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
View or 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

- Income tax policy models need a microdata input file that represents the population of interest
- PSL TaxData prepares & enhances 2 such files for Tax-Calculator:
  - A PUF-based version that enhances the IRS Public Use File
  - A CPS version that enhances the Current Population Survey
- Enhanced PUF represents taxpaying universe well for many purposes. (Does not include non-taxable benefits. And not perfect.)
- PUF costs ~$10k and requires legal agreement.
- CPS is missing important income components & does not represent high-income taxpayers well. Tax liability differences vs. PUF can be significant (e.g., 20%). However, CPS-based file includes non-taxable benefits -- important for some analyses.

Synthetic data

- Data “created” by means other than direct measurement - e.g., by algorithm. Intended to preserve key statistical characteristics of true data: means, variances, correlations, patterns of missingness, while not containing confidential information that must be suppressed.
- “Fully synthetic” data have no real data (unlike data with imputations).
- Goals - non-confidential data that mimic true data well enough that:
  - Analysts can practice and learn how to use a restricted-access confidential data source.
  - Analysts can test out programs and get “pretty good” results prior to using confidential “gold standard” data in a safe (restricted) environment.
  - Holy Grail: Results can be treated as if they are based on gold-standard data. (Not clear whether/when this can be done.)
- Ex: Synthetic versions of Longitudinal Business Database, SIPP

Free unrestricted synthetic tax data can allow:

- Much wider use given that it is free
- Testing of programs without restriction
- Desktop analysis of federal income tax policies
  - Students, professors -- public policy, economics
  - Fiscally oriented data journalists
  - Cash-strapped nonprofit fiscal analysis groups
  - State policymakers/advisers concerned about federal tax impacts
- No blurring of Tax-Brain results
- Ability to focus TaxData resources on a single data approach
- Spinoff state-specific synthetic data files
  - Another OSPC-incubated project: Could allow analysis of state income taxes, as well as federal tax impact on a state

Our goals are to create synthetic tax data that:

1. Can satisfy IRS/SOI disclosure review
2. Can be made available for free
3. Without legal agreements
4. Is usable as input to tax-calculator models
5. Is useful for some kinds of tax policy analyses
6. We can identify good uses for, and dangerous uses

We have a version that fully satisfies 1-3 and partially satisfies 4-6.

We are learning how to improve it, and plan to do so.

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.

Steps in the process

1. Create synthetic data file (details in a minute)
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 steps

1. Create one or more “seed” variables
2. Construct synthetic records, variable by variable:
    a. Construct a model for the variable using only RHS variables that have been modeled or “seeded”
    b. Predict synthetic values for the variable based upon predicted values of variables modeled so far
    c. Repeat until all variables are synthesized

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

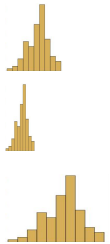
TRUE

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

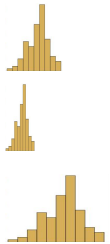
TRUE

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

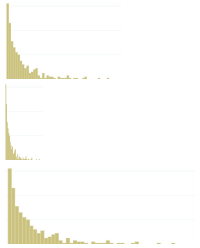
TRUE

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

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

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

- By definition and design, synthesized records are not actual data. Every record is constructed from a model, not by adding noise to data.
- However, it is possible to produce by chance a synthesized record that matches a PUF record on all variables. This is more likely to occur for a trivially simple record that occurs multiple times in the data set, such as a record for which all income and deduction variables are zero, than for a complex record that is unique.
- We were asked to follow the rule below. We followed this rule and obtained certification that the file meets this requirement:

