URBAN INSTITUTE · ELEVATE · THE · DEBATE

SYNTHETIC INDIVIDUAL INCOME TAX DATA

Leonard Burman
Urban Institute/Tax Policy Center
Syracuse University (Emeritus)

2022 FTA Revenue Estimation and Tax Research Conference

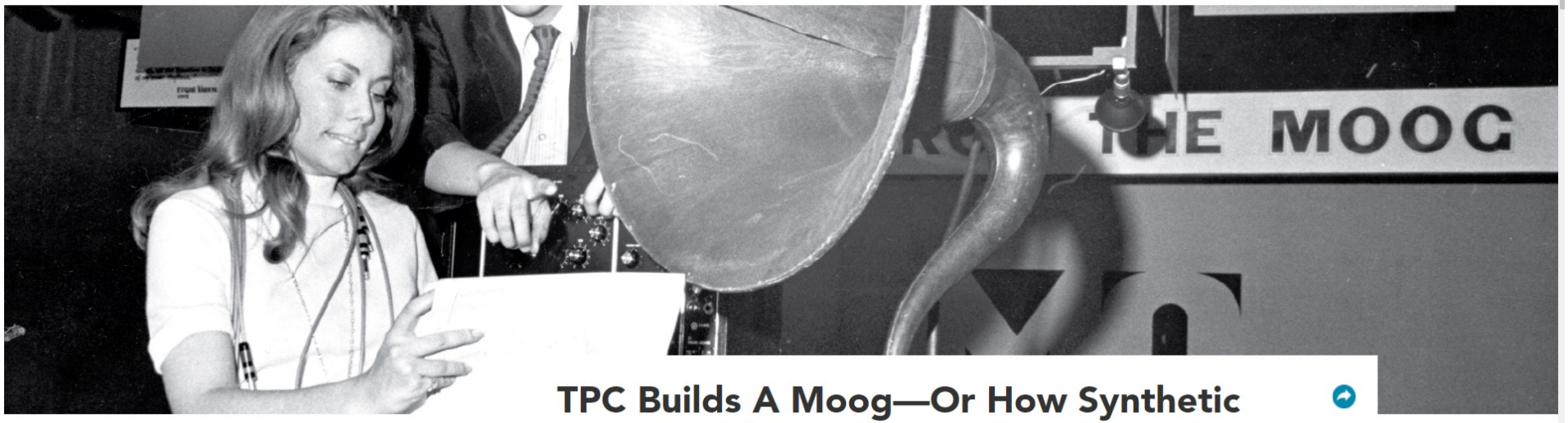
October 25, 2022

Project Team

- Andres F. Barrientos – Assistant Professor, Florida State University
- Claire Bowen – Lead Data Scientist, Urban Institute
- Victoria Bryant – Senior Economist, Statistics of Income, IRS
- Len Burman – Institute Fellow, Tax Policy Center, Urban Institute
- John Czajka - Senior Fellow, Mathematica Policy Research
- Surachai Khitatrakun – Senior Research Methodologist, Tax Policy Center, Urban Institute
- Graham MacDonald – Chief Data Scientist, Urban Institute
- Rob McClelland – Senior Fellow , Tax Policy Center, Urban Institute
- Sybil Mendonca – Associate Director of Data and Analysis, Urban Institute
- Josh Miller – Lead Drupal Developer, Urban Institute
- Maddie Pickens – Data Scientist, Urban Institute
- Livia Mucciolo – Research Assistant, Tax Policy Center, Urban Institute
- Clayton Seraphin - Web Programmer, Urban Institute
- Joshua Snoke – Statistician, RAND Corporation
- Deena Tamaroff – Senior UX Design Specialist, Urban Institute
- Silke Taylor – Senior Software Engineer, Urban Institute
- Erika Tyagi – Senior Data Engineer, Urban Institute
- Aaron R. Williams – Data Scientist, Income and Benefits Policy Center, Urban Institute
- Doug Wissoker – Senior Fellow, Statistical Methods Group, Urban Institute

TaxVox

The voices of Tax Policy Center's researchers and staff



TPC Builds A Moog—Or How Synthetic Data Could Transform Policy Research

And provide a safe way for data stewards to share sensitive data with researchers and the public

URBAN INSTITUTE

Synopsis

- With SOI division of IRS, we are working to create a high-quality fully synthetic Public Use File (PUF) as a safe replacement for traditional PUF plus validation server that researchers could use to safely access confidential data
- Research supported by foundations:
  - Alfred P. Sloan Foundation
  - Arnold Ventures
  - NSF/NCSES
- Usual disclaimers apply

What are synthetic data?

