TPC TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Protecting Privacy and Expanding Access in a Modern Administrative Tax Data System

Leonard Burman, Barry Johnson, Victoria Bryant,
Graham MacDonald, Robert McClelland

NTA Spring Symposium

May 10, 2024

Project Team

Nikki Airi

Conrado Arroyo

Andres Barrientos

Claire Bowen

Victoria Bryant

Len Burman

John Czajka

Derek Gutierrez

Barry Johnson

Surachai Khitatrakun

Graham MacDonald

Rob McClelland

Josh Miller

Gabe Morrison

Maddie Pickens

Chris Rexrode

Clayton Seraphin

Joshua Snoke

Deena Tamaroff

Silke Taylor

Erika Tyagi

Giang Trinh

Aaron Williams

Doug Wissoker

Overview

- Old system: tables, PUF, access to confidential data for very small number of researchers
  - PUF is a stratified sample of anonymized tax returns with various measures taken to protect against disclosure
    - *Increasing amount of suppression to protect confidentiality has made PUF less useful over time*
  - Researchers could access confidential data through SOI's JSRP
    - *Requires labor-intensive manual review of all output before release*
- New system: tables, synthetic PUF (starting TY16), validation server, JSRP
  - Systematic approach that can expand research access while strengthening privacy

Tiered access to tax data

Tier	Access	To Whom
1	Tabular data and reports	Anybody – via website and published reports
2	Synthetic individual income tax return data	Anybody who needs it – upon request to SOI
3	Validation server: Automated system allows researchers to access confidential tax return information in an environment that protects against disclosure	Researchers vetted by SOI with a research plan that could not be completed using tier 1 or tier 2 access.
4	Access to confidential microdata	Researchers approved for access through the Joint Statistical Research Program.