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

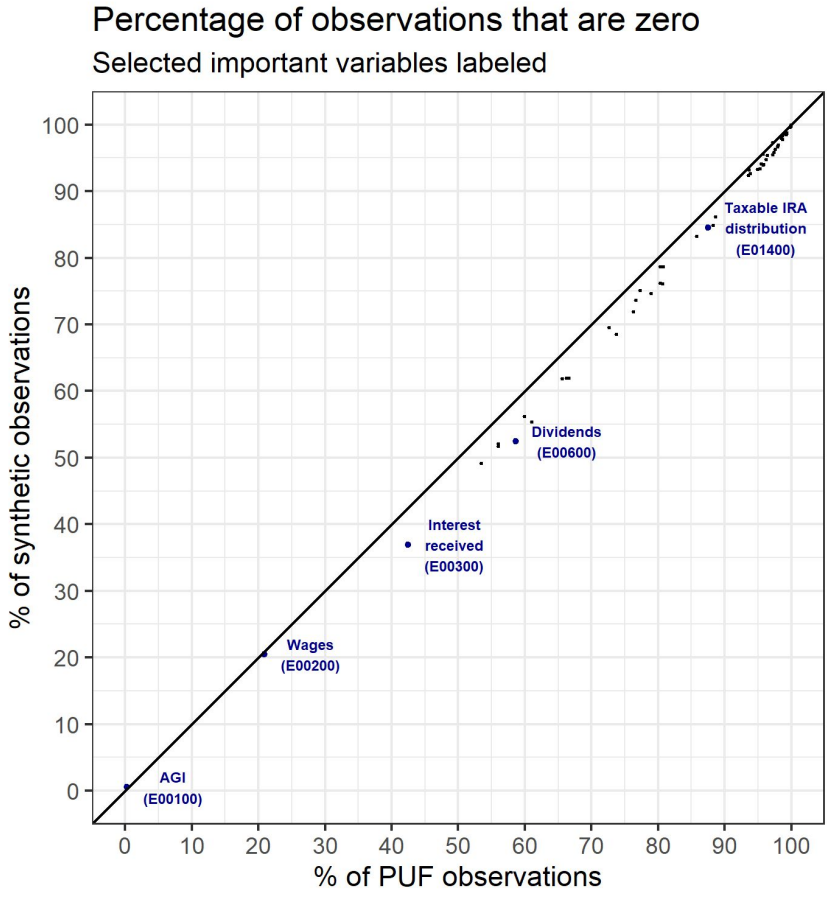
Evaluating the synthetic file (before weighting)

Examining synthesis results (unweighted)

- Individual variables
  - Complicated: 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 variable by categorical variable groupings (e.g., by marital status)