- Fake data, designed to mimic confidential data
- Data points are drawn at random from multivariate empirical distribution
- Two types of synthetic data: full and partial

Why synthetic tax data?

- Protecting confidentiality in public datasets has never been more challenging.
- Emerging literature on privacy reveals threats are greater than previously understood.
- Fully synthetic data are a way to safely expand access.

Validation Server

- Synthetic data could be useful for many purposes (such as running a tax model), but may not produce reliable estimates for complex statistical models
- Validation server allows the execution of statistical programs developed and debugged on synthetic data to run on the confidential data with noise added to estimates to preserve privacy
  - Methodology generates statistically valid estimates with robust measurable privacy guarantee
- Synthetic data plus validation server will allow wider research access to tax data with more robust privacy guarantee and lighter demands on IRS staff

Current status of project

- We have created a fully synthetic nonfiler database based on information returns.
  - The first database of its kind ever released publicly.
- We have produced a preliminary 2012 synthetic public use file (PUF) for testing and evaluation (not public).
- We are starting to work on a 2013 synthetic PUF and plan to continue to test and refine our methodology on later years.
  - If all goes well, SOI will release the 2016 PUF publicly.
- We have developed a prototype validation server.

Tax Data

Administrative tax datasets

- Government researchers and select outside scholars have used tax data for research
  - SOI Joint Statistical Research Program
- Tax data are useful in many fields (not just public economics)
- Access limited by privacy laws (IRC §6103) and IRS resource constraints

The Public Use File

- PUF is useful for tax model simulations of current policies and proposed alternatives
- Synthetic datasets could include information suppressed on PUF
  - E.g., include state of residence, more detail on business income, AMT
- Could also produce other PUFs
  - E.g., nonfiler database, already released

Threats to administrative data releases

- Massive amounts of personal data and computing power raise the risk of matching those data with tax return info
- Identity disclosure
- Attribute disclosure
- Inferential disclosure

Existing Protection

- SOI takes many steps to protect confidential data, but those measures distort the data in ways that may undermine its research value
- Current protections may not be robust to future threats

Synthetic Data

Synthetic data

- Goal is to simulate the statistical process that produces the administrative data
- Potential for very good synthetic file with no disclosure risk

Random Rapture Theorem ⟹ Good Synthetic Data Exist

- Suppose 1 in 1,000 people are raptured, along with their history on earth.
- But their income tax returns remain.
- The random rapture sample is an ideal synthetic data set.
- The only problem: how to find it.

Fully synthetic data

- In a fully synthetic dataset, all data are synthesized, in steps.
  - If there are *k* variables, $Y_1, \ldots, Y_k$, create synthetic $\widehat{Y_1}$ drawn from empirical distribution of Y_1; $\widehat{Y_2}$ conditional on $\widehat{Y_1}$ and empirical distribution of Y_2; and so on until $\widehat{Y_k}$ is synthesized based on $\widehat{Y_1}, \ldots, \widehat{Y_{k-1}}$
- We use non-parametric CART models to sequentially synthesize each variable based on previously synthesized variables as predictors.

Classification and Regression Trees (CART)

- Non-parametric model by Breiman, Friedman, Olshen, and Stone (1984)
- Used to synthesize data in Reiter (2005)
- Good for data that don't fit common distributions
- May capture complex nonlinear relationships between variables

