Generating Synthetic Microdata

SDSS 2022

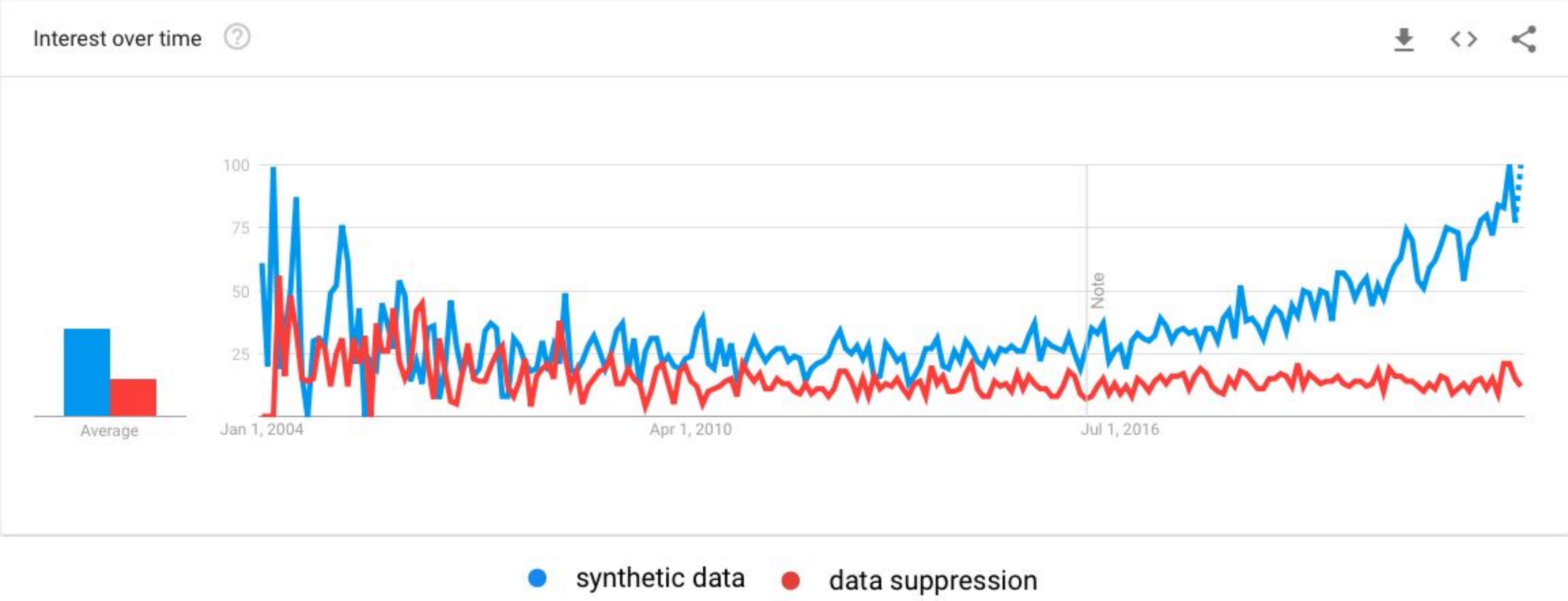
Statistical Data Privacy Techniques for Sharing Sensitive Data

Short Course: Part 3

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
BEYOND BIG DATA: INFLUENCING SCIENCE, TECHNOLOGY, AND SOCIETY
PITTSBURGH, PA • JUNE 7–10, 2022

RAND CORPORATION

Synthetic Data Has Started to Replace Older Methods



Source: google trends

Assumptions about the Data for This Lesson

- Assume we have microdata on individuals
- Can handle mixture of categorical and continuous data
  - Incl. many categories and skewed variables
- For example, a Census extract:
  - Extensive demographics
  - Geographic information

SYLLS
SYNTHETIC DATA ESTIMATION FOR
UK LONGITUDINAL STUDIES

Synthetic Data Has Roots in Imputation

- Proposed based on multiple imputation
  - Rubin (1993), Little (1993)
- Newer methods have evolved to include approaches less connected to MI
  - Single synthetic data sets common

All or Only Some of the Values Can be Synthesized

- Fully or completely synthetic data
  - All released values are synthesized
  - Can use unreleased values to augment models
- Partially or incompletely synthetic data
  - Only some released variables are synthesized (common)
  - Only some released values are synthesized (less common)

Why use Synthetic Data to Protect Data?