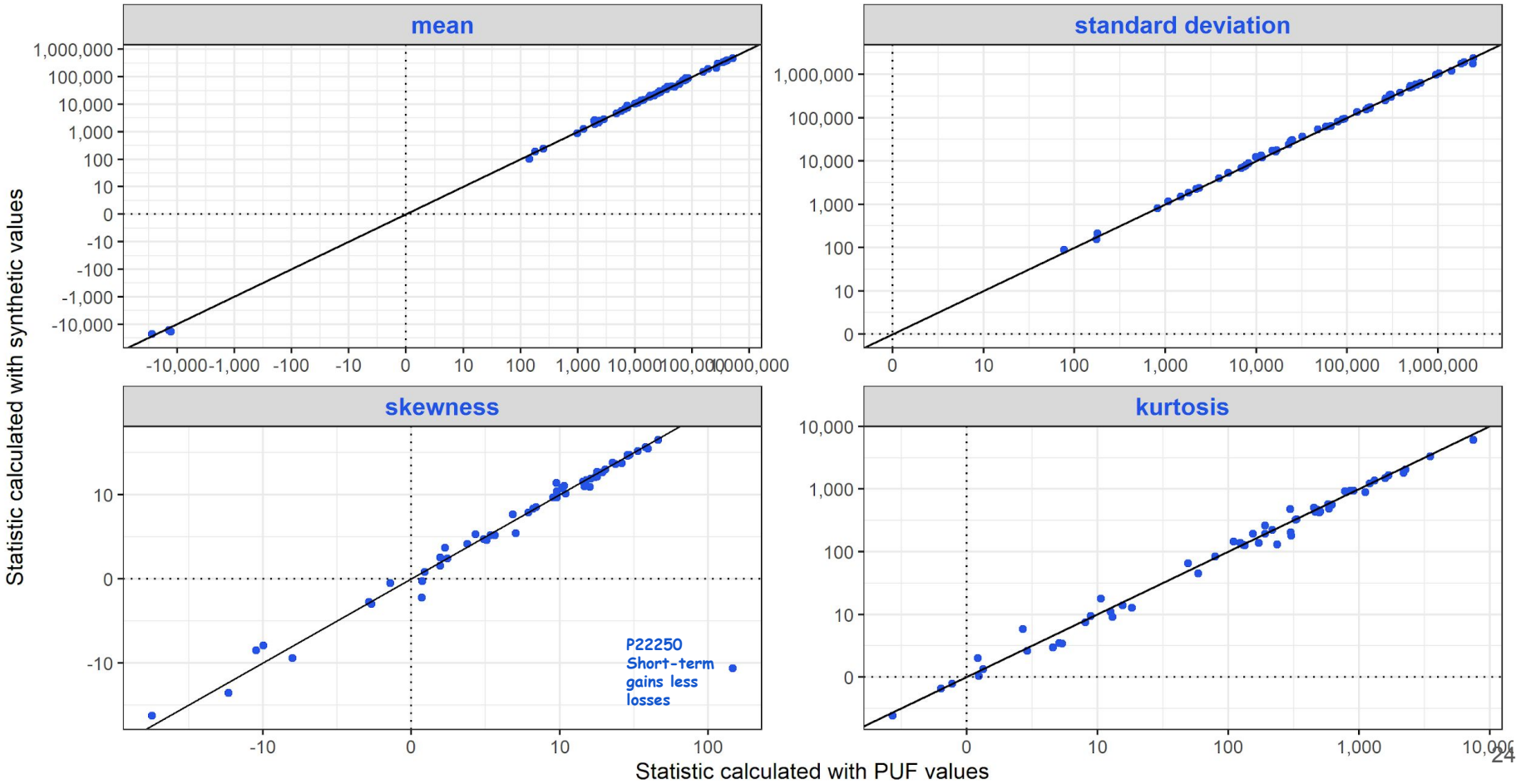
We (slightly) under-synthesize zero-values

- Each observation (dot) in the scatterplot is one of our 55 continuous variables
- The horizontal axis is the % of this variable's values in the PUF that are zero
- The vertical axis is the % of this variable's values in the synthetic file that are zero
- If a dot is on the 45-degree line, then synthetic and PUF files have the same % of observations that are zero.
- Below the line means the synthetic file has a smaller % of nonzero values.

