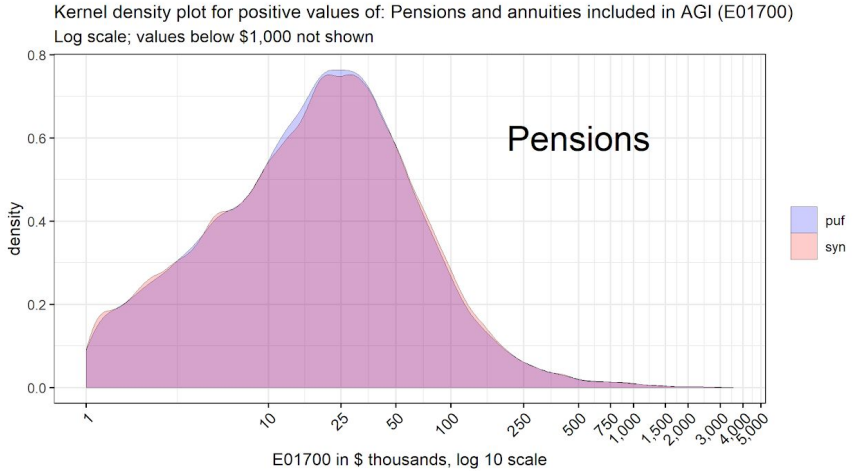


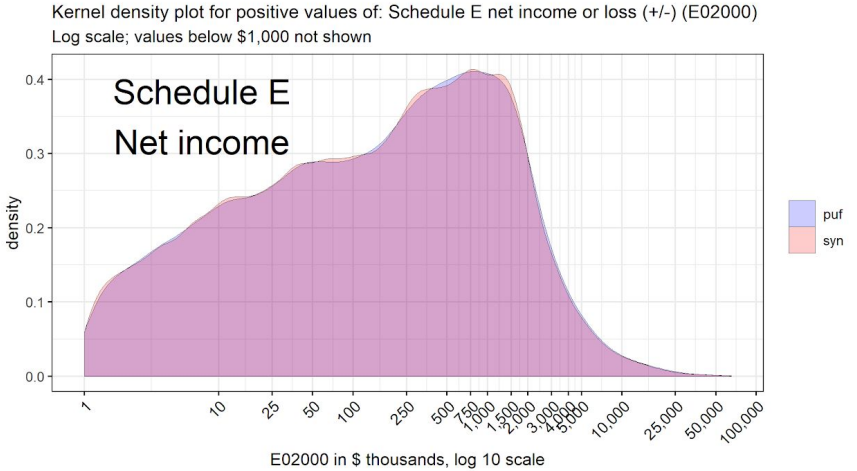

Note: Untransformed AGI has very steep hump near $60k & extremely long right tail.

