URBAN INSTITUTE · ELEVATE · THE · DEBATE

Thursday, August 24th, 2024

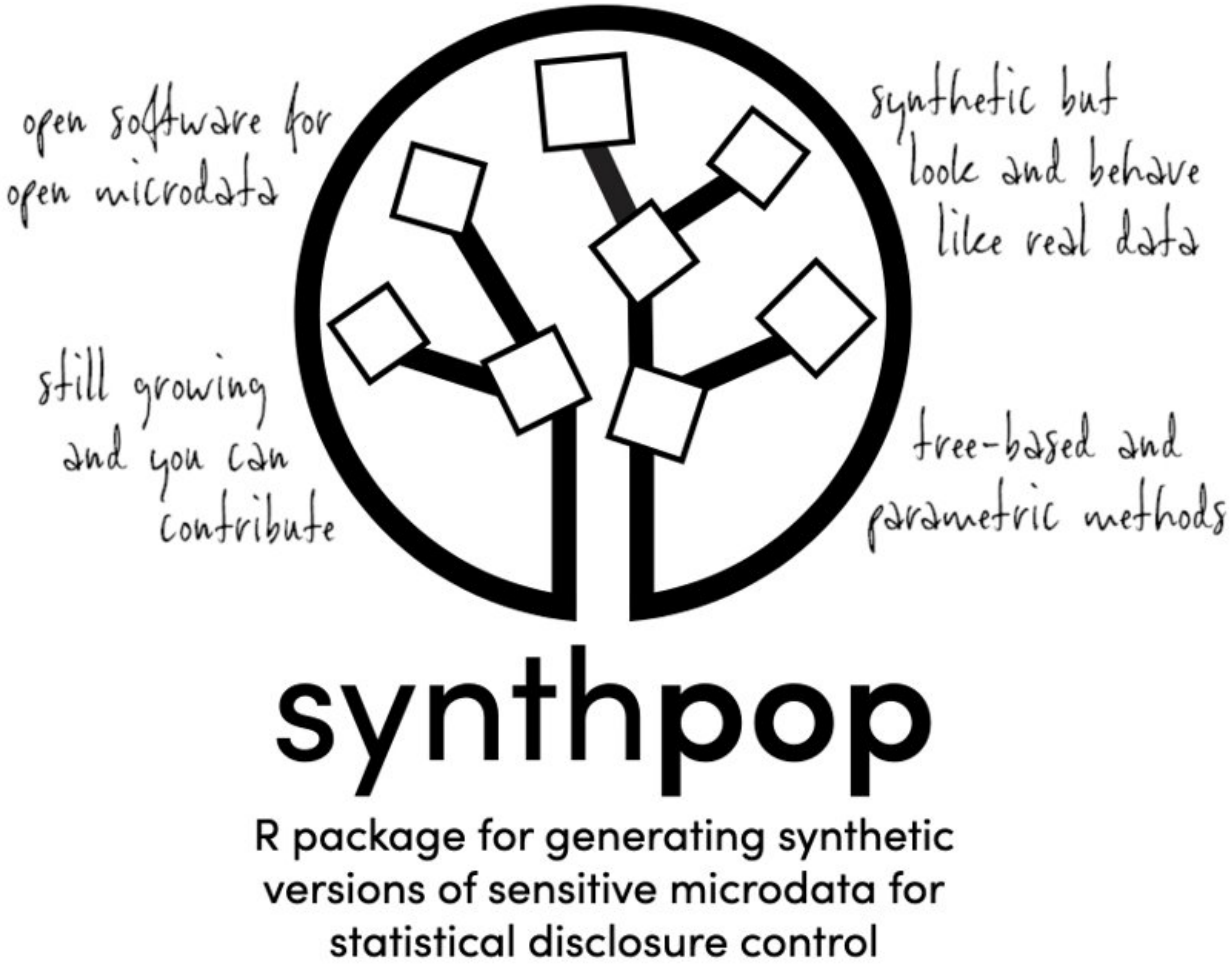
Introducing library(tidysynthesis) and using it to create an IRS synthPUF

Gabriel Morrison

library(tidysynthesis)

Existing Functionality is Groundbreaking, but Limited

- Limited to a small set of methods
- Difficult to extend



R Packages

- **tidysynthesis:** Flexible tools for generating fully and partially synthetic data.
- **syntheval:** Utility and disclosure risk evaluation of synthetic data.