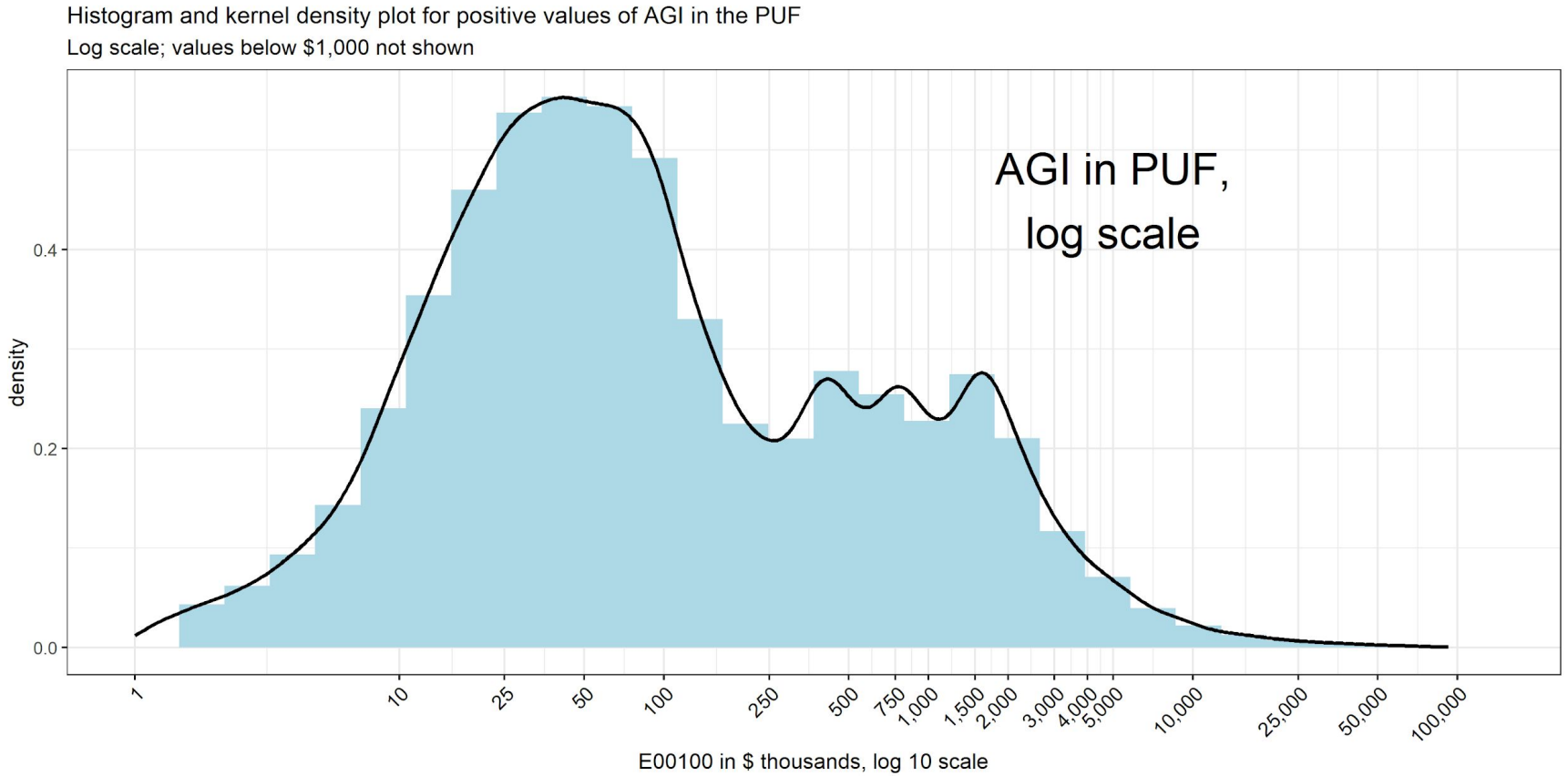


Statistics calculated using nonzero values, log scales, with 45-degree line

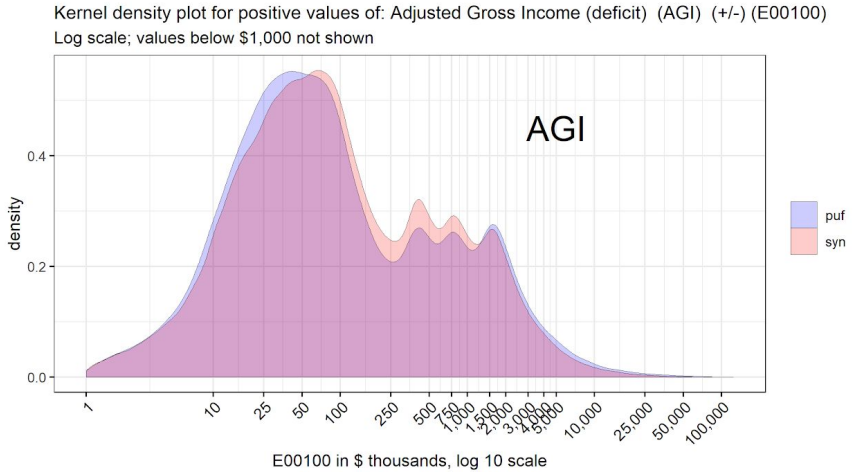
Each point gives the PUF and synthetic-file statistic for a continuous variable