# Kernel density plots, 4 large income variables (unweighted)

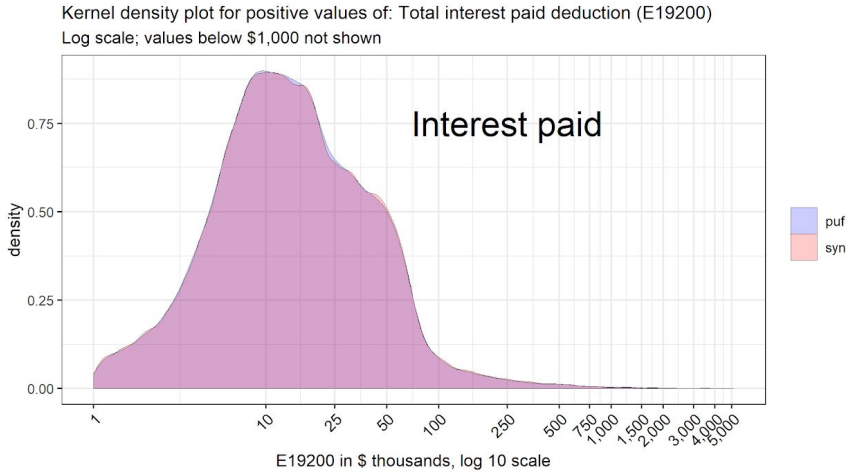


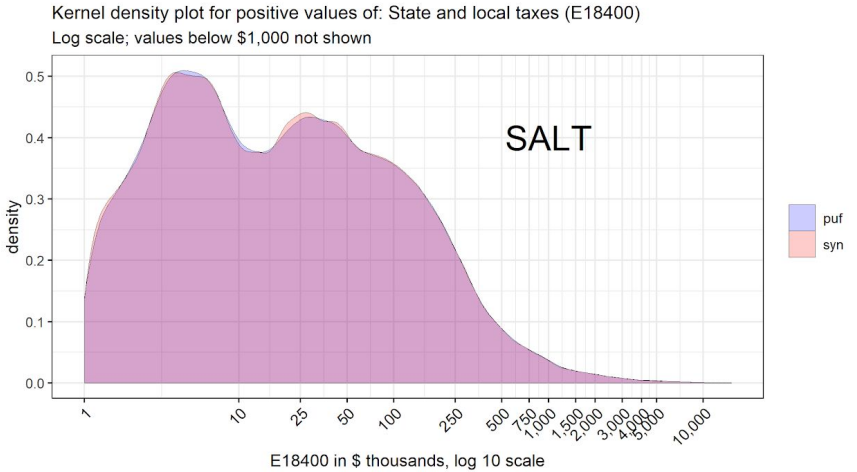



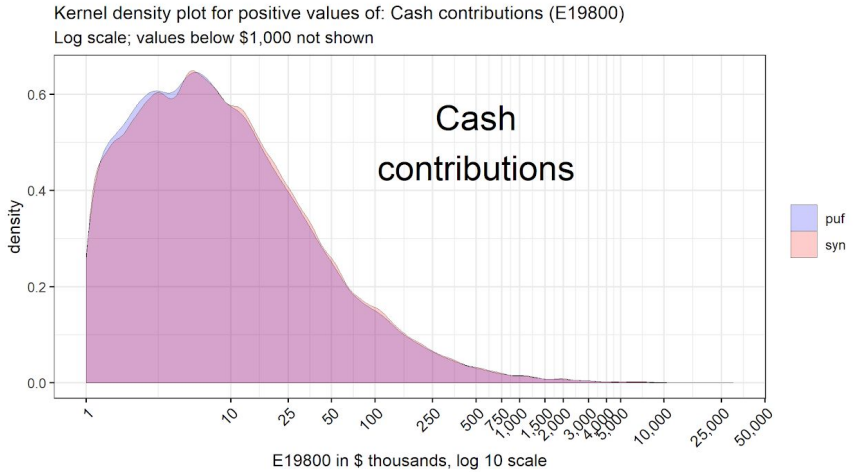



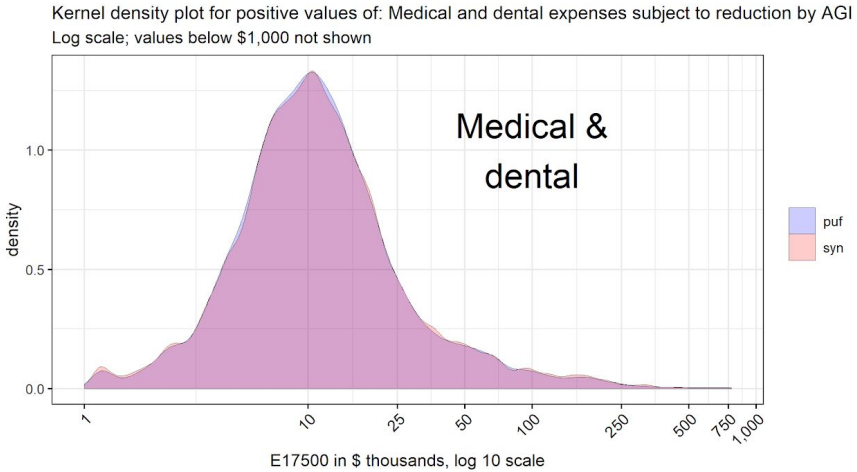



# Kernel density plots, 4 large deduction variables (unweighted)



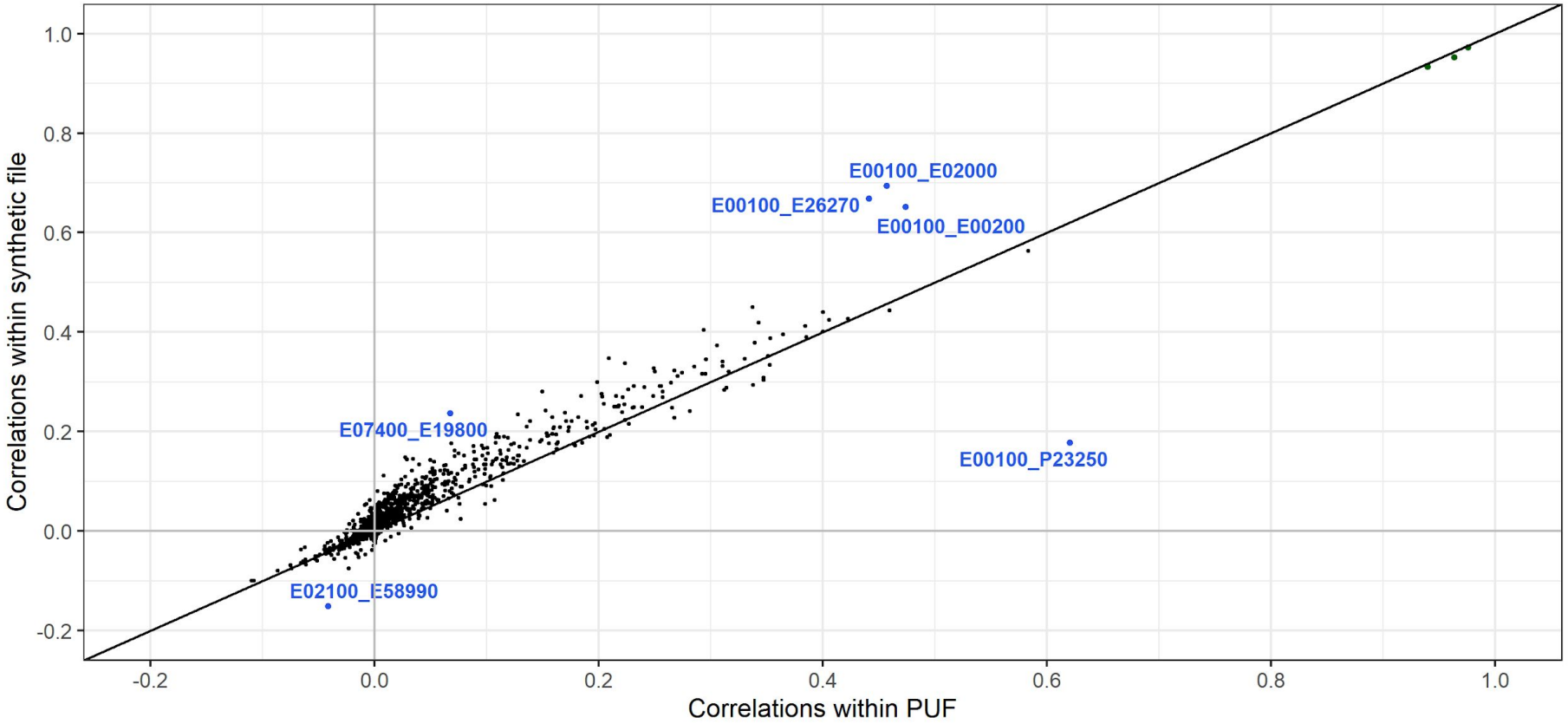



# Correlations of variable pairs within each file

Selected outlier correlation pairs labeled (see table)



# 10 worst correlation differences, sorted by absolute difference

Out of ~1,600 correlations, almost all differences are < 0.115

| Variable pair | PUF correlation | Synthetic correlation | Difference | Variable 1 | Variable 2 |
|---|---|---|---|---|---|
| E00100_P23250 | 0.621 | 0.178 | -0.442 | Adjusted Gross Income (deficit) (AGI) (+/-) | Long-term gains less losses |
| E00100_E02000 | 0.457 | 0.694 | 0.237 | Adjusted Gross Income (deficit) (AGI) (+/-) | Schedule E net income |
| E00100_E26270 | 0.441 | 0.669 | 0.228 | Adjusted Gross Income (deficit) (AGI) (+/-) | Combined partnership and S corporation net income |
| E00100_E00200 | 0.474 | 0.651 | 0.178 | Adjusted Gross Income (deficit) (AGI) (+/-) | Salaries and wages |
| E07400_E19800 | 0.067 | 0.237 | 0.170 | General business credit | Cash contributions |
| E03240_E07400 | 0.209 | 0.348 | 0.138 | Domestic Production Activities deduction | General business credit |
| E00100_E07400 | 0.150 | 0.281 | 0.132 | Adjusted Gross Income (deficit) (AGI) (+/-) | General business credit |
| E07400_E62900 | 0.027 | 0.149 | 0.121 | General business credit | Alternative tax foreign tax credit |
| E07300_E07400 | 0.028 | 0.145 | 0.116 | Foreign tax | General business credit |
| E00100_E03240 | 0.223 | 0.338 | 0.115 | Adjusted Gross Income (deficit) (AGI) (+/-) | Domestic Production Activities deduction |