Our Approach

1. Embrace design philosophy from the tidyverse and tidymodels
2. Flexible
3. Modular
4. Extensible

library(tidymodels)

- *All the power of library(tidymodels) for data synthesis, concisely, with a few special tools.*

- Full predictive modeling toolkit

Flexibility

- Express different recipes, predictive models, and samplers for each variable
- Hyperparameter tuning
- Additional noise
- Mid-synthesis constraints
- Synthesize missing data
- Parallel computation with futures/furrr

Modular

- Everything is handled through objects with classes
  - Interchangeable objects that act like building blocks
- Robust testing suite
- Manage computation
  - Lazy evaluation and checks that catch errors before computation

Extensibility

- Ability for someone else to add the thing we haven't thought of

Demonstration

Synthetic data

Confidential data

select	species	island	sex	bill_length_mm	…
TRUE	Adelie	Torgersen	male	39.1	…
FALSE	Adelie	Torgersen	female	39.5	…

Synthetic data

select	species	island	sex	bill_length_mm	…
TRUE					
TRUE					

🎉 Synthetic data! 🎉

Synthetic data

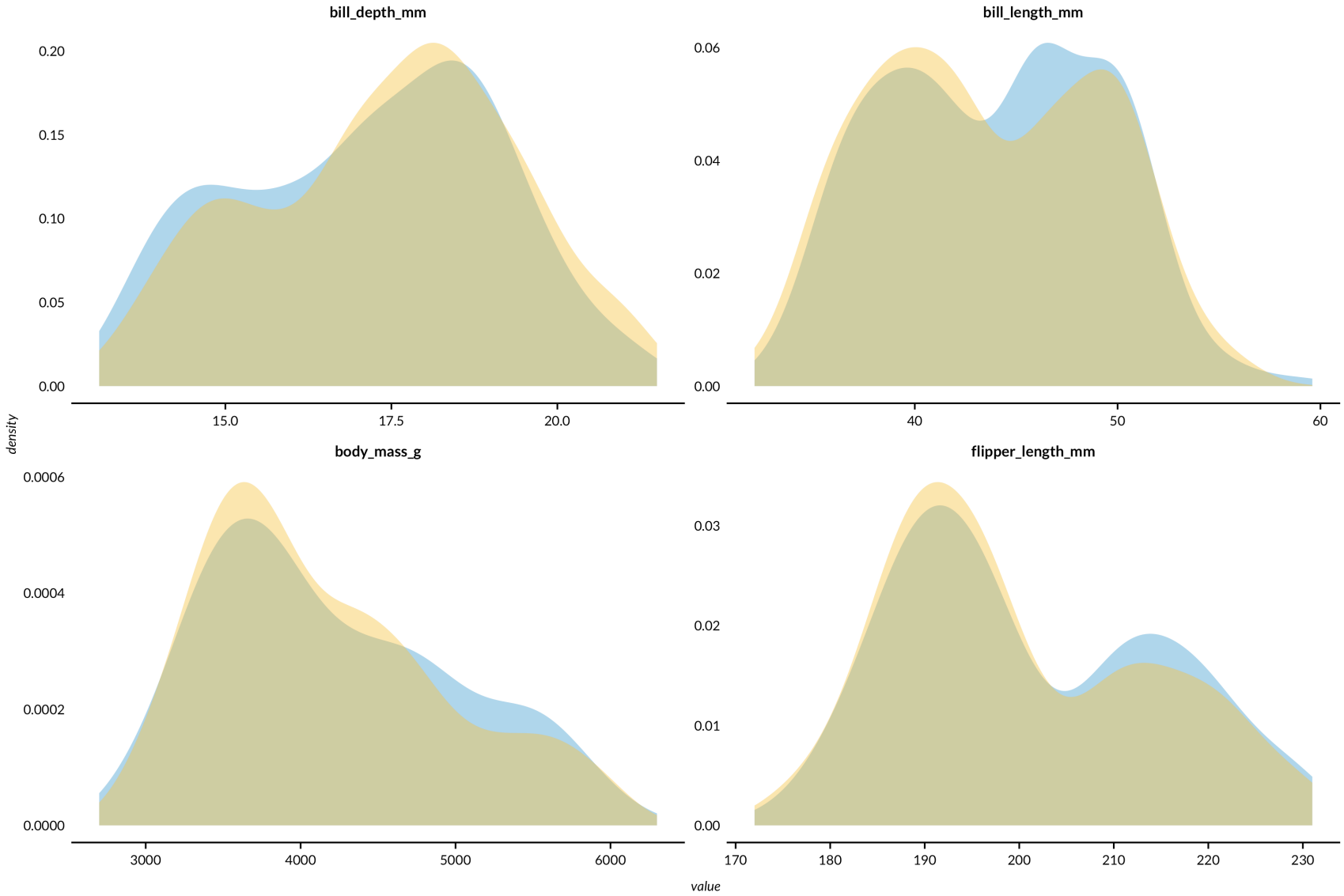
Confidential data

select	species	island	sex	bill_length_mm	…
TRUE	Adelie	Torgersen	male	39.1	…
FALSE	Adelie	Torgersen	female	39.5	…

