November 2, 2021

Synthetic Public Use File of Administrative Tax Data

Methodology, Utility, and Privacy Implications

URBAN INSTITUTE · ELEVATE · THE · DEBATE

Claire McKay Bowen, Ph.D
Principal Research Associate
cbowen@urban.org www.clairemckaybowen.com @ClaireMKBowen

Our Project Team

- Andrés Felipe Barrientos – Assistant Professor, Dept. of Statistics, Florida State University
- Claire Bowen – Principal Research Associate, Urban Institute
- Len Burman – Institute Fellow, Tax Policy Center, Urban Institute
- John Czajka – Senior Fellow, Mathematica Policy Research
- Lillian Hunter – Research Assistant, Tax Policy Center, Urban Institute
- Surachai Khitatrakun – Senior Research Methodologist, Tax Policy Center, Urban Institute
- Graham MacDonald – Associate Vice President, Technology & Data Science, Urban Institute
- Rob McClelland – Senior Fellow, Tax Policy Center, Urban Institute
- Joshua Snoke – Statistician, RAND Corp.
- Silke Taylor – Lead Software Engineer, Technology & Data Science, Urban Institute
- Aaron R. Williams – Senior Data Scientist, Income and Benefits Policy Center, Urban Institute
- Doug Wissoker – Senior Fellow, Statistical Methods Group, Urban Institute

Overview

1. **Motivation:** Why synthesize taxpayer data?
2. **Background:** What are non-filer data?
3. **Methodology:** How did we generate the synthetic data? How did we evaluate the quality and privacy?
4. **Future:** What are our next steps?

Motivation


UNITED STATES
Internal
Revenue
Service
Building


MONDAY
TUESDAY

Background

Administrative Tax Data

Master File: A massive tax database of about 145 million unedited tax returns for 2012, but cannot be used in its current state due to:

- Size
- Timing of completion (e.g., late filers)
- Item content
- Potential data inconsistences

INSOLE: Stratified sample with weights to represent the U.S. taxpayer population.

What do we mean by non-filer data?

Any U.S. resident who did not file a federal tax return, had no obligation to file, and was not claimed as a dependent in 2012 but had income reported to the IRS on at least one information return for the 2012 tax year:

- 10-in-9,999 sample of the confidential data
- Some variables are excluded and others are top-coded and/or recoded
- About 26,000 records
- 19 variables after data preprocessing

Methodology

What are our goals?

Produce a fully synthetic data file with the same record layout as IRS Administrative Data that:

- Protects the confidentiality of tax return information
- May be used for statistically valid analysis for certain research purposes
- May be used as a “training dataset” to develop programs to run on confidential data

What are our goals?

Produce a fully synthetic data file with the same record layout as IRS Administrative Data that:

- Works well for tax microsimulation
- Avoids logical inconsistencies
- Avoids disclosure risk

How did we generate the synthetic data?

- Applying synthetic data generation, a data privacy and confidentiality method that is widely accepted across many disciplines.
- Train on Classification and Regression Trees (CART) models on the confidential data
- Sequentially synthesize variables in order of linear correlation with wages (highest to lowest) with previously synthesized variables as predictors.