# Constructing weights for the file

# Overview of our weighting approach

1. Divide PUF into mutually exclusive subsets by AGI and marital status
2. Calculate weighted PUF values for all subsets
3. Divide synthetic file similarly into subsets
4. For each synthetic-file subset:
    a. Choose weights that minimize the squared differences between weighted synthetic values and weighted PUF values (assign some targets higher priority than others)

- 124 subsets x ~128 targets each → ~16k targets

More details in the appendix.

# Enhancing the file via TaxData

# TaxData enhancement - major elements

Brings in selected information and relationships from the CPS ASEC, and prepares the file for Tax-Calculator:

- Create tax units from the CPS file
- Match tax units between the CPS and IRS-SOI
- Split certain income items between prime and spouse
- Add non-filers from the CPS to the matched file
- Extrapolate data to future years

# Comparing enhanced synthetic & PUF files

- Currently, our best comparative synthetic file:
  - Has the CPS additions
  - With a few small workarounds, mostly relating to prime-spouse splits of Schedule E and F income. (Our next synthesis will address these issues.)
  - Extrapolated to the first tax year in Tax-Calculator: 2013.
- For comparison we construct enhanced PUF with same CPS additions and workarounds, also extrapolated to 2013.
- This allows tax-analysis comparison of:

apples that are slightly bruised (synthetic)
to
apples intentionally bruised in the same way (PUF)

# Evaluating the weighted & enhanced synthetic file:  Tax-Calculator analyses

# Baseline calculations for 2017 law

# Number of returns (2017 tax law, 2013 income)
## Enhanced synthetic file and enhanced PUF, comparability-adjusted

| AGI range | # of returns in millions | | | |
|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference | % Difference |
| Negative | 1.4 | 1.4 | -0.0 | -1.0 |
| >= $0 to < 25k | 74.5 | 74.5 | 0.0 | 0.0 |
| >= $25k to < 50k | 35.2 | 35.2 | -0.0 | -0.1 |
| >= $50k to < 100k | 31.9 | 31.9 | -0.0 | -0.1 |
| >= $100k to < 200k | 15.6 | 15.6 | -0.0 | -0.1 |
| >= $200k to < 1m | 4.7 | 4.7 | 0.0 | 1.0 |
| >= $1m | 0.3 | 0.3 | -0.0 | -0.1 |
| Total | 163.6 | 163.5 | -0.0 | -0.0 |

# Adjusted gross income (2017 tax law, 2013 income)
## Enhanced synthetic file and enhanced PUF, comparability-adjusted

| AGI range | AGI in $ billions | | | |
|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference | % Difference |
| Negative | -62.0 | -60.6 | 1.4 | -2.2 |
| >= $0 to < 25k | 760.4 | 760.5 | 0.2 | 0.0 |
| >= $25k to < 50k | 1,272.9 | 1,271.6 | -1.3 | -0.1 |
| >= $50k to < 100k | 2,275.2 | 2,272.1 | -3.1 | -0.1 |
| >= $100k to < 200k | 2,086.8 | 2,081.5 | -5.3 | -0.3 |
| >= $200k to < 1m | 1,573.5 | 1,589.4 | 15.9 | 1.0 |
| >= $1m | 831.1 | 829.8 | -1.4 | -0.2 |
| Total | 8,737.9 | 8,744.3 | 6.4 | 0.1 |

# Regular tax before credits (2017 tax law, 2013 income)
## Enhanced synthetic file and enhanced PUF, comparability-adjusted

| AGI range | Regular income tax before credits in $ billions: Enhanced PUF | Synthetic file | Difference | % Difference |
|---|---|---|---|---|
| Negative | 0.0 | 0.0 | 0.0 | N/A |
| >= $0 to < 25k | 13.2 | 13.3 | 0.1 | 0.6 |
| >= $25k to < 50k | 76.3 | 76.2 | -0.0 | -0.1 |
| >= $50k to < 100k | 213.1 | 213.5 | 0.3 | 0.2 |
| >= $100k to < 200k | 265.0 | 264.8 | -0.1 | -0.1 |
| >= $200k to < 1m | 320.5 | 322.2 | 1.7 | 0.5 |
| >= $1m | 243.2 | 241.7 | -1.5 | -0.6 |
| Total | 1,131.3 | 1,131.7 | 0.4 | 0.0 |

This will come back to haunt us.