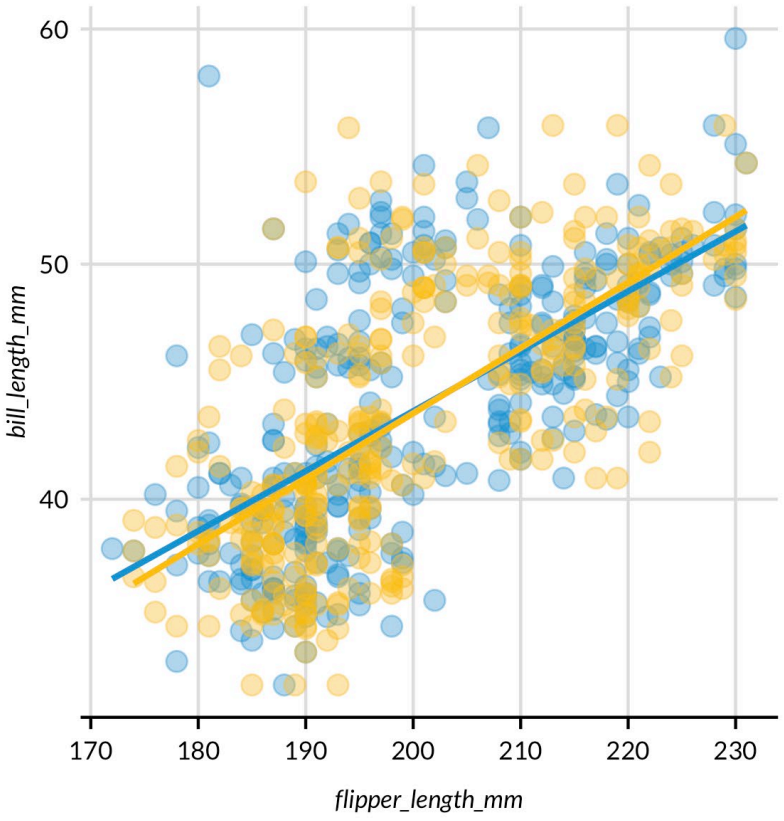
Synthetic data

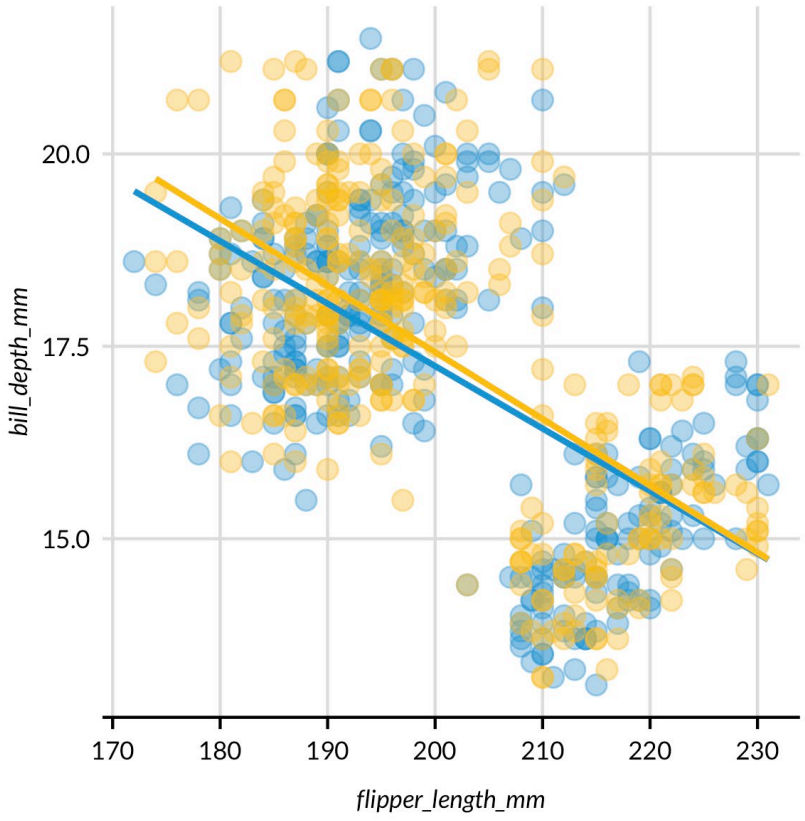
select	species	island	sex	bill_length_mm	…
