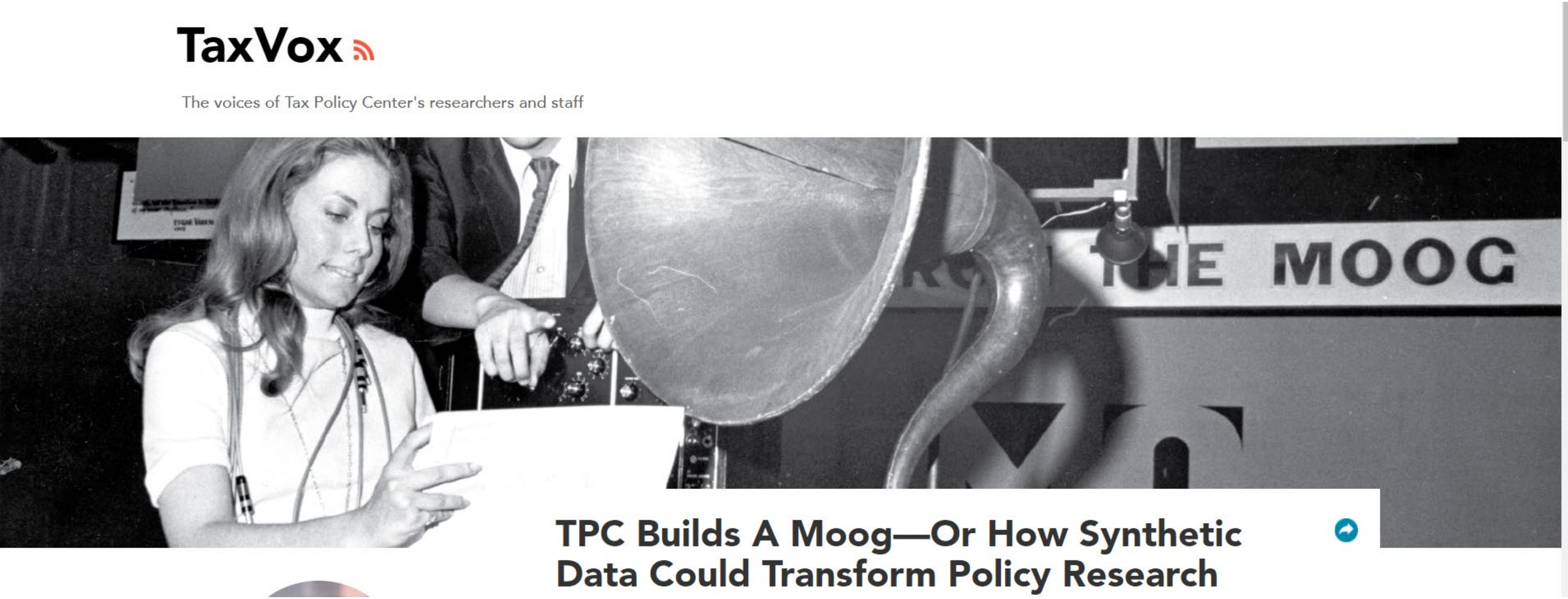
URBAN INSTITUTE · ELEVATE · THE · DEBATE

Synthetic Data and Validation Server: Safely Expanding Research Access to Sensitive Data

Len Burman
Urban Institute/Tax Policy Center and Syracuse University/Center for Policy Research

Advisory Committee on Data for Evidence Building
February 19, 2021



TaxVox

The voices of Tax Policy Center's researchers and staff

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

Project Team

- 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
- Silke Taylor – Senior Software Engineer
- Kyle Ueyama – 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
- Noah Zwiefel – Research Assistant, Tax Policy Center, Urban Institute

Intro

- Synthetic Public Use File (PUF) plus Validation Server
- Collaboration with Statistics of Income (SOI) @ IRS
- Research supported by:
  - Alfred P. Sloan Foundation
  - Arnold Ventures
  - NSF/NCSES
- Usual disclaimers apply

What are synthetic data?

- Fake data, designed to look like confidential data
- Draw data points at random from empirical distribution
- Two types of synthetic data: full and partial
- We are working with SOI/IRS to produce fully synthetic tax datasets

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 protection
  - Risk, however, of an excessively stringent privacy standard
- Synthetic data plus validation server will allow wider research access to tax data with more robust privacy guarantee and lighter demands on IRS staff

Comparison with Synthetic SIPP Beta

- Synthetic SIPP Beta and Validation Server
    - Synthetic SIPP is a partially synthetic file (sensitive variables are imputed)
    - Census runs programs debugged on the synthetic SIPP on the “gold standard” files
    - Statistical results returned to users after manual disclosure review
- Our approach differs in three key ways—designed to provide more rigorous privacy protection and quicker turnaround
    - Synthetic PUF is fully synthetic
    - Validation server will return statistical estimates perturbed to protect privacy
    - Validation server process will be fully automated

Tax Data

IRS data releases

- Annual PUF
- Detailed tables
- Government researchers and select outside scholars have used tax data for research
- Access limited by privacy laws (IRC §6103) and IRS resource constraints

Administrative tax datasets

- SOI has constructed many valuable datasets
  - Include high-quality information useful for research
- PUF is useful for tax model simulations of, for example, presidential candidates’ tax plans
- Synthetic datasets could include information suppressed on PUF

Benefits of tax data for nontax research

- Tax data are useful in many fields (not just public economics)
- Broader research could be done by linking with other datasets, subject to legal constraints
    - Potential to use secure multiparty computing for data under auspices of NSDS

Threats to administrative data releases

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

Fully synthetic data

- In a fully synthetic dataset, all data are synthesized, in steps.
  - If there are *k* variables, Y_1, … , 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 ε_2; and so on until $\widehat{Y_k}$ is synthesized based on $\widehat{Y_1}$, … , $\widehat{Y_{k-1}}$
- 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)
- 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

Disclosure risks

- Identity disclosure
- Attribute disclosure
- Inferential disclosure

Synthetic Supplemental PUF (nonfiler database)

- With SOI, we developed a synthetic version of data for individuals who did not file, and were not obliged to file
- Based on information returns for tax year 2012
- Ultimate sample size ~ 26,000, 19 variables
- The SSPUF closely matched the distribution of the underlying data and presented negligible disclosure risk
- See references on last slide for details

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

- Defining an appropriate privacy standard
- Developing and implementing privacy protections consistent with that standard
- Measuring and allocating the privacy budget
- Educating researchers about the privacy budget
- Building a useful, general program interface for researchers
- Ensuring reasonable processing time

For more information, see

Claire Bowen, Leonard E. Burman, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Philip Stallworth, Kyle Ueyama, Aaron R. Williams and Noah Zwiefel. 2020. “A Synthetic Supplemental Public-Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications.” Tax Policy Center. https://www.taxpolicycenter.org/publications/synthetic-supplemental-public-use-file-low-income-information-return-data-methodology

“TPC Builds A Moog—Or How Synthetic Data Could Transform Policy Research.” https://www.taxpolicycenter.org/taxvox/tpc-builds-moog-or-how-synthetic-data-could-transform-policy-research

Contact: lburman@urban.org