Understanding Synthetic Data

Using pseudo-records to maintain privacy in publicly released data

Madeline Pickens, Jennifer Andre, and Gabe Morrison

Synthetic data replace actual records in a confidential dataset with **statistically representative pseudo-records.** Synthetic data enable data curators to release data that would otherwise be too sensitive for public release.

Synthetic data are typically generated from **probability distributions** or **models** that are identified as being representative of the confidential data. **Synthetic data can be postprocessed** to account for real-world constraints.

The quality of synthetic data can be evaluated by **comparing distributions** of the original and synthetic datasets and by measuring the suitability of the synthetic data for a **specific analysis**.

Stakeholder input is crucial for data curators to **understand the potential applications** of the synthetic data, which in turn informs decisions about **what constitutes sufficient quality and privacy protections.**

Synthetic data replace actual records in a dataset with statistically representative pseudo-records. The goal of most data synthesis is to closely mimic the underlying distributional and statistical properties of the original, confidential data.

WHY SYNTHETIC DATA?

Researchers, service providers, and other stakeholders benefit from access to individual-level data safeguarded by governments or organizations. However, the public release of more granular (disaggregated) data could expose the people represented in that data to privacy violations. This risk has been exacerbated by increased computing power, the availability of auxiliary datasets and information, and the development of powerful new statistical methods. Data curators (individuals responsible for the safekeeping of an organization's data) must navigate these increased risks when determining which datasets or statistics to release publicly and how to obscure private information before these releases. Data synthesis is a statistical technique that allows data curators to release record-level data; this benefits stakeholders who might not otherwise have access to the confidential data while maintaining privacy protections.

GENERATING SYNTHETIC DATA

Synthetic data are typically generated from probability distributions or models identified as being representative of the confidential data. Once values are generated, additional noise can be added to enhance privacy, and constraints can be applied to ensure the new values are realistic in the context of the dataset. Often, multiple versions, or implicates, of the synthetic dataset are generated so data curators can release the dataset version that best balances utility and privacy.

Synthetic data can be partially or fully synthetic. Partially synthetic data synthesize only some columns of a dataset (generally the most sensitive columns from a privacy perspective), retaining a one-to-one mapping between the original and synthetic product. Fully synthetic data, in contrast, synthesize all values in the original dataset and do not necessarily maintain a one-to-one mapping. Fully synthetic data provide stronger privacy protections than partially synthetic data, but preserving dataset properties can be more difficult in the full synthesis process.

TRUSTING SYNTHETIC DATA

Evaluating Synthetic Data Quality

Data curators can evaluate how well synthetic data capture the properties of the underlying confidential data using two main types of metrics:

- General (sometimes called global) utility metrics measure distributional similarity between the original and synthetic data. Some examples of these metrics include comparisons of summary statistics, correlation fit between variables, and discriminant-based metrics, which measure how difficult it is to distinguish between original and synthetic observations.
- Specific utility metrics measure the suitability of a dataset for a specific analysis. These vary by dataset but could include measurements of confidence interval overlaps for regression coefficients or microsimulation results.

Ensuring Privacy in Synthetic Data Releases

Data curators should evaluate synthetic datasets for risk of identity disclosure (i.e., the ability to associate a known individual with a synthetic record) and attribute disclosure (i.e., the ability to determine some new characteristic of an individual based on the information in the released data).

Synthetic Data and Stakeholder Input

Stakeholder input is crucial throughout the synthesis process so data curators can understand

- ideal uses of the released synthetic data;
- the measure of synthetic data quality (general and specific) that must be maximized in the synthesis; and
- the "acceptable" level of disclosure risk and "acceptable" loss of data quality.

Each of these elements can vary by dataset and use case and will have substantial impact on the decisions made throughout the synthesis process. The more feedback stakeholders can provide, the more the final synthetic product can enable applications of the data that might otherwise be impossible without access to the confidential data.

ABOUT THE AUTHORS

Madeline Pickens
mpickens@urban.org

Madeline Pickens is a data scientist in the Office of Technology and Data Science at the Urban Institute. Her research focuses on applications of data science methodology in data privacy.

Jennifer Andre
jandre@urban.org

Jennifer Andre is a data scientist in the Center on Labor, Human Services, and Population. Her research focuses primarily on financial well-being.

Gabe Morrison
gmorrison@urban.org

Gabe Morrison is a data scientist in the Office of Technology and Data Science. His research focuses on visualization and analysis of spatial data.

ADDITIONAL READING

Personal Privacy and the Public Good: Balancing Data Privacy and Data Utility
Claire Bowen **https://urbn.is/3krfGeQ**

A Synthetic Supplemental Public-Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications
Claire Bowen, Victoria L. Bryant, Leonard E. Burman, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Philip Stallworth, Kyle Ueyama, Aaron R. Williams, Noah Zwiefel **https://urbn.is/2ZS4RIx**

This fact sheet is a product of the Urban Institute's Racial Equity Analytics Lab, which operates with the generous support of the Ballmer Group, the Bill & Melinda Gates Foundation, the Salesforce Foundation, and Urban's general support donors. Lead funding for this fact sheet was provided by the Bill & Melinda Gates Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Further information on the Urban Institute's funding principles is available at **urban.org/fundingprinciples**.