# Now, simple across-the board rate cuts vs. 2017 law

Tables show:

- $ billions change from 2017 law, PUF and synthetic
- Synthetic estimated change vs. PUF estimated change, in $ billions and as % of PUF change
- % change from 2017 law, PUF and synthetic

All tables have the same structure so we go through the first carefully.

# Across-the-board rate cuts compared with 2017 law as baseline
# Regular tax before credits, $ billions

| AGI range | Estimated impacts | | | | % change from baseline | |
|---|---|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference in estimated impacts | Difference as % of PUF estimate | Enhanced PUF | Synthetic file |
| Negative | 0.0 | 0.0 | 0.0 | N/A | N/A | N/A |
| >= $0 to < 25k | -6.0 | -6.1 | -0.0 | 0.7 | -45.6 | -45.6 |
| >= $25k to < 50k | -30.0 | -30.0 | 0.0 | -0.1 | -39.3 | -39.3 |
| >= $50k to < 100k | -70.8 | -71.0 | -0.2 | 0.2 | -33.2 | -33.3 |
| >= $100k to < 200k | -70.3 | -70.4 | -0.1 | 0.1 | -26.5 | -26.6 |
| >= $200k to < 1m | -44.9 | -45.2 | -0.3 | 0.7 | -14.0 | -14.0 |
| >= $1m | -15.0 | -15.0 | -0.0 | 0.1 | -6.2 | -6.2 |
| Total | -237.0 | -237.6 | -0.6 | 0.2 | -21.0 | -21.0 |

# Next, a complex reform with winners and losers vs. 2017 law

- Eliminate all standard and itemized deductions
- Decrease regular income tax rates
- Increase capital gains rates and pass-through-income rates

# Complex winners-losers reform compared with 2017 law as baseline
# Regular tax before credits, $ billions

| AGI range | Estimated impacts | | | | % change from baseline | |
|---|---|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference in estimated impacts | Difference as % of PUF estimate | Enhanced PUF | Synthetic file |
| Negative | 0.0 | 0.0 | 0.0 | N/A | N/A | N/A |
| >= $0 to < 25k | -1.9 | -2.0 | -0.0 | 1.9 | -14.7 | -14.9 |
| >= $25k to < 50k | -16.3 | -16.4 | -0.1 | 0.6 | -21.4 | -21.6 |
| >= $50k to < 100k | -40.8 | -41.1 | -0.3 | 0.8 | -19.1 | -19.3 |
| >= $100k to < 200k | -41.2 | -42.1 | -0.9 | 2.1 | -15.6 | -15.9 |
| >= $200k to < 1m | -11.7 | -10.1 | 1.6 | -13.3 | -3.6 | -3.1 |
| >= $1m | 17.1 | 17.8 | 0.7 | 3.8 | 7.0 | 7.4 |
| Total | -94.9 | -94.0 | 0.9 | -0.9 | -8.4 | -8.3 |

# Losers under complex winners-losers reform vs 2017 baseline law
# Regular tax before credits, $ billions

| AGI range | Estimated impacts | | | | % change from baseline | |
|---|---|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference in estimated impacts | Difference as % of PUF estimate | Enhanced PUF | Synthetic file |
| Negative | N/A | N/A | N/A | N/A | N/A | N/A |
| >= $0 to < 25k | 3.7 | 3.7 | -0.0 | -0.1 | 438.9 | 457.6 |
| >= $25k to < 50k | 6.1 | 6.0 | -0.1 | -1.9 | 108.6 | 107.5 |
| >= $50k to < 100k | 11.1 | 10.9 | -0.2 | -1.7 | 51.1 | 51.5 |
| >= $100k to < 200k | 12.4 | 12.9 | 0.5 | 4.1 | 35.3 | 39.9 |
| >= $200k to < 1m | 24.9 | 26.9 | 2.1 | 8.3 | 26.3 | 28.5 |
| >= $1m | 33.6 | 35.1 | 1.5 | 4.5 | 26.6 | 29.0 |
| Total | 91.7 | 95.5 | 3.8 | 4.1 | 32.3 | 34.7 |