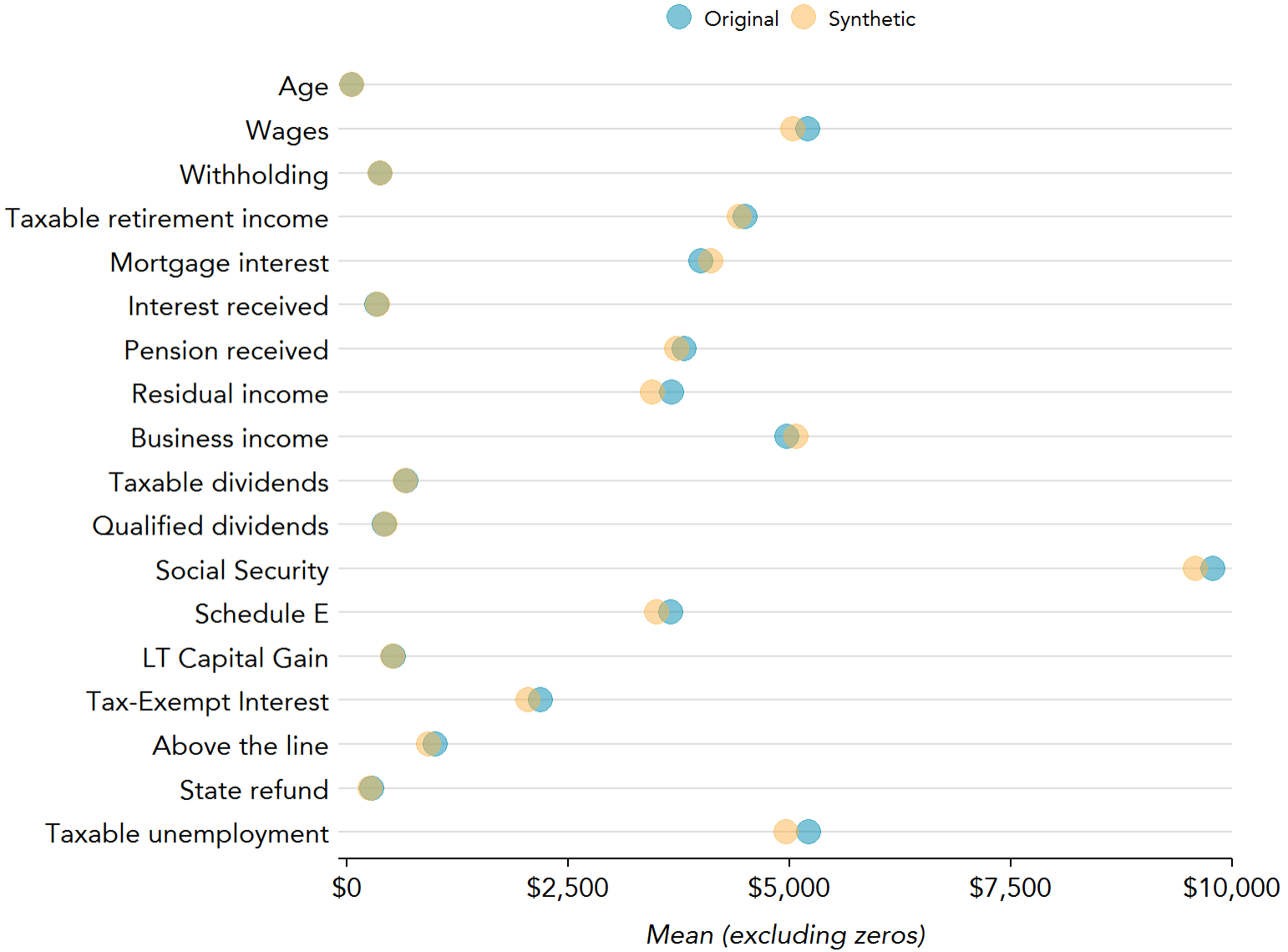
Step	Outcome	Predictors
1	Sex	Random sampling with replacement
2	Age	Sampled Sex
3	Wages	Sampled Sex, Synthetic Age
4	Withholding	Sampled Sex, Synthetic Age, Synthetic Wages
…	…	…
n	nth variable	Sampled Sex, Synthetic Age, Synthetic Wages, …, n-1st synthetic variable

How did we measure our disclosure risks?