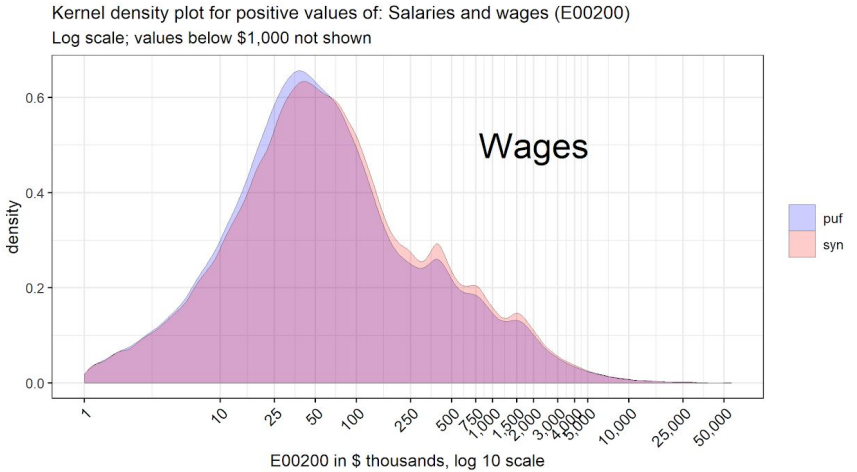


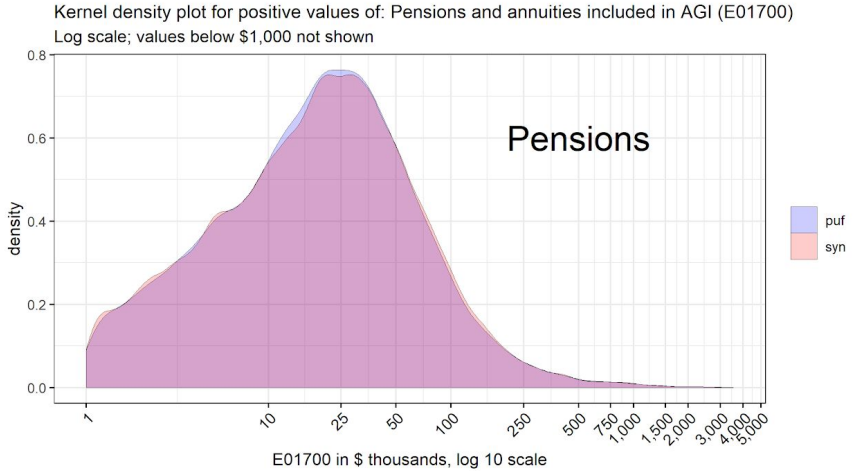
Kernel density plot is like a smoothed histogram. Useful for comparing distributions.

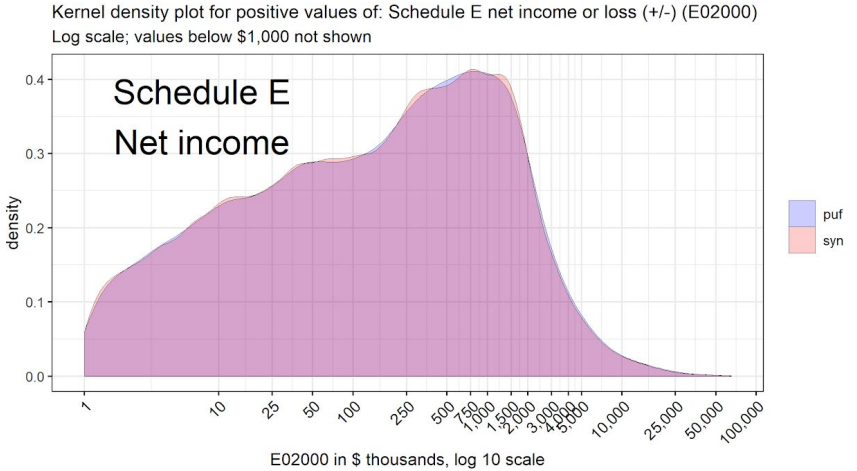


Kernel density plots, 4 large income variables (unweighted)