# Winners under complex winners-losers reform vs 2017 baseline law
# Regular tax before credits, $ billions

| AGI range | Estimated impacts | | | | % change from baseline | |
|---|---|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference in estimated impacts | Difference as % of PUF estimate | Enhanced PUF | Synthetic file |
| Negative | N/A | N/A | N/A | N/A | N/A | N/A |
| >= $0 to < 25k | -5.6 | -5.7 | -0.0 | 0.6 | -45.4 | -45.3 |
| >= $25k to < 50k | -22.4 | -22.4 | 0.0 | -0.1 | -31.7 | -31.7 |
| >= $50k to < 100k | -51.9 | -52.0 | -0.1 | 0.3 | -27.1 | -27.1 |
| >= $100k to < 200k | -53.7 | -55.1 | -1.4 | 2.6 | -23.4 | -23.7 |
| >= $200k to < 1m | -36.5 | -37.0 | -0.5 | 1.4 | -16.2 | -16.3 |
| >= $1m | -16.5 | -17.3 | -0.9 | 5.3 | -14.1 | -14.3 |
| Total | -186.6 | -189.5 | -2.9 | 1.6 | -22.0 | -22.1 |

Drill down into
single filers
who are winners

# Single winners under complex reform vs 2017 baseline
# Regular tax before credits, $ billions

| AGI range | Estimated impacts | | | | % change from baseline | |
|---|---|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference in estimated impacts | Difference as % of PUF estimate | Enhanced PUF | Synthetic file |
| Negative | N/A | N/A | N/A | N/A | N/A | N/A |
| >= $0 to < 25k | -4.8 | -4.9 | -0.0 | 0.7 | -43.0 | -42.9 |
| >= $25k to < 50k | -14.0 | -14.0 | 0.0 | -0.2 | -29.5 | -29.5 |
| >= $50k to < 100k | -19.7 | -19.7 | -0.0 | 0.1 | -26.9 | -26.9 |
| >= $100k to < 200k | -8.7 | -9.1 | -0.5 | 5.5 | -23.1 | -23.8 |
| >= $200k to < 1m | -4.4 | -4.4 | -0.0 | 0.1 | -17.1 | -17.1 |
| >= $1m | -1.9 | -2.3 | -0.4 | 20.4 | -14.3 | -15.7 |
| Total | -53.5 | -54.4 | -0.9 | 1.7 | -25.7 | -25.8 |

How much do these bother us (for a first effort)?

# Dan Feenberg ran 108 tax reforms via NBER taxsim, on a synthetic file version and the PUF

- He adjusted, in isolation, each tax or clawback rate by 1 % point, each $ threshold (e.g., brackets) by $1k, each income, adjustment, or deduction component by 10%, & abolished each credit.
- Compared revenue impacts (% difference synthetic file vs. PUF):
  - 55% of synthetic impacts were within 5% of PUF impact
  - Tax rate and bracket reforms did very well
  - Changes to large income items (e.g., wages) generally did well
  - Reforms affecting items that few taxpayers have did very poorly (e.g., deductions for domestic production activities, self-employment health insurance)

# Lessons and next steps

# Lessons (1)

- Unweighted files: Univariate statistics, distributions, and correlations generally look quite good. Problems with gains and losses. Diagnostics will help us determine fixes.
- Weighted file, baseline tax: Synthetic file is close to enhanced PUF on many aggregates. Problems in highest income ranges and AMT.
- Simple rate cut: Easy: we just have to get weighted distribution of AGI and taxable income right, not income components.
- Complex reform: Did well in aggregate, but:
  - Top 2 income ranges gave trouble. Likely related to AGI mis-distribution. Could be a weighting or a synthesis problem.
  - Drilling deeply into winners/losers and marital status reveals some large % errors, albeit small relative to overall tax liability.

# Lessons (2)

- Dan's multitude of reforms: Items affecting small groups of taxpayers are problematic.
- Status: Not ready for all-purpose real-world decisions (aspirational). Useful for selected reforms. Useful for professors and students for policy analysis.
- These analyses teach us how to:
  - Improve our synthesis
  - Improve our weighting
  - Provide guidance about safe and unsafe uses

Improved files will open up more potential uses.

# Near term next steps

- Run current synthetic file fully through TaxData so that it can be used by others as *experimental* file in Tax-Calculator and Tax-Brain.
- When we do this will depend on press of other things, available resources, and level of interest.
- Seek support for improvements in all aspects of project.

Please let us know what you think!