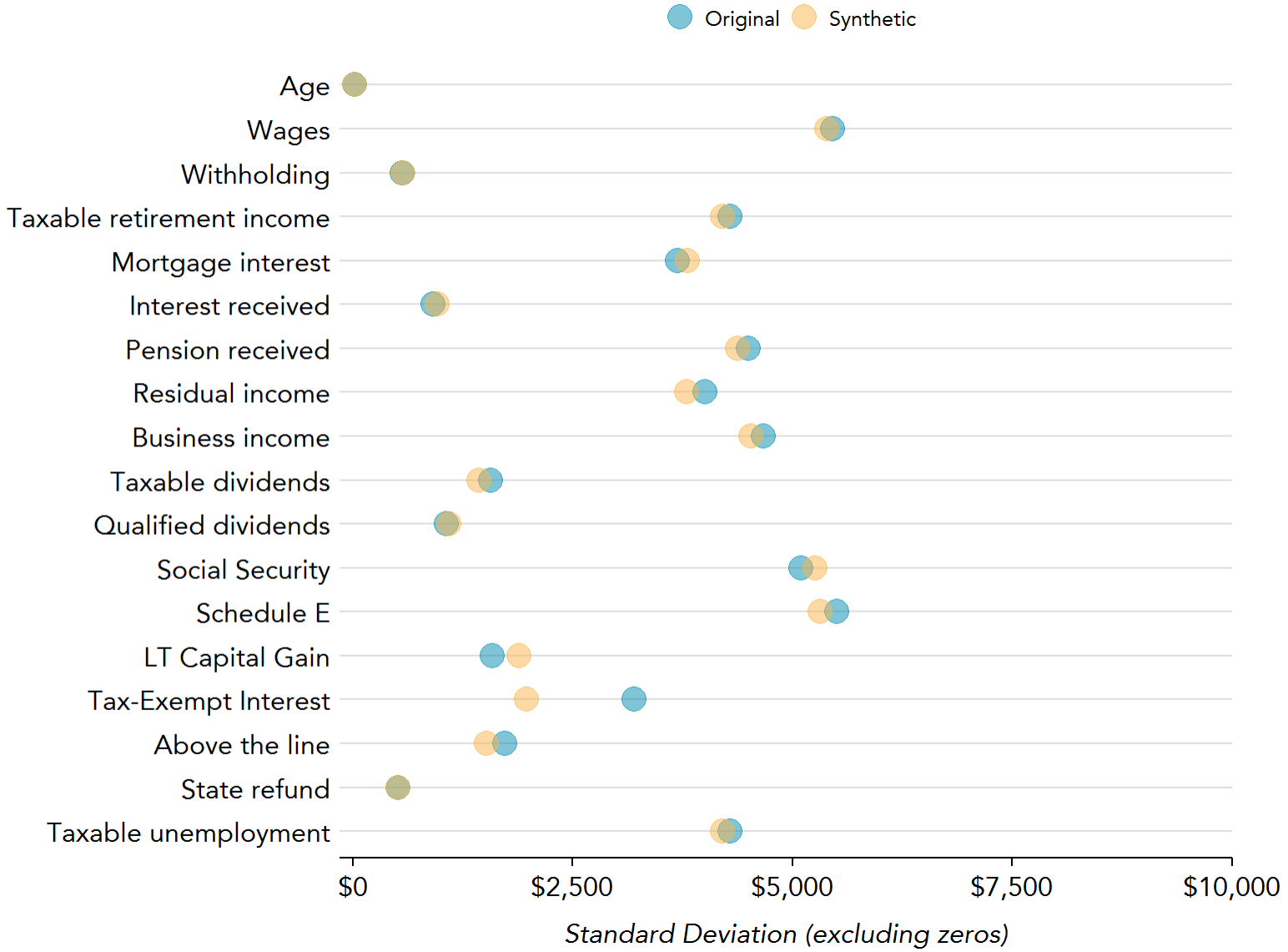
- Number of Unique-Uniques
- Row-Wise Squared Inverse Frequency
- l-Diversity of Final Nodes in the CART Algorithm