Streamlined application process

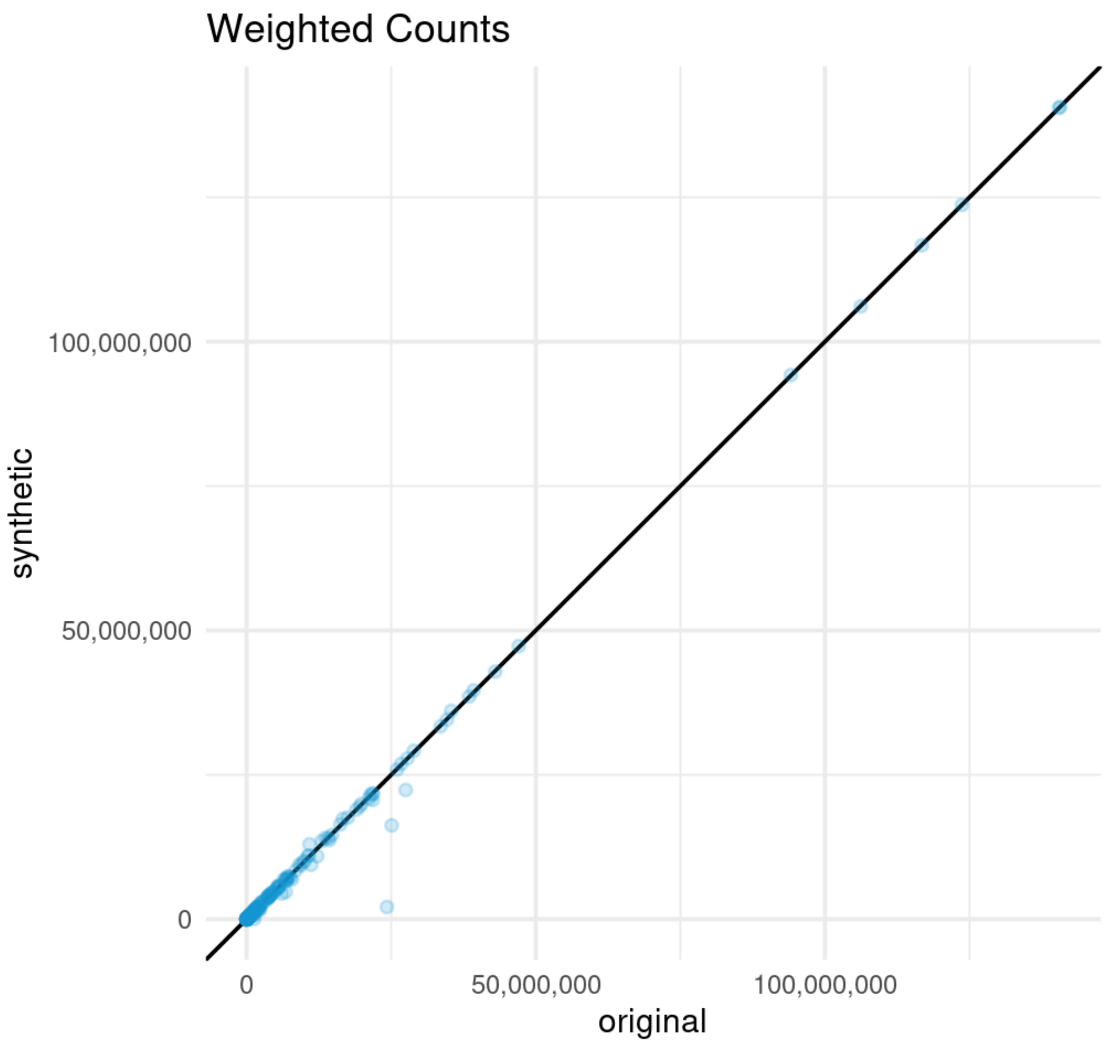
Synthetic Tax Data

Creating synthetic tax data

- Synthetic data drawn from an empirical joint distribution function
- CART model: nonparametric machine-learning tool that can capture highly irregular distributions like tax data
    - Trained on actual tax data, but output is completely synthetic
    - Each variable synthesized based on variables synthesized before
    - Noise added—more in sparse parts of the distribution
- UI team developed new tools—*tidysynthesis* and *syntheval*—to capture the idiosyncrasies of tax data and make other improvements to *synthpop*
- Quality of synthetic PUF comparable to traditional PUF and better in some ways

Synthesis Quality

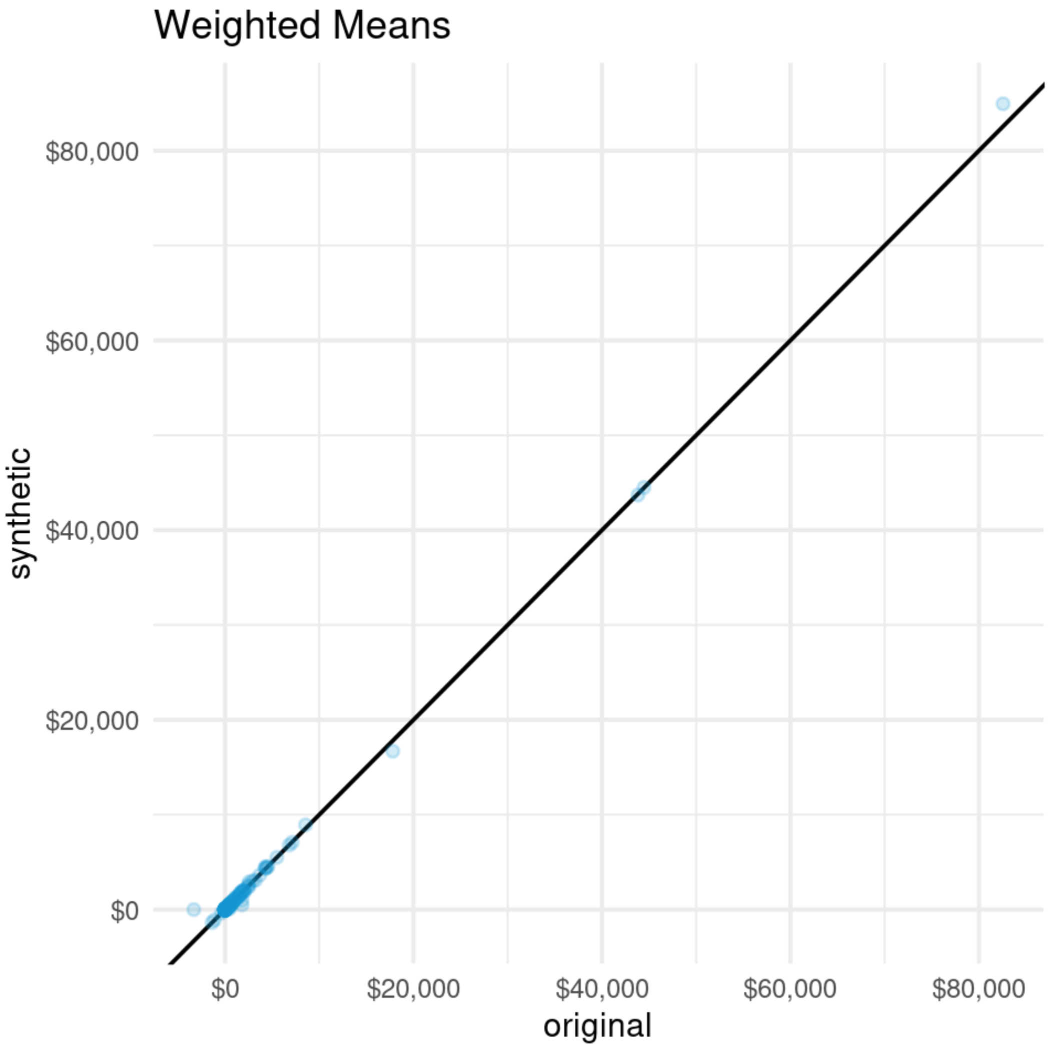
The synthetic data match the weighted counts of tax variables (tax year 2012)