- Synthesized values are drawn from random distributions
- Fully synthetic data breaks the link to real individuals
- Data can be generated with very similar distributions to the confidential data

What are Drawbacks to Synthetic Data?

- The risk model is not easy to define
- Developing appropriate generative models can be difficult
- Models can overfit the confidential data
  - Potential problem for both privacy and inference
- Fundamentally: you get out what you put in

How Do We Measure Risk for Synthetic Data?

- No clear consensus exists on defining risk
  - Though proposals exist
- Generally measured based on attribute closeness or exact replicates
- The influence of outliers can be an issue
- Not covering this in-depth today

Different Types of Synthetic Data Models Exist

- Most approaches fall into one of these four categories
- Examples shown for each category

	Joint	Sequential
Parametric	Multivariate Normal	Sequential GLM
Non-Parametric	GANs	Sequential CART

- Today focusing on sequential models

Joint vs. Sequential Synthetic Data Models

- Joint assume a known multivariate distribution for all variables
  - E.g., multivariate normal
- Sequential take advantage of the law of total probability
  - I.e., $P(X, Y, Z) = P(X)P(Y|X)P(Z|X,Y)$
  - Fit a sequence of models with each variable conditional on those prior

Sequential are Flexible and More Common

- An example can help understand how a sequential model works
  1. Take a bootstrap sample of `age` → `syn_age`
  2. Fit a logistic model: `sex` ~ `age`
     a. Predict new values using `syn_age` → `syn_sex`
  3. Fit a linear model: `income` ~ `age` + `sex`
     a. Predict new values using `syn_age` and `syn_sex` → `syn_income`

Non-Parametric Models Can be Easily Substituted in the Sequence

- In the previous example, each variable was modeled using a GLM
- Instead, use regression trees or other non-parametric models
- Increases flexibility if distributions are not known

Other Important Topics We Not Covering

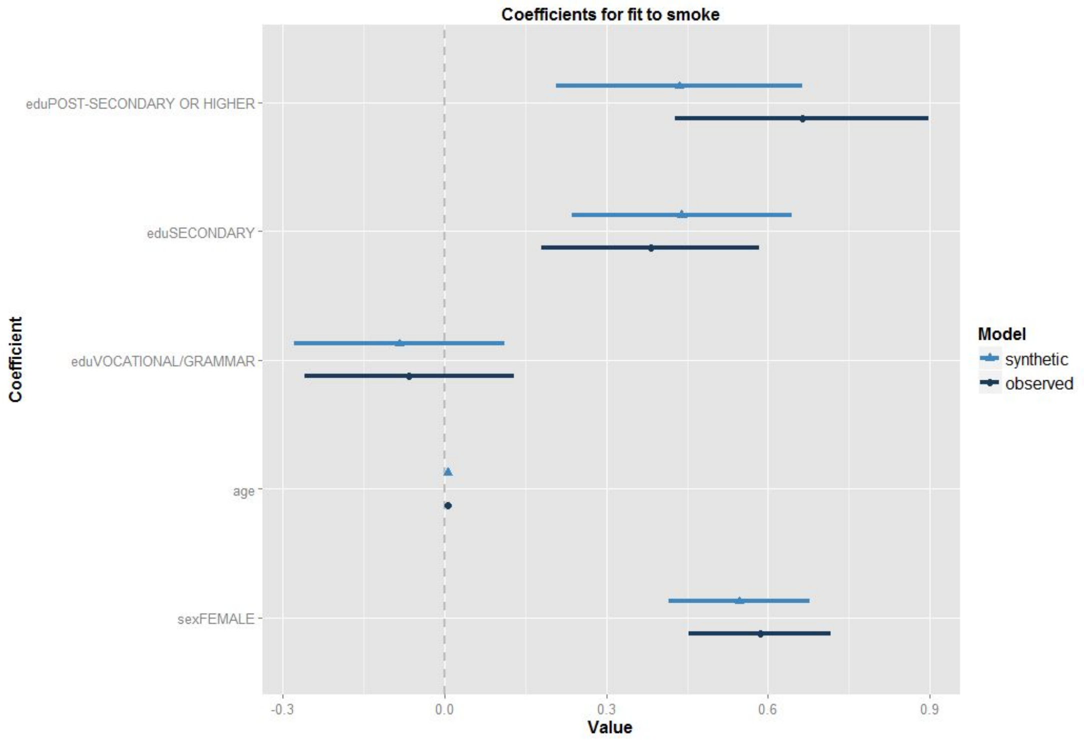
- Approaches to partially synthetic data
  - Data augmentation models
- Joint synthetic data models
- Detailed inference rules for different types of synthesis
  - Combining rules for multiple synthetic data sets

How Can We Measure Utility?