# Appendix:

- Additional details
- Bibliography

# Examples

- Census Bureau's Synthetic Longitudinal Business Database: "Unless validated, there is no guarantee results from the SynLBD reflect results from the underlying confidential data"
- Scottish Longitudinal Study synthetic data: "Bespoke synthetic extracts are produced using the R package ***synthpop***...Variables are synthesised one by one using sequential regression modelling...The SLS synthetic dataset will look and behave relatively similarly to the real data. However, there is no guarantee of the validity of the results...Any analysis for publication must be run on the real dataset within the SLS Safe Setting"
- TPC goals: Fully synthetic tax files for testing programs; full programs to be run on IRS computers in safe environment.

# This is different from and complements the Tax Policy Center's project on synthetic data

- TPC project is far more extensive: substantial edits to underlying IRS and SSA data and construction of a far larger fully synthetic PUF directly from source tax data; many more schedules and items than current PUF.
- TPC synthetic PUF likely to be richer and more faithful to tax data.
- Our project is limited to creating synthetic data directly from current PUF, and *data will be available much sooner*.
- Depending on how the projects evolve a TPC fully synthetic PUF, when ready, could be an improved replacement for synthetic data we create.

# Specific data and methods choices

- Data: 2011 raw PUF. Delete the 4 aggregate records. Select and synthesize 65 variables (55 continuous, 10 categorical) needed for Tax-Calculator.
- Enforce relationships:
  - Separately synthesize pensions and annuities included in AGI (e01700) and those not included; construct their sum (e01500).
  - Use a variant on this approach for qualified and ordinary dividends.
- Seeds:
  - Use marital status and record weight as is
  - Use 9 additional variables on RHS for all estimations/predictions but drop from output: AGI (e00100), exemption amount (e04600), deduction amount (p04470), taxable income (e04800), AMT income (e62100), income tax before credits (e05800), income tax after credits (e08800), earned income for the EIC (e59560), and non-passive income (e26190). Much later, calculate from components.
- Visit sequence: largely is descending size by sum of weighted values in PUF
- Output: 5x as many synthetic records as PUF records -- 750-800k records

# Weighting approaches considered

- Reweighting:
  - When we created the data, we synthesized (modeled and predicted) record weights based on relationships to other variables
  - Reweighting starts with these synthesized weights and adjusts:
    - So that the weighted file hits or comes close to targets based on PUF aggregates,
    - While penalizing adjustments - the less adjusted, the better
- “Weighting from scratch”:
  - Throw away (ignore) the synthesized weights, and instead...
  - Choose new weights from whole cloth in a way that minimizes difference between PUF-based targets and weighted file values

# Currently we use “weighting from scratch”

1. Divide PUF into subsets of ~1-2k records based on marital status and AGI -- 124 subsets.
2. For each subset, for each of ~55 variables, calculate 4 measures: weighted numbers of positive values & negative values, weighted sums of positive & negative values -- i.e., ~220 potential targets.
3. Remove certain redundant targets. This gives ~128 targets per subset.
4. Divide synthetic file into subsets using same cutpoints (~5-10k records each).
5. For each subset, choose weights for each record that minimize a priority-weighted sum of squared differences between actual PUF targets and computed synthetic sums (using the chosen weights).
6. Highest priority in the sum of squared differences is given to AGI and a few other critical variables. This is based on judgment.

# Simplified example for one file subset

- Let's say PUF file subset # 17 (out of 124 subsets) has 1k married-joint PUF records with AGI in the $10-20k range and:
  - Total AGI of $580 million
  - Total negative capital gains of -$50 million
  - ...and 126 other targets
  - We want AGI target to be 20x as important as capital gains target
- Corresponding synthetic subset has 5k married-joint records with $10-20k AGI. We choose 5k weights ws[i] that minimize the penalty function:

```
      20 * {sum(ws[i] * agi[i]) - 580}^2
+      1 * {sum(ws[i] * (capgains[i] if negative)) -  -50}^2
+       priority-weighted squared differences from ~126 additional targets
```

Note: We also do some scaling, not shown here.