Not for public release.

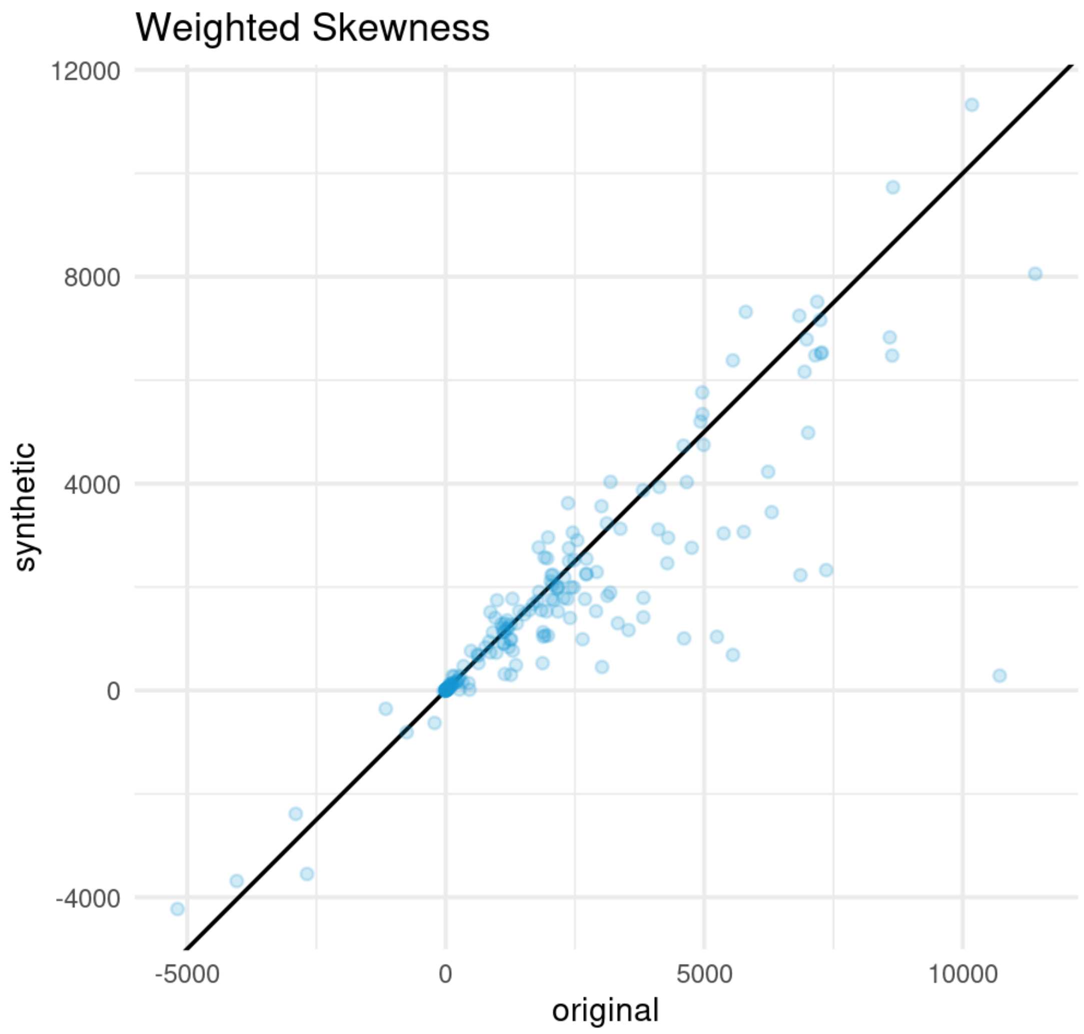
The synthetic data closely match the means