TRUE	Chinstrap	Dream	male	48.4	…
TRUE	Gentoo	Biscoe	male	51.4	…

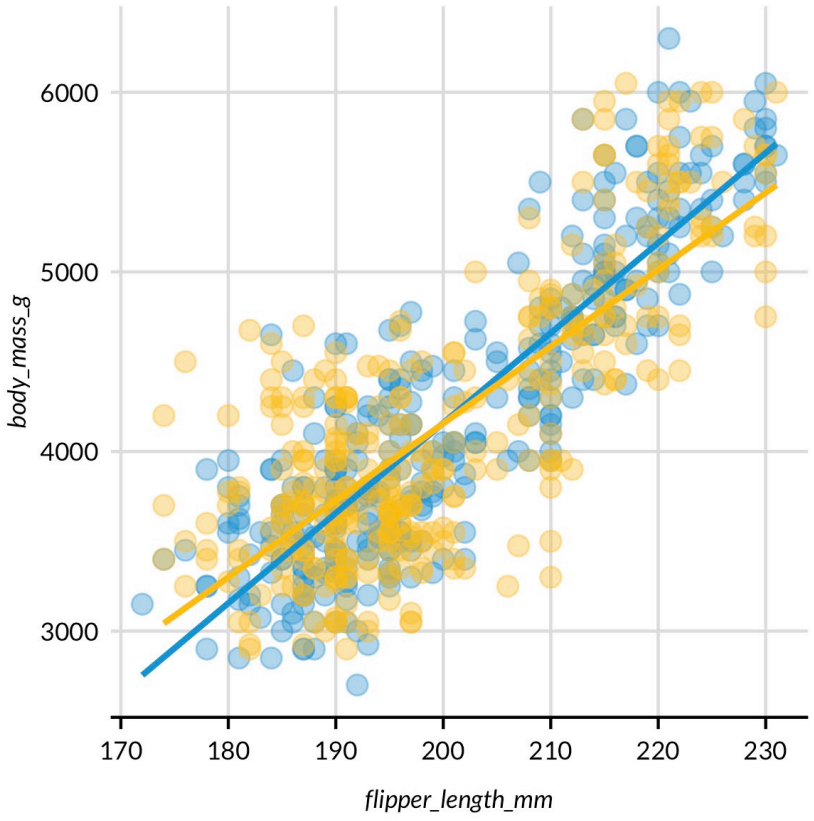
The Synthetic Data are Similar to the Confidential Data