Results

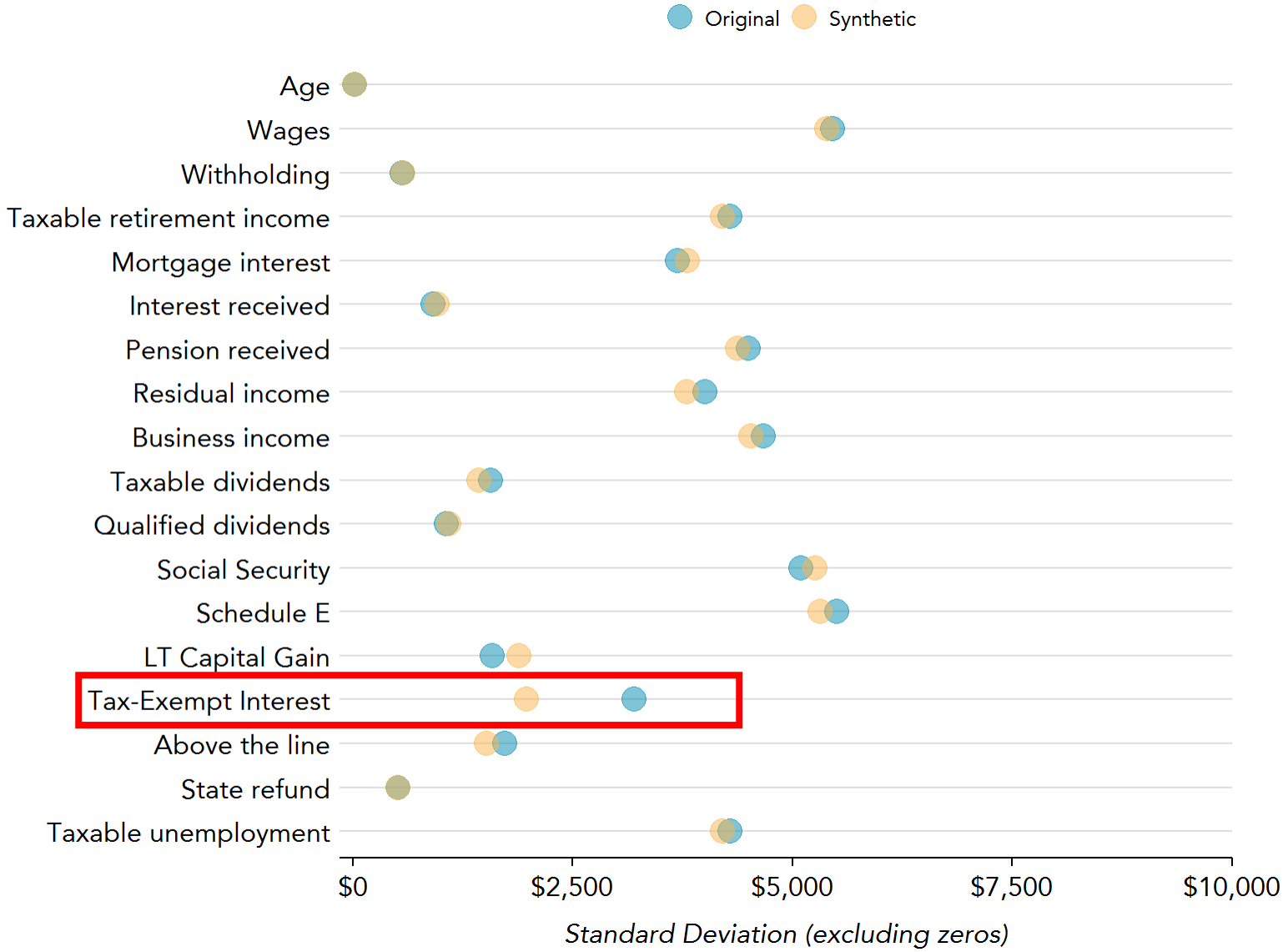
Means


Original
Synthetic
Age
Wages
Withholding
Taxable retirement income
Mortgage interest
Interest received
Pension received
Residual income
Business income
Taxable dividends
Qualified dividends
Social Security
Schedule E
LT Capital Gain
Tax-Exempt Interest
Above the line
State refund
Taxable unemployment
$0
$2,500
$5,000
$7,500
$10,000
Mean (excluding zeros)