Goals for synthetic data quality

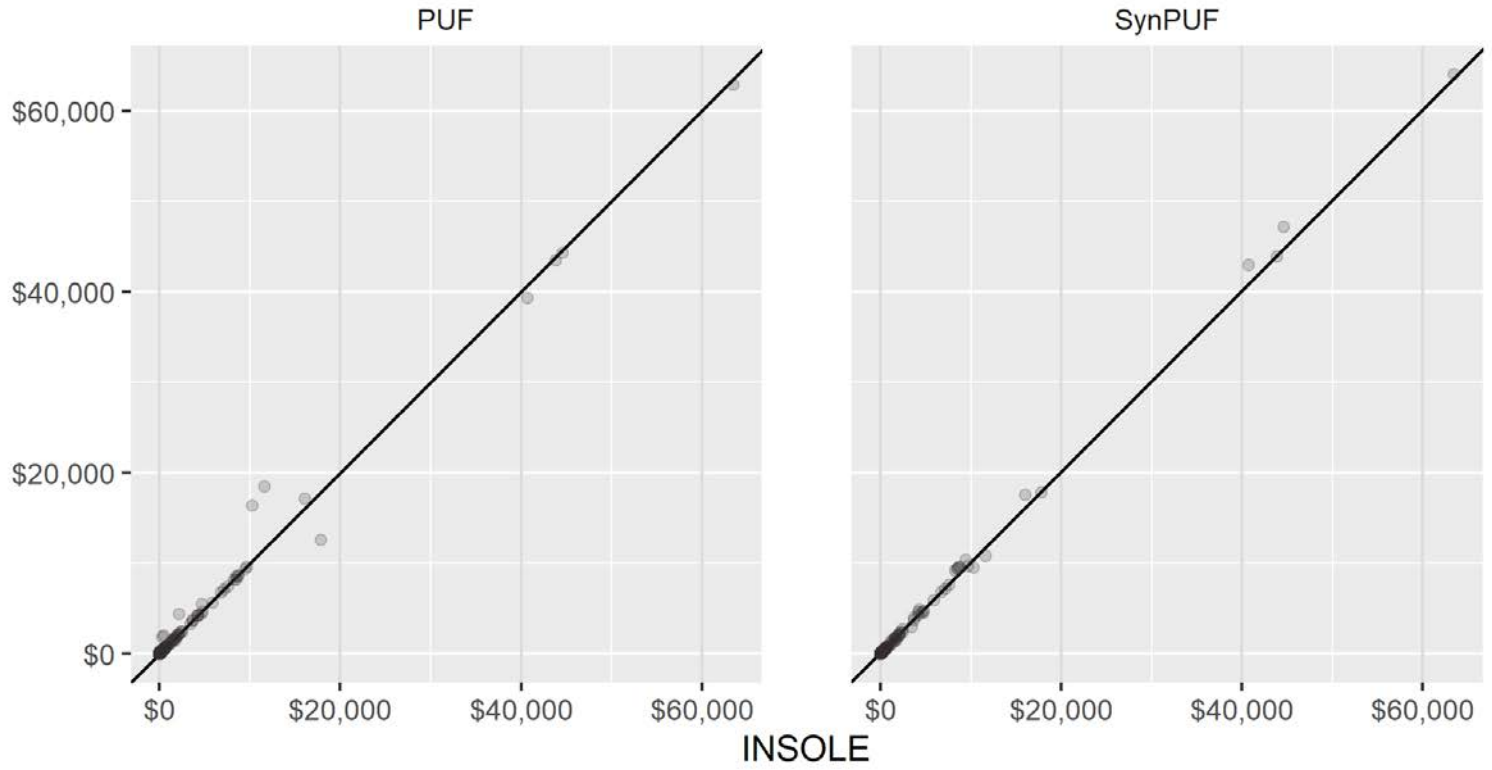
- General Utility
  - Distribution of synthetic data is close to the distribution of the original data
- Specific Utility
  - Results of *an analysis* from the synthetic data are similar to those using the original data