The Synthetic Data are Similar to the Confidential Data







PUF

The PUF is critical for tax policy analysis

- Urban-Brookings Tax Policy Center
- American Enterprise Institute
- National Bureau of Economic Research

Administrative Tax Data

Master File
(145 million records)

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

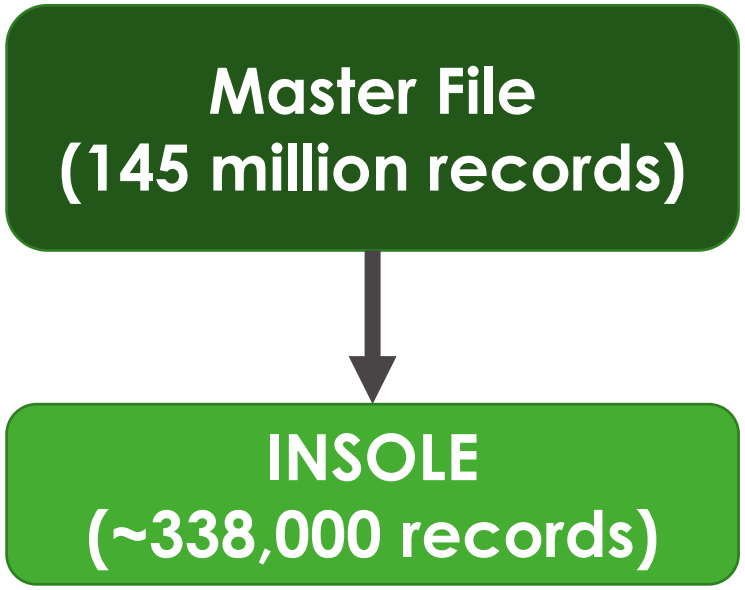
INSOLE: Stratified sample with weights to represent the U.S. taxpayer population.

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.