- The synthetic data also match most weighted percentiles (not shown)



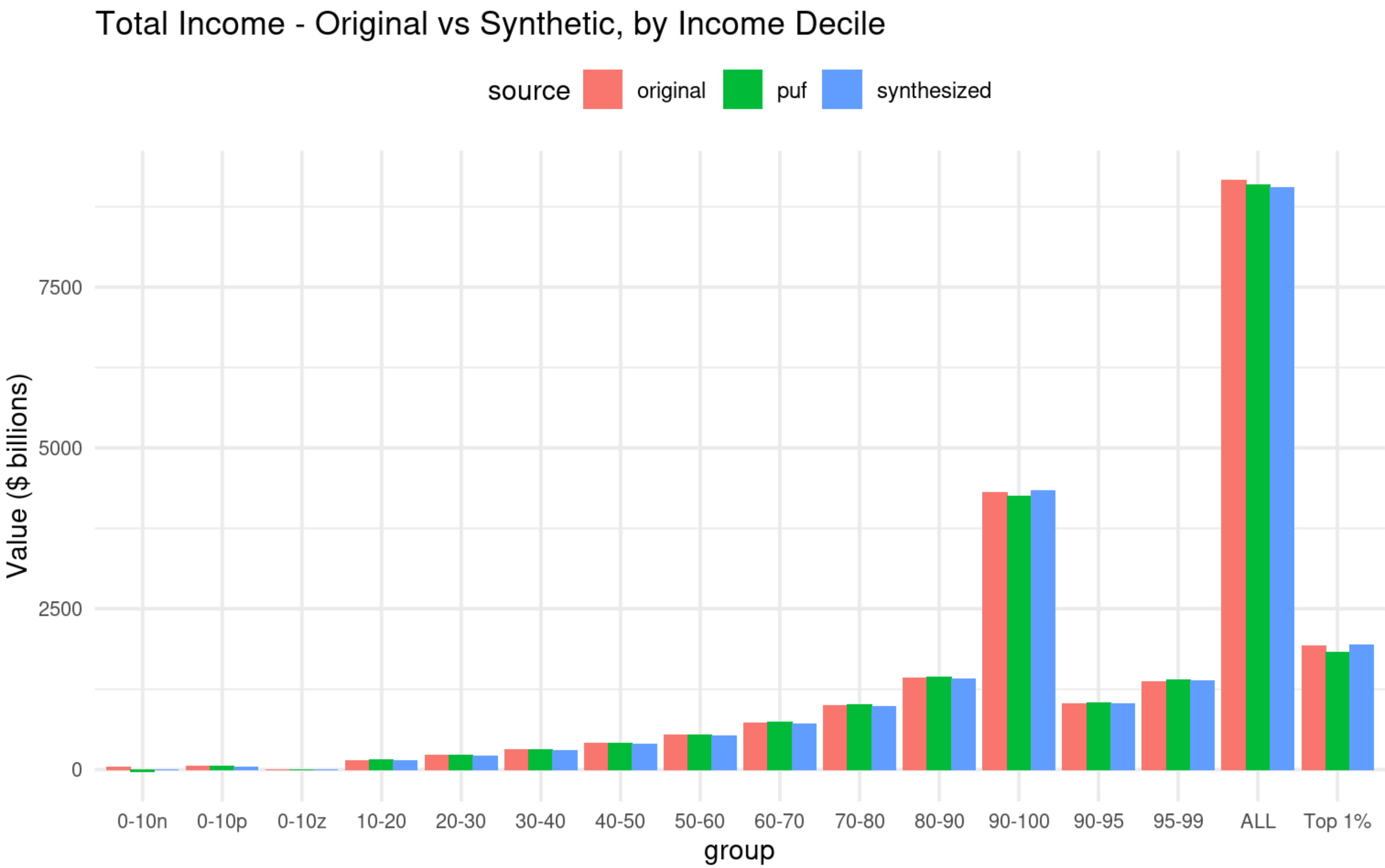
Not for public release.

The synthetic data even capture the skewness of many tax variables



Distribution of Total Income by AGI Group (ty 2012)

Traditional PUF v Synthetic



Not for public release.

Synthesis summary

- Overall, quality of synthetic PUF is pretty good, and comparable to traditional PUF
- Synthetic data protect privacy well
- For more information, see:

Bowen, et al. 2022. “Synthetic Individual Income Tax Data: Promises and Challenges.” *National Tax Journal*, *75*(4), 767-790.