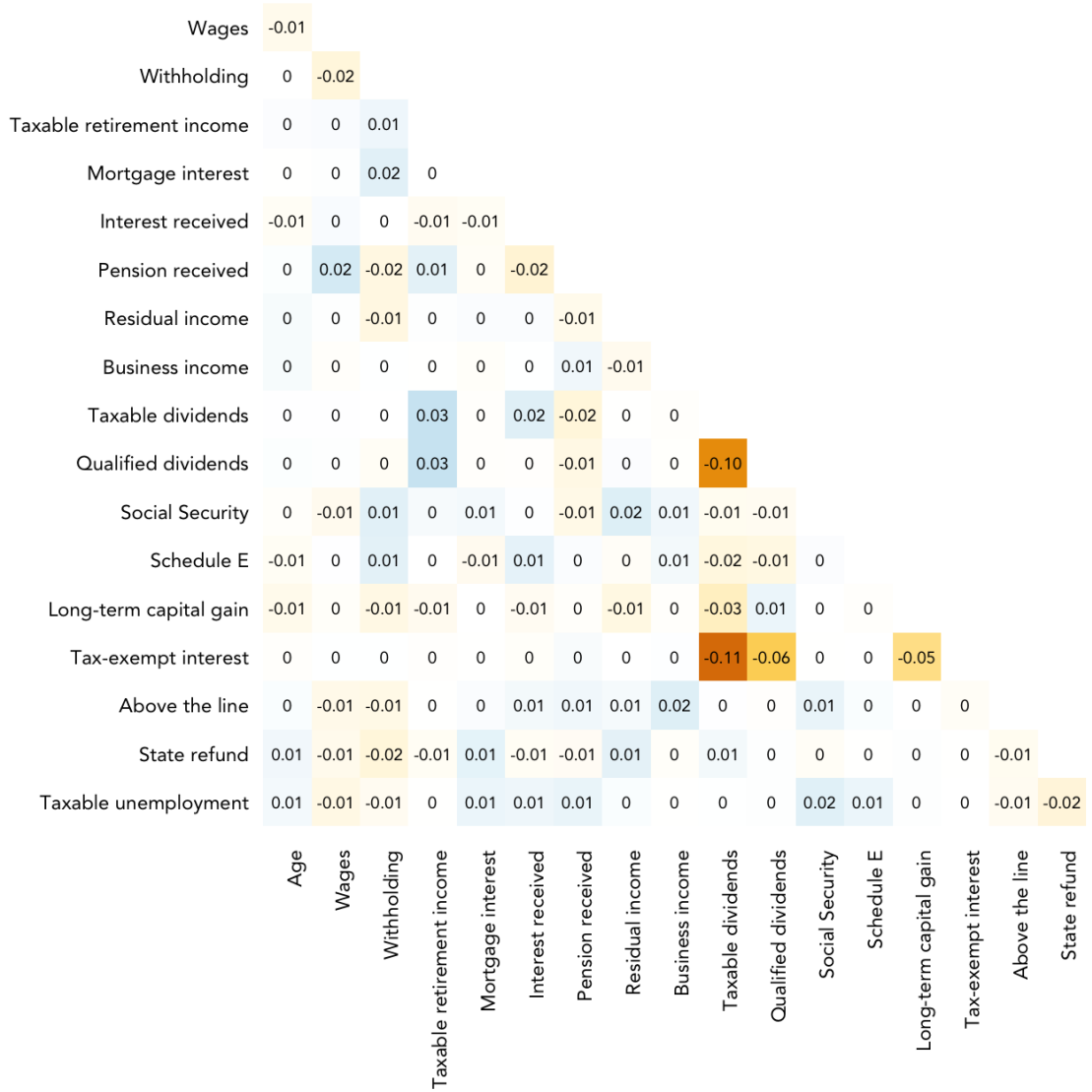
Standard Deviations



Standard Deviations



Linear Correlations



Tax Calculator

The Future

The Future

1. Additional non-filer data releases

```
# generate synthetic data set
syn1 <- nonfilers.generate_final_synthesis(df_original,
                                           threshold_ratio = 4,
                                           irp_filter = TRUE,
                                           drop_missing = TRUE,
                                           visit_sequence = "proportion",
                                           smoothing = "ntile",
                                           method = "cartpc",
                                           cartpc.minbucket = 75,
                                           cartpc.ntile = 200,
                                           seed = 8675309,
                                           top_code = TRUE)

# generate report with utility and privacy metrics
nonfilers.report(syn1$syn_nonfilers,
                 syn1$data_original,
                 report = "summary",
                 output_file = "synthesis_summary.html",
                 output_dir = "reports")
```