Measures of Quality

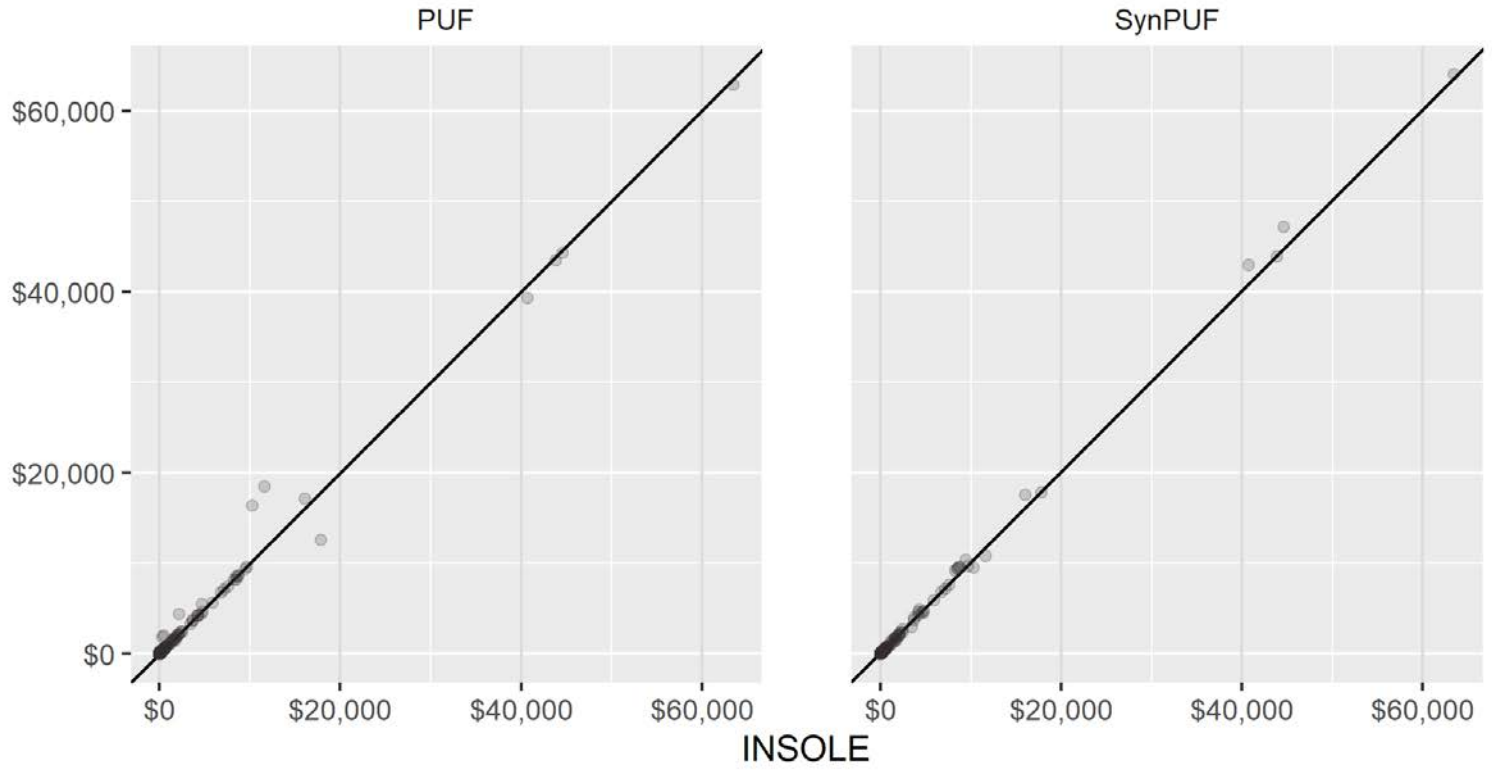
Current Results

- By some metrics, the quality is very high
  - Pairwise correlations on SynPUF closely match underlying data
  - Weighted means for most variables closely match data
  - Overall distribution of AGI is close
- By others, improvements can be made
  - Standard deviations for some variables don't match
  - Too much income at the top 1 percent of AGI