# Solving the weighting problem

- We have:
  - 124 file subsets
  - ~5-10k synthetic records per subset, choose weight for each record
  - ~128 targets per subset
  - Thus we have ~16k targets (124 x 128) for the entire synthetic file
  - File subsets are mutually exclusive and can be solved in parallel
- Penalty function is nonlinear. No constraints (they are built into objective function). Bounds on the weights (e.g., >=1).
- Currently we solve with MMA (method of moving asymptotes)
- Solves in about an hour, in parallel (6 cores), with 500 iteration limit
- Generally suboptimal but very good solutions; could run longer

# Additional baseline calculations for 2017 law

# Total regular tax + AMT (2017 tax law, 2013 income)

## Enhanced synthetic file and enhanced PUF, comparability-adjusted

| AGI range | Total regular and AMT tax before credits in $ billions | | | |
|---|---|---|---|---|
| | Enhanced PUF | Synthetic file | Difference | % Difference |
| Negative | 3.7 | 0.0 | -3.7 | -100.0 |
| >= $0 to < 25k | 13.9 | 13.3 | -0.6 | -4.6 |
| >= $25k to < 50k | 76.7 | 76.3 | -0.5 | -0.6 |
| >= $50k to < 100k | 214.3 | 213.7 | -0.6 | -0.3 |
| >= $100k to < 200k | 267.4 | 266.4 | -0.9 | -0.4 |
| >= $200k to < 1m | 338.5 | 340.5 | 2.0 | 0.6 |
| >= $1m | 249.8 | 248.3 | -1.5 | -0.6 |
| Total | 1,164.4 | 1,158.6 | -5.8 | -0.5 |

# Selected Relevant Reports and Papers (1)


Benedetto, Gary, Jordan C Stanley, and Evan Totty. “The Creation and Use of the SIPP Synthetic Beta v7.0.” U.S. Bureau of the Census, November 2018. https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnicalv7.pdf.

Benedetto, Gary, and Martha Stinson. “Disclosure Review Board Memo: Second Request for Release of SIPP Synthetic Beta Version 6.0.” Survey Improvement Research Branch, Social, Economic, and Housing Statistics Division (SEHSD), U.S. Bureau of the Census, January 15, 2015.

Bryant, Victoria L. “General Description Booklet for the 2011 Public Use Tax File.” Individual Statistics Branch, Statistics of Income Division, Internal Revenue Service, August 2016.

Burman, Leonard E, Alex Engler, Surachai Khitatrakun, and Sarah Armstrong. “A Synthetic Income Tax Return Data File: Tentative Work Plan and Discussion Draft,” June 30, 2017, 42. https://www.taxpolicycenter.org/sites/default/files/publication/142421/2001396-a-synthetic-income-tax-return-data-file-tentative-work-plan_and_discussion_draft.pdf.

Burman, Leonard, Surachai Khitatrakun, and Philip Stallworth. “Proposed Methodology for Creating a Fully Synthetic Dataset and Privacy Implications.” For presentation at Syracuse University Maxwell School Center for Policy Research, October 2018. https://www.maxwell.syr.edu/uploadedFiles/cpr/events/cpr_seminar_series/Synthesis%20methodology%20and%20privacy%20Center%20for%20Policy%20Research.pdf.

Caiola, Gregory, and Jerome P Reiter. “Random Forests for Generating Partially Synthetic, Categorical Data.” *Transactions on Data Privacy* 3 (2010): 16.

# Selected Relevant Reports and Papers (2)


Drechsler, Jörg. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer Science & Business Media, 2011.

Drechsler, Jörg, and Jerome P. Reiter. “An Empirical Evaluation of Easily Implemented, Nonparametric Methods for Generating Synthetic Datasets.” *Computational Statistics & Data Analysis* 55, no. 12 (December 2011): 3232–43. https://doi.org/10.1016/j.csda.2011.06.006.

Reiter, Jerome P. “Satisfying Disclosure Restrictions With Synthetic Data Sets.” *Journal of Official Statistics*, 2002, 19.

Reiter, J P. “Using CART to Generate Partially Synthetic Public Use Microdata.” *Journal of Official Statistics* 21, no. 3 (2005): 22.

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. “General and Specific Utility Measures for Synthetic Data.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181, no. 3 (June 2018): 663–88. https://doi.org/10.1111/rssa.12358.

Winglee, Marianne, Richard Valliant, Jay Clark, Yunhee Lim, Michael Weber, and Michael Strudler. “Assessing Disclosure Protection for a SOI Public Use File.” Presented at the 2002 American Statistical Association: Internal Revenue Service, 2002. https://www.irs.gov/pub/irs-soi/weber.pdf.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation.” *Journal of Privacy and Confidentiality* 1, no. 1 (April 1, 2009). https://doi.org/10.29012/jpc.v1i1.568.

Zayatz, Laura. “Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update.” Research Report Series, August 31, 2005.