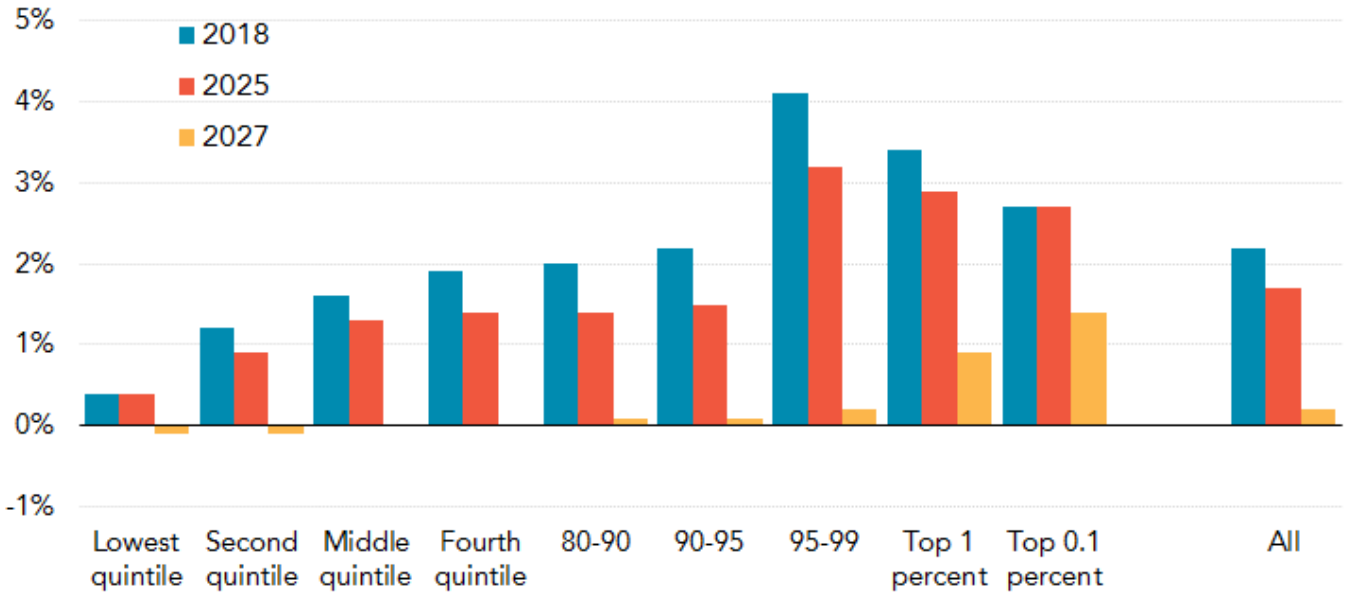
The Future

1. Additional non-filer data releases
2. Synthetic IRS PUF

FIGURE 1

Percent Change in After-tax Income of the Conference Agreement for the Tax Cuts and Jobs Act

By expanded cash income percentile, 2018, 2025, and 2027



Source: Urban-Brookings Tax Policy Center Microsimulation Model (version 0217-1).