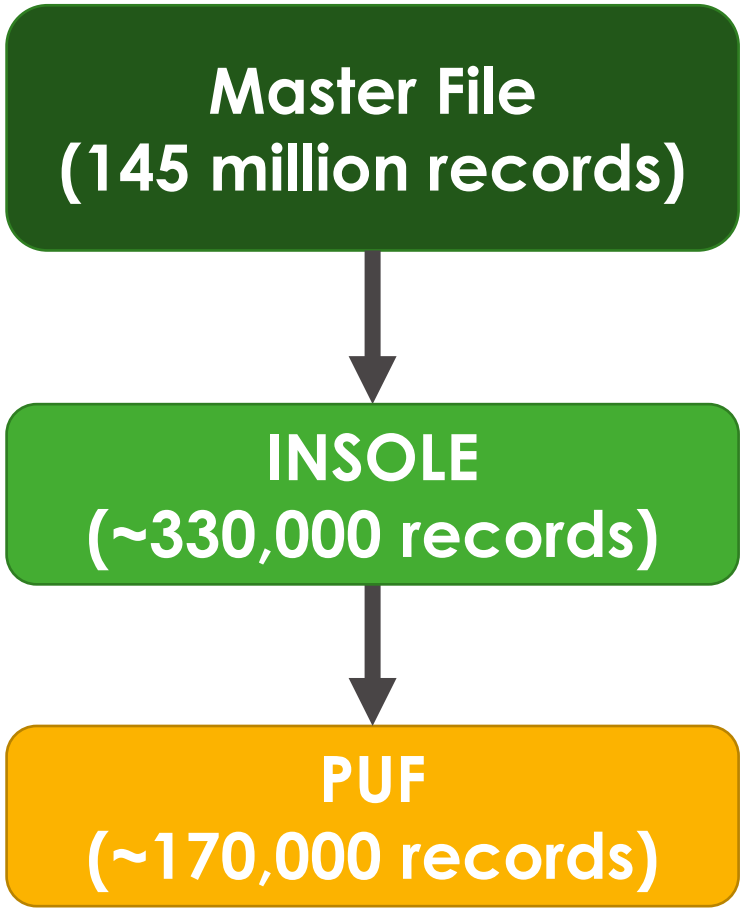


Administrative Tax Data

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

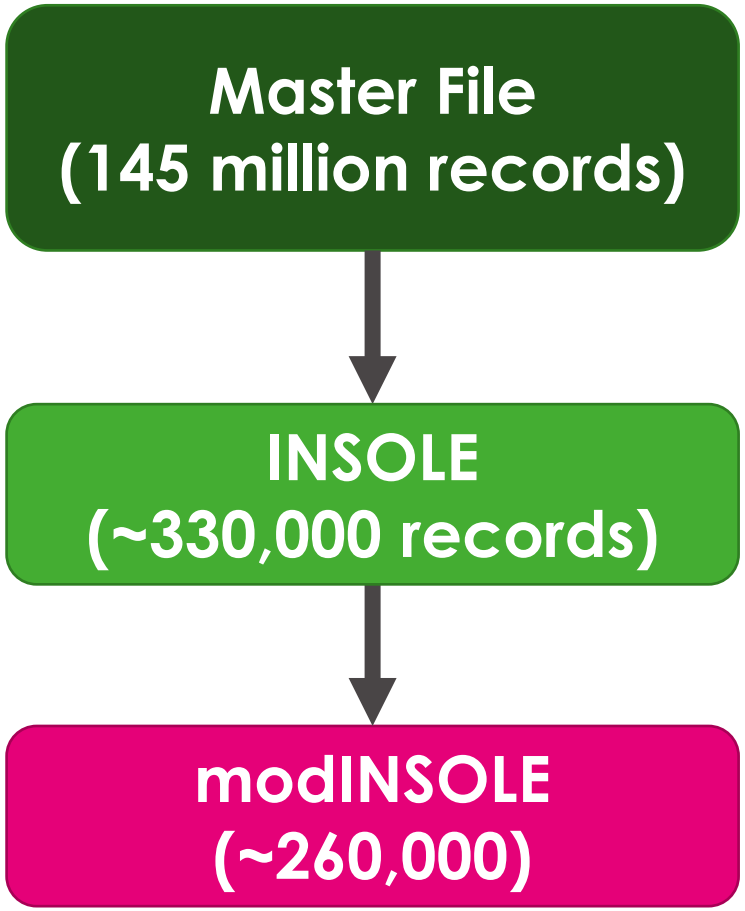
INSOLE: Stratified sample with weights to represent the U.S. taxpayer population.



SynPUF: Synthesis Preparation

modINSOLE: create a base dataset of the 207 variables (INSOLE has over 3,000) that keeps more of the original records than the PUF