- Cannot measure utility based on how much data are released
  - Unlike suppression
- Instead two common approaches:
  - *Specific utility:* Similarity between statistics estimated on the confidential and synthetic data
  - *General utility:* Distributional closeness

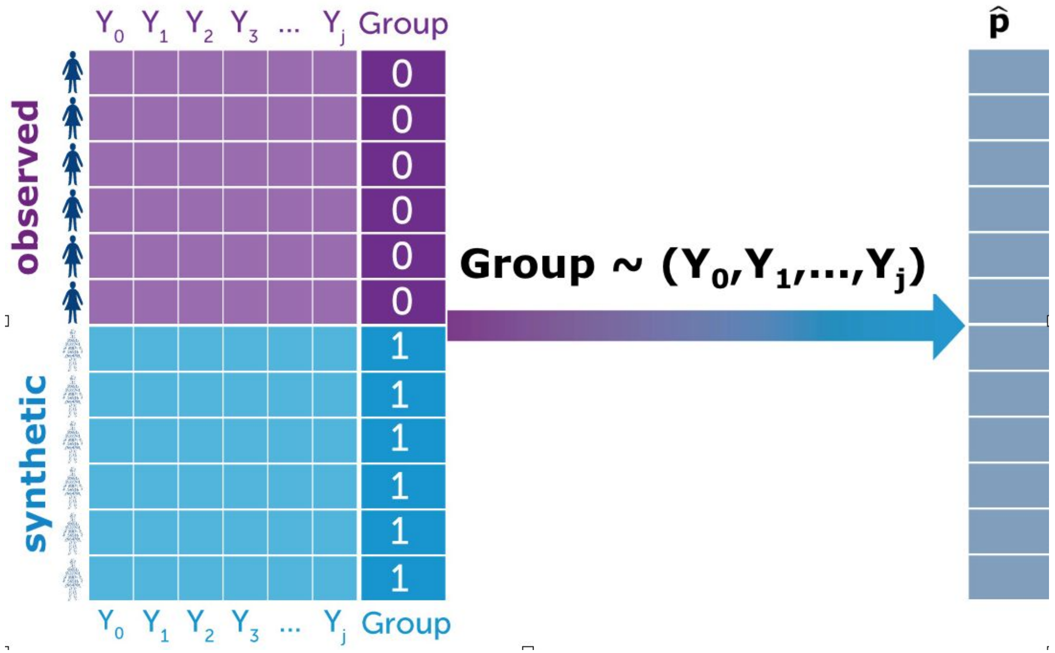
Specific Measures of Utility

- Distance between statistics
  - L1, standardized difference
- CI measures
  - Overlap, ratio
- Inference
  - Coverage of population parameter
- These cannot be generalized about the entire data



General Measures of Utility

- Direct distance measures
  - K-L Divergence
  - Empirical CDF
- Discriminant approaches
  - Probability of group membership
  - Requires good baseline
  - Relates to GANs
- Generalizable but does not inform about specific results



Takeaways from This Lesson

- Synthetic data offers a flexible means of releasing data with similar distributions to the confidential data
- Highly “tuneable”
- Risk is hard to define in this context
- Utility depends on the goals of the data

Further Reading

- Raghunathan, Trivellore E., Jerome P. Reiter, and Donald B. Rubin. "Multiple imputation for statistical disclosure limitation." *Journal of official statistics* 19, no. 1 (2003): 1.
- Drechsler, Jörg. *Synthetic datasets for statistical disclosure control: theory and implementation*. Vol. 201. Springer Science & Business Media, 2011.
- Reiter, Jerome P. "Using CART to generate partially synthetic public use microdata." *Journal of official statistics* 21, no. 3 (2005): 441.
- Raab, Gillian M., Beata Nowok, and Chris Dibben. "Practical data synthesis for large samples." *Journal of Privacy and Confidentiality* 7, no. 3 (2016): 67-97.
- Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. "General and specific utility measures for synthetic data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181, no. 3 (2018): 663-688.
- Nowok, Beata, Gillian M. Raab, and Chris Dibben. "synthpop: Bespoke creation of synthetic data in R." *Journal of statistical software* 74 (2016): 1-26.