Means in SynPUF and PUF versus INSOLE, 2012

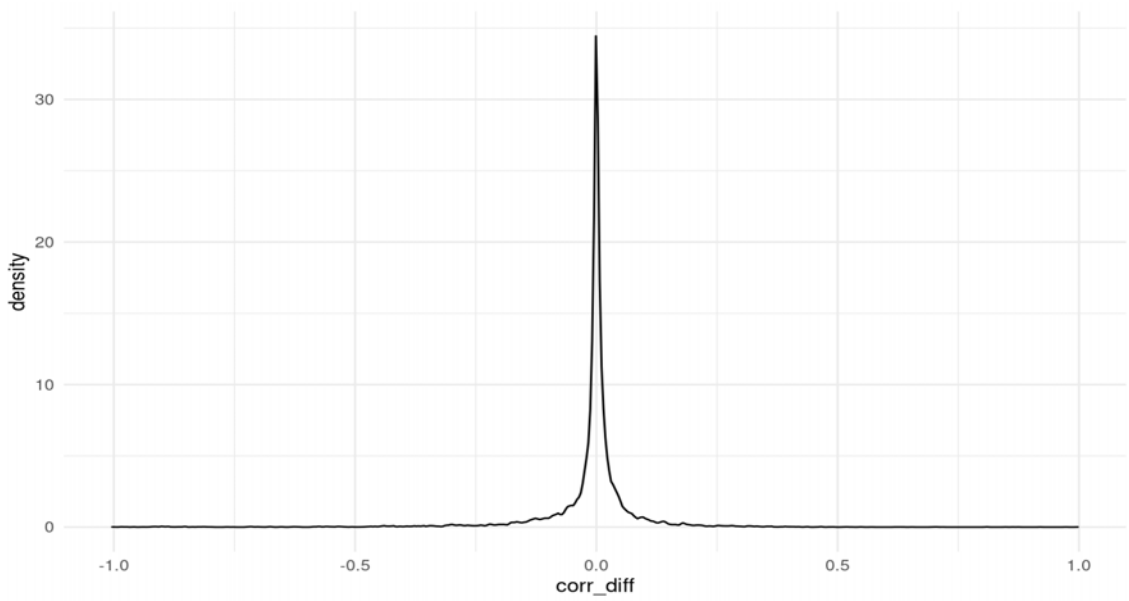


Standard Deviations in SynPUF and PUF versus INSOLE, 2012



Most Correlations are Replicated in the SynPUF

Difference between confidential and synthetic correlation coefficients

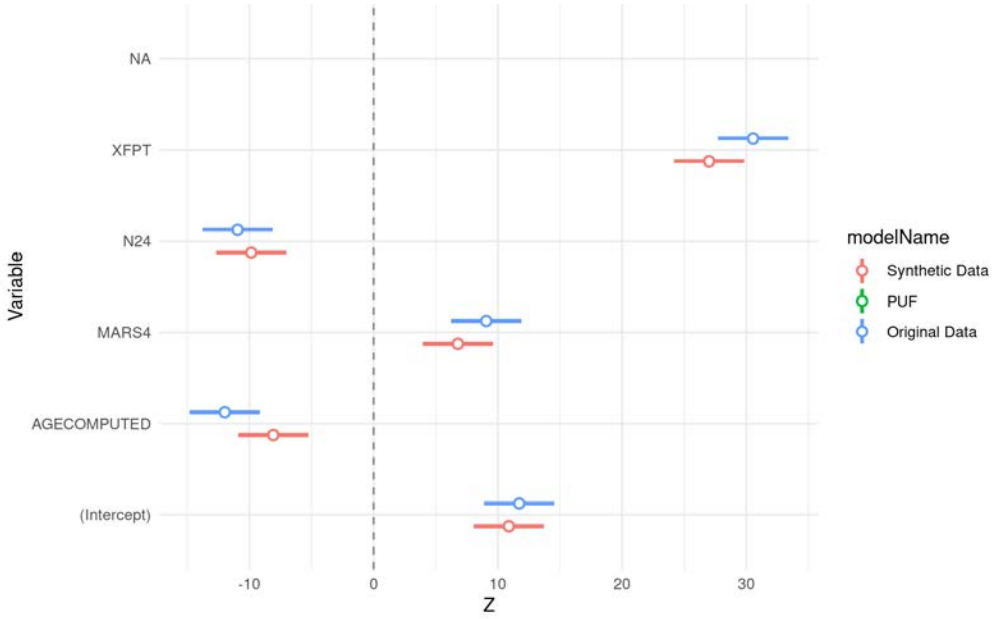


synth0077

Some Multivariate Relationships are Replicated in the Synthetic Data

Regression Fit Test (weighted)

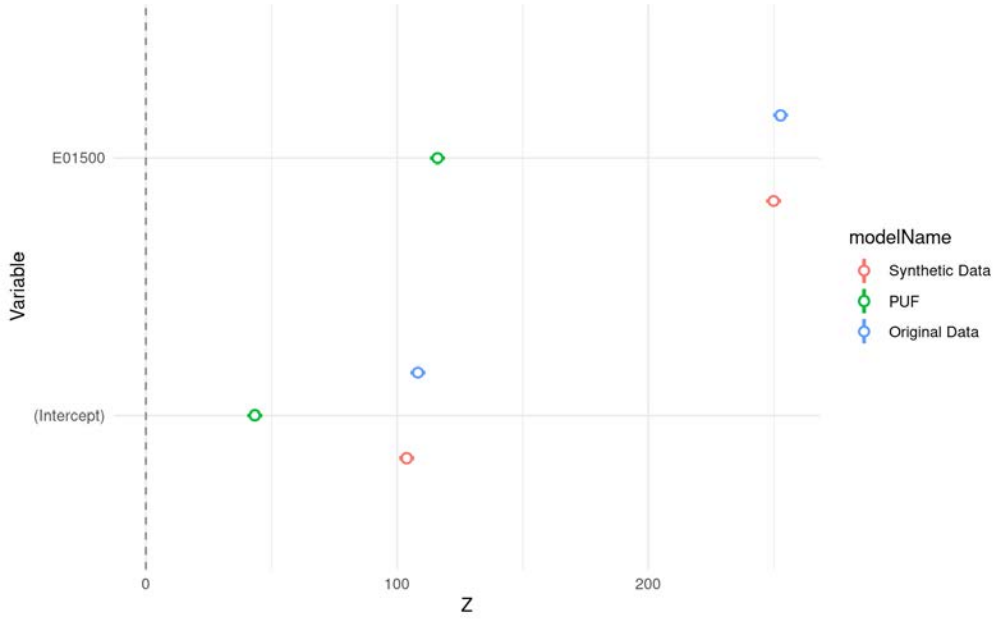
- Salaries and Wages regressed against age, filing status, number of children eligible for child tax credit, and primary taxpayer exemption (i.e., not a dependent)