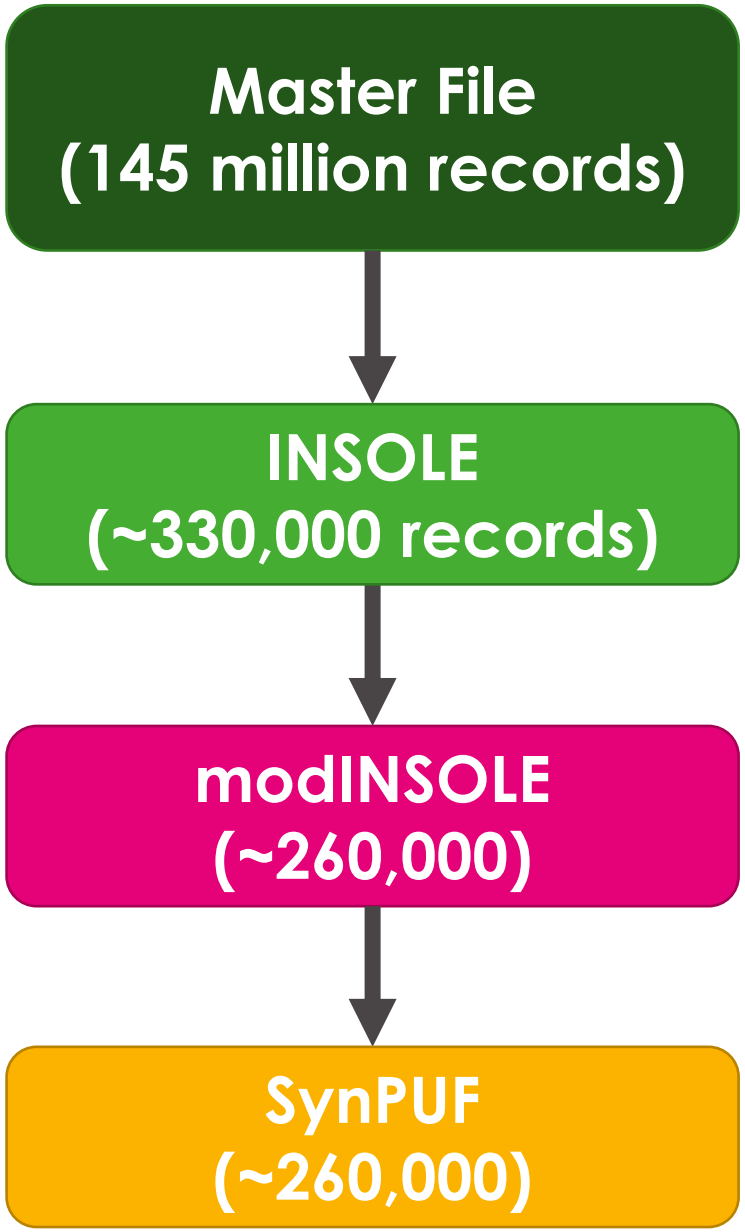
- create new survey weights (98 strata to 25 strata)
- sample within strata



SynPUF: Synthesis Preparation

modINSOLE: create a base dataset of the 207 variables (INSOLE has over 3,000) that keeps more of the original records than the PUF

- create new survey weights (98 strata to 25 strata)
- sample within strata



Using tidysynthesis to create a synthPUF

How tidysynthesis helps Urban and IRS synthesize the PUF

1. Constraints to follow tax policy

How tidysynthesis helps Urban and IRS synthesize the PUF

1. Constraints to follow tax policy
2. Recipes with pre-processing to transform skewed variables

How tidysynthesis helps Urban and IRS synthesize the PUF

1. Constraints to follow tax policy
2. Recipes with pre-processing to transform skewed variables
3. Noise infusion to preserve privacy

How tidysynthesis helps Urban and IRS synthesize the PUF

1. Constraints to follow tax policy
2. Recipes with pre-processing to transform skewed variables
3. Noise infusion to preserve privacy
4. Flexibly specify different models for different variables

Privacy considerations

Procedures to protect confidentiality

1. Sampling before synthesis
2. Fully synthetic data
3. Node heterogeneity
4. IRS rounding rules

Not for public release.

Privacy Tests

- Duplicates
- Unique-uniques
- L-diversity
- Identity disclosure
- Attribute disclosure

Release schedule

Release schedule

- **library(syntheval):** Released and ready to use!
- **library(tidysynthesis):** Spring 2025
- **2013 synthPUF:** October 31 internally, never public
- **2015 synthPUF:** March 31, 2025 to approved external testers
- **2016 synthPUF:** TBD

Safe Data Technologies Project Landing Page

Contact Me

gmorrison@urban.org

Estimating the multivariate distribution of the data

- Goal is to approximate the empirical multivariate distribution function for the data
- Joint multivariate probability distribution can be represented as the product of sequential, conditional probability distributions:

$$f(\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_k \mid \theta_1, \theta_2, \ldots, \theta_k) =$$

$$f_1(Y_1 \mid \theta_1) \cdot f_2(Y_2 \mid Y_1, \theta_2) \cdots f_{\mathrm{k}}(Y_k \mid Y_1, Y_2, \ldots, Y_{k-1}, \theta_k)$$

  - where Y_i the variables and θ_i are vectors of model parameters