Bowen, et al. forthcoming. “Safe Data Technologies: Safely Expanding Access to Administrative Tax Data.” in *Handbook of Sharing Confidential Data: Differential Privacy, Secure Multiparty Computation, and Synthetic Data* (CRC Handbooks of Modern Statistical Methods)

Advantages of synthetic PUF over traditional PUF

- Safe, systematic privacy protection
- Some variables may be more accurately represented
- More data may safely be included in the synthetic PUF
- More timely release
- More synthetic datasets may be produced
- No charge for synthetic PUF

Limitations of synthetic PUF

- Statistics derived from synthetic PUF may be unreliable for some purposes
- Some kinds of data cannot be accurately represented in a PUF of any kind

Plans for release of 2016 synthetic PUF

- First, user testing of 2015 synthetic PUF, scheduled for fall 2024
    - Trusted users may compare synthetic with traditional PUF in microsimulation models and for other purposes
    - 2015 synthetic PUF will *not* be publicly released because parallel traditional and synthetic PUF creates unnecessary privacy risk
- Fully synthetic 2016 PUF will be released after approval by IRS disclosure review board

Validation Server

Validation Server

- Automated validation server is a digital tool that allows a researcher to access confidential data and receive statistically valid estimates without seeing the underlying data
- Users would develop and test statistical programs using synthetic data, and then submit them remotely to run on the confidential data
- The validation server adds random noise with mean zero and variance calibrated to protect privacy before returning results
- The amount of noise depends on the sensitivity of estimates to outliers
- User faces a privacy budget that limits the number of estimates that may be generated and released
- Algorithm based on MOS methodology developed by Chetty and Friedman (a more flexible relaxation of differential privacy)

Current capabilities of internal prototype

- Simple univariate statistics like count, mean, and variance
- Some multivariate statistics such as OLS, logistic regression
- Some machine-learning models

Limitations

- Researcher can’t inspect the confidential data
- Weighted estimates not currently supported
- Only a limited set of statistical models currently supported
- Addition of noise means some relationships between estimates may no longer hold. For example, sum of components will not add up exactly to the total
- Trade off between accuracy and number of statistics released
- Data mining and p-hacking would quickly exhaust privacy budget

Advantages

- No cumbersome security clearance process required (although SOI will need to approve access to the server)
- Eventually, the process can be completely automated, meaning no labor-intensive review of results required by SOI
- Many more researchers will be able to access tax data than under current arrangements

Future plans

- Users apply for access to validation server—and allocation of a privacy budget—through an agency (SOI) or as part of NSDS
    - Bigger budget for testing—including unreleased statistics—than for released statistics
- Beyond estimation, potentially public tax microsimulation models could be designed to run in a validation server.
    - This could include models with underlying datasets that couldn't be accurately represented in synthetic data, such as a corporate income tax model

Challenges

- How to show useful error messages without unnecessary expenditure of privacy budget
- Calculating privacy budget for more complex kinds of estimators
- Speeding up complex analyses in big datasets without compromising privacy
- Improving the privacy algorithm while maintaining a comprehensible interface

Other Issues

Need for user education

- Explaining tiered access and why least restrictive access might not be appropriate
- Aligning research practices with privacy budgets
- Explain that synthetic PUF is not a sample of tax returns; certain statistics may be unreliable
  - Bias is also a growing problem in the traditional PUF
- Researchers will have to carefully test and debug statistical programs before running on validation server to avoid exhausting finite privacy budget.

Limits to formal privacy methods

- Differential privacy is mathematically elegant but based on extreme assumptions and only applies to a small set of statistics
- MOS more realistic and flexible, but computationally intensive
- Impossible to enforce a privacy budget when some users need unrestricted access to confidential data
- Reasons why formal privacy models overestimate privacy risk
  - Complexity of statistical models
  - Cost of attempting to hack a private statistic (models measure only probability some information *could* be inferred, not the cost of doing it)
- Little guidance to data stewards about how to set the privacy budget

Questions

- The tax community already uses simulations (e.g., to model tax legislation), algorithms (e.g., to select returns for audit) and big data sets (e.g., for transfer pricing). Is AI just bigger and faster than what we do today, or it is different?
- How quickly will AI and associated tools change the tax landscape? And how?
- What dangers do you see with respect to AI in the tax world?
- Will AI and tax develop separately within nations, or uniformly across nations?