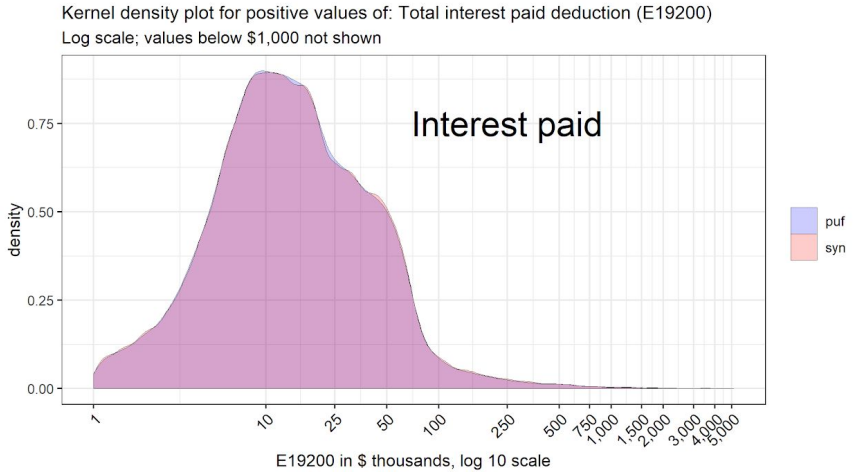


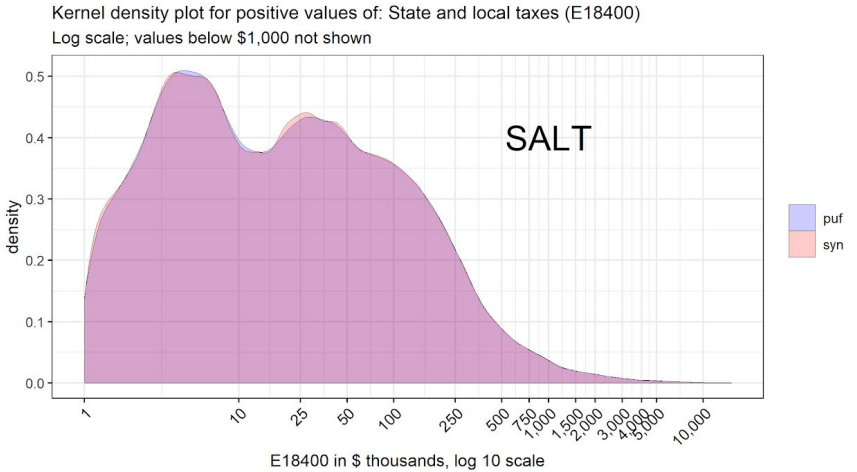


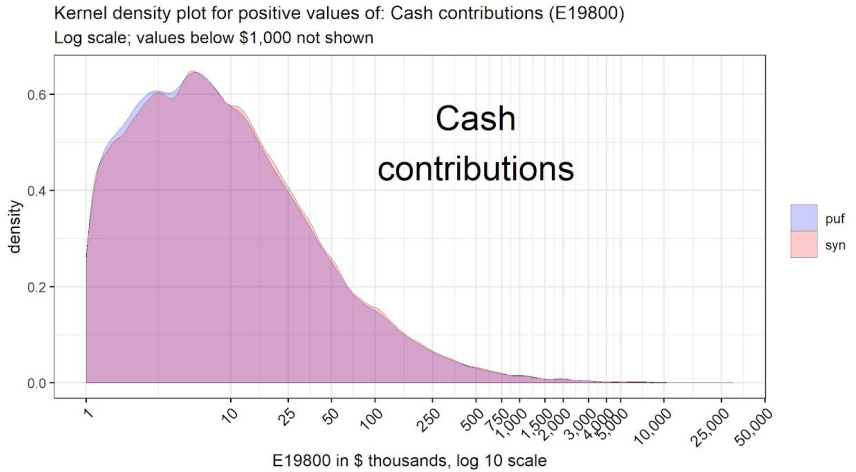


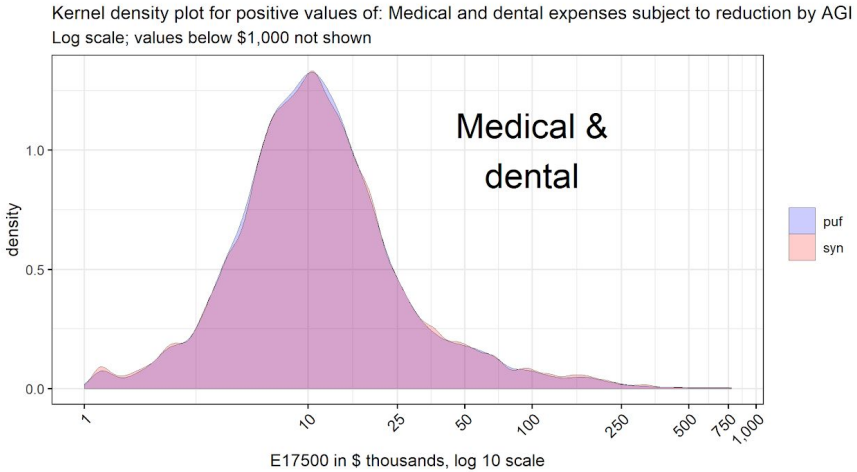


Kernel density plots, 4 large deduction variables (unweighted)