synth0077

In some cases, synthetic data are better than the PUF

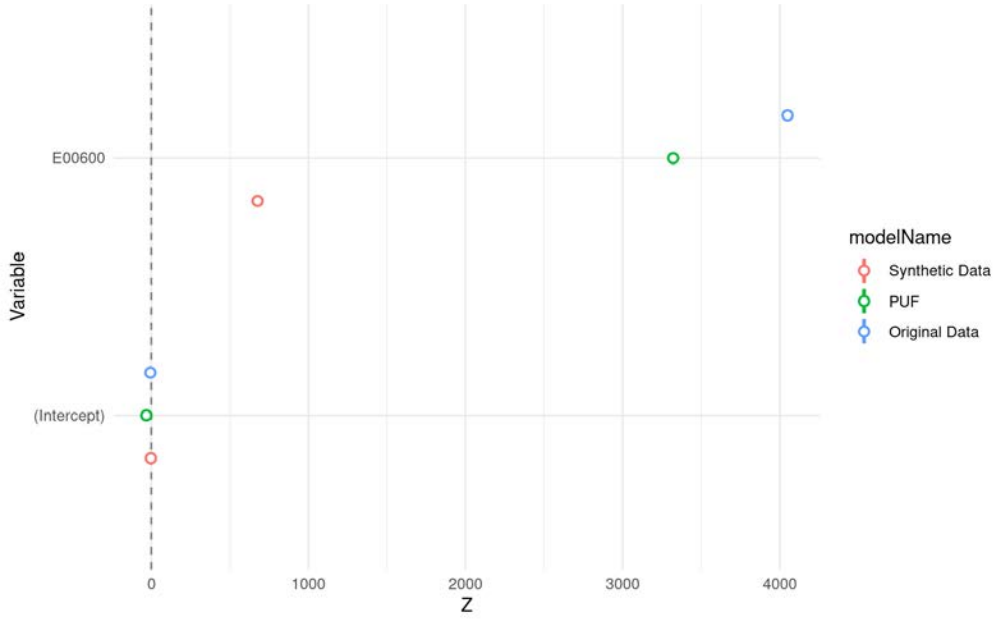
Pensions and annuities included in AGI as share of total pensions and annuities received (weighted regression)



synth0077

In some cases, synthetic data are worse than the PUF

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



synth0077

The Validation Server

Why a validation server?

- Synthetic data might be useful for tax modeling, but may not provide reliable answers to particular kinds of questions or accurate estimates for complex statistical models (e.g. regression discontinuity) or for analysis of small subpopulations.
- Automate traditional SDC process for researchers, enabling more research using sensitive data.
  - Avoid potentially lengthy clearance process
  - Enforces consistency in privacy protection without spending valuable senior staff time for review

What is a validation server?

A system that can:

- Accept submitted research programs
- Automatically calculate and return privacy-preserving results
- Provide information about and enforce the “privacy budget” of released results for each researcher and across all users
- Educate the researcher about the privacy budget and its tradeoffs, and empower the researcher to manage their privacy budget
- **NOTE:** synthetic data are an essential complement to the validation server, allowing testing and debugging code before submitting to the server

Challenges

- In a feasibility study, privacy algorithms performed well for simple statistics but performed poorly on regressions
- Finding algorithms for general use
- Measuring and allocating the privacy budget
- Educating researchers about the privacy budget
- Building a useful, general program interface for researchers
- Ensuring reasonable processing time

Help may be coming for States

Help for States

ACDEB
Advisory Committee on Data for Evidence Building

- ACDEB report just released recommends significant help for state, local, territorial, and tribal governments that want to apply these and other privacy-preserving technologies to their data releases.
  - Block grants to states.
  - NSDS would provide technical support for data stewards and training and guidance for data users.

https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf

For more information, see

https://www.urban.org/projects/safe-data-technologies

Contact: lburman@urban.org