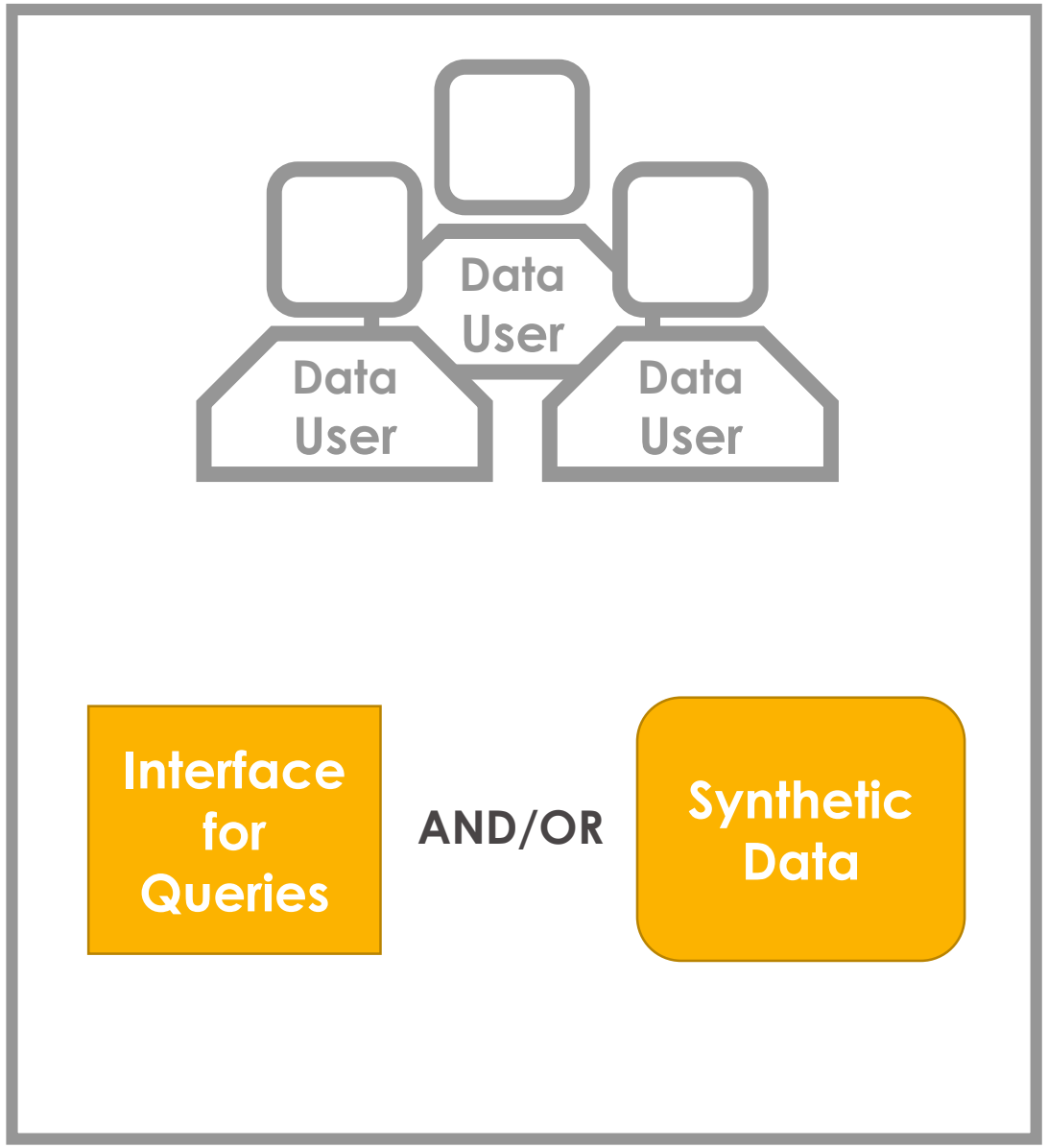
The Future

1. Additional non-filer data releases
2. Synthetic IRS PUF
3. `library(tidysynthesis)`

```
# visit_sequence
visit_sequence <- visit_sequence(conf_data = penguins_complete,
                                 type = "correlation",
                                 factor_var = c("species", "island", "sex"),
                                 cor_var = "bill_length_mm")

# roadmap
roadmap <- roadmap(conf_data = penguins_complete,
                   start_data = starting_data,
                   visit_sequence = visit_sequence)

# synth_spec
penguins_rec <- construct_recipes(conf_data = penguins_complete,
                                  other_vars = c("species", "island", "sex"),
                                  synth_order = visit_sequence$visit_sequence)

lm_mod <- parsnip::linear_reg() %>%
  parsnip::set_engine("lm")

synth_spec <- synth_spec(roadmap = roadmap,
                         synth_algorithms = lm_mod,
                         recipes = penguins_rec,
                         predict_methods = sample_lm)

# noise
# don't add noise to predictions
noise <- noise(roadmap = roadmap,
               add_noise = FALSE,
               exclusions = 0)

# constraints
# don't impose constraints
constraints <- constraints(roadmap = roadmap,
                           constraints = NULL,
                           max_z = 0)

# replicates
replicates <- replicates(replicates = 1,
                         workers = 1,
                         summary_function = NULL)

# create a presynth object
presynth1 <- presynth(roadmap = roadmap,
                      synth_spec = synth_spec,
                      noise = noise,
                      constraints = constraints,
                      replicates = replicates)

# synthesize!
set.seed(1)
synthesize(presynth1)
```

The Future

1. Additional non-filer data releases
2. Synthetic IRS PUF
3. `library(tidysynthesis)`
4. Validation server



URBAN INSTITUTE · ELEVATE · THE · DEBATE

Contact Me

cbowen@urban.org

www.clairemckaybowen.com

/in/bowenclaire

@ClaireMKBowen