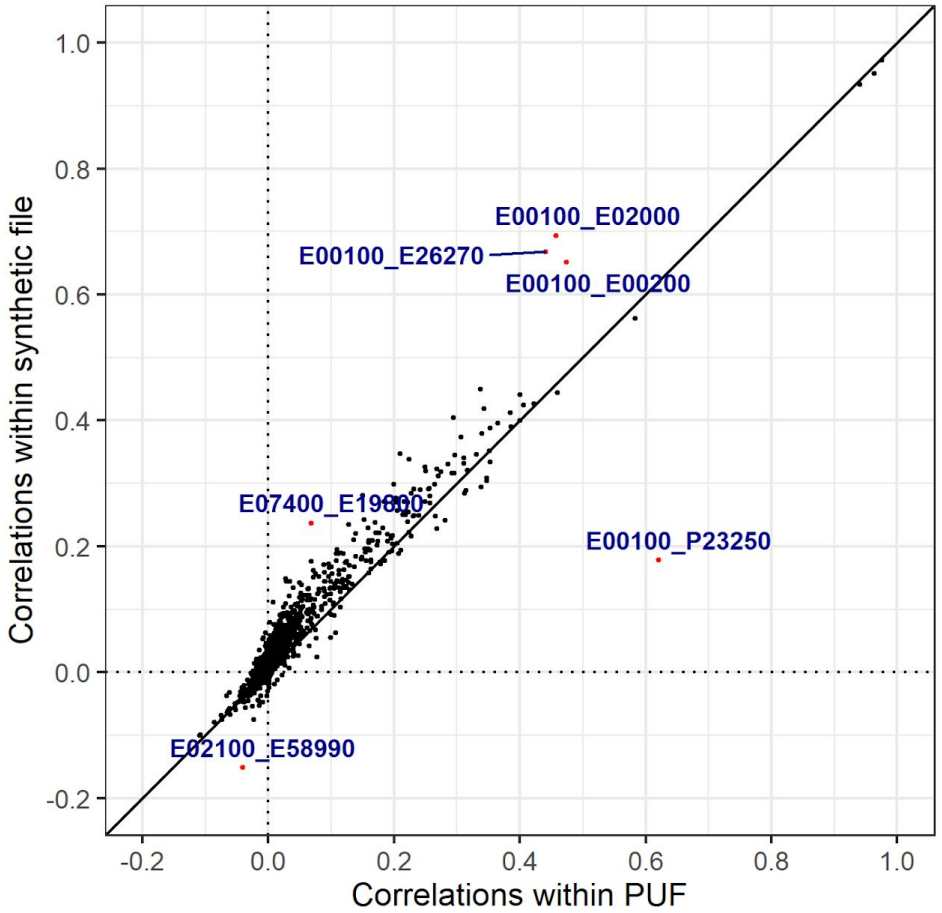






Correlations of variable pairs within each file, compared across files

Selected outlier correlation pairs labeled



Variables included in labeled correlation pairs:

E00100: Adjusted Gross Income (deficit) (AGI) (+/-)
E00200: Salaries and wages
E02000: Schedule E net income or loss (+/-)
E02100: Schedule F net profit/loss (+/-)
E07400: General business credit
E19800: Cash contributions
E26270: Combined partnership and S corporation net income/loss (+/-)
E58990: Investment income (Form 4952 part 2 line 4g)
P23250: Long-term gains less losses

10 worst correlation differences (difference >= ~0.12)

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 or loss (+/-)
E00100_E26270	0.441	0.669	0.228	Adjusted Gross Income (deficit) (AGI) (+/-)	Combined partnership and S corporation net income/ loss (+/-)
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

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

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
- Develop factors needed to extrapolate data to future years:
  - Growth factors for wages, dividends, and other variables
  - New weights that allow the file to hit targets for future years
  - Ratios to adjust distribution of certain interest income

So far we have succeeded in applying some but not all TaxData enhancements

- Harder than I expected
- Several small issues have taken time to identify and resolve.
- Extrapolating to future years with 5x as many records requires careful examination and review to ensure good results.
- Issues are resolvable; our next synthesis will nip some in the bud.

Comparing enhanced synthetic & PUF files

- Currently, our best comparative file:
  - Has the CPS additions
  - With a few small workarounds, mostly relating to prime-spouse splits of Schedule E and F income.
  - Extrapolated to the first tax year in Tax-Calculator: 2013.
- We construct an enhanced PUF that has the 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 analysis

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.

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

Lessons from our baseline analysis

- Enhanced synthetic file is close to enhanced PUF on many aggregates.
- Further off than I would like in upper-income ranges
- We are far off on AMT and thus further than we would like on total income tax liability (regular + AMT).
- We know some reasons for differences. I believe the problems generally will be solvable.
- Some of the solution will come from better weighting.
- Some may come from better synthesis.
- It is a lot of work to track the reasons down and then solve the issues. But it is hard work, not impossible work.

Now we examine two tax reforms, vs. 2017 law

1. Simple across-the board rate cuts

2. A complex reform with winners and losers
   a. Eliminate all standard and itemized deductions
   b. Decrease regular income tax rates
   c. Increase capital gains rates and pass-through-income rates

For the complex reform, we drill down into winners and losers, including by marital status.

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

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

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

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

Drill down further and look at 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)?

Lessons from these tax-reform analyses

- Simple rate cut: Did very well. Easy: we just have to get weighted distribution of AGI and taxable income right, not income components.
- Complex reform:
  - In aggregate - did well
  - Top 2 income ranges gave trouble. Likely related to AGI mis-distribution in these ranges. Could be a weighting or a synthesis problem. Should be improvable.
  - Drilling deeply into winners/losers and marital status reveals some large % errors, albeit small relative to overall tax liability.

In addition, Dan Feenberg ran 108 tax reforms on a version of the synthetic file 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.
- % difference, synthetic file revenue impact vs. PUF impact:
  - 55% of synthetic impacts were within 5% of PUF (45% worse)
  - 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

- Synthetic file clearly is useful for some kinds of tax reforms
- Potentially unacceptable results when we drill into small slices of the data or examine reforms that affect narrow areas of the income tax.
- These analyses teach us how to:
  - Improve our synthesis
  - Improve our weighting
  - Provide guidance about safe and unsafe uses
- Not ready for real-world policy decisions. It can be useful for first cuts at analyses and for pedagogical and “practice” uses.
- We are learning how to make it much more useful.

Next steps

- Run current synthetic file fully through TaxData so that it can be used as *experimental* file in Tax-Calculator and Tax-Brain.
- “When” depends on press of other things, resources, and interest.
- Seek outside support for improvements in all aspects of project.

Please let us know what you think!