FULLY SYNTHETIC DATA FOR THE AMERICAN COMMUNITY SURVEY

Michael Freiman, Evan Totty, Aref Dajani

U.S. Census Bureau

To continue providing access to high-quality statistics while protecting respondent confidentiality, the U.S. Census Bureau is researching the feasibility of creating fully synthetic microdata for the American Community Survey (ACS). A parallel line of research is on a service allowing users to validate their results against a lightly perturbed version of the original internal Census Bureau data. Since the data are fully synthetic, a person or household record in the synthetic data should not correspond to a record in the internal data, but we intend for a wide range of statistical estimates about collections of people or households to give similar results on the synthetic data and the internal data.

We begin generating the synthetic data by drawing the geographic locations of the synthetic records from those observed in the original data, with replacement. For subsequent variables, synthesis follows a sequential tree-based method, building on [1]: we construct a decision tree to predict each variable from previously synthesized variables, then draw a synthetic value for each record at random from the appropriate leaf. We construct some summary variables, such as the median household income in a person or household's county, to use as additional tree predictor variables to better preserve relationships among variables. To maintain intra-household relationships, we synthesize some "stub variables" for households and people within them together so that, for example, we can capture the relationship between a householder's age and the age(s) of the householder's spouse or children, or between the number and characteristics of the people in a household and the size of the housing unit. The stub variables are a small set of selected variables whose intra-household relationships seem particularly important to maintain. The rest of each person and household record is synthesized separately because of sparsity concerns. For larger households, a separate matching routine adds additional synthetic person records to the synthetic household record until it reaches the appropriate size. The synthesis uses a modified version of Knexus Research Corporation's (KRC) CenSyn package, and some of the code for handling the stub variables also builds on code by KRC.

Synthetic data with validation is a different tier of access from the existing Public Use Microdata Sample (PUMS), but with some potential advantages. The PUMS includes coarsening, top-coding and perturbation [2], reducing its utility for some use cases, and the extent of these protection measures is likely to increase over time as re-identification threats proliferate. Together, synthetic data and the validation service could afford greater availability and accuracy of survey estimates while maintaining or improving confidentiality.

To assess the accuracy of the synthetic data, we are developing a question bank including thousands of modeled and descriptive statistics. We compare the results of these analyses on the internal data, the current version of the synthetic data, and the PUMS. Of approximately 1,700 modeled statistics in the question bank, almost 300 cannot be produced from the PUMS at all because of coarsening or top-coding. Of the modeled statistics that can be produced with the PUMS, we find that statistical conclusions (statistically significant positive effect, statistically significant negative effect, or no statistical significance at .05) agree on the internal data and the synthetic data 90% of the time, only a small decrease from the 94% agreement between the internal data and the PUMS on the same analyses.[1] While the Census Bureau will encourage data users to use the validation service for final results from any future ACS fully synthetic dataset, this research suggests that the synthetic data are comparable to the PUMS for developing an analysis.

To assess the confidentiality of the synthetic data, we use a variation of the re-identification studies that the Census Bureau has conducted on the ACS and other surveys. We simulate an attack where an intruder with access to the unprotected internal data matches these data to other government or commercial files that include personally identifiable information. We use a crosswalk to confirm which of these households were correctly re-identified. We assess whether the inclusion or exclusion of re-identified households in the training data for synthesis affects the proportion of re-identified records for which at least one synthetic record is similar, both in the state as a whole and in the Public Use Microdata Area (PUMA). In all three states studied

[1] The Census Bureau has reviewed this information product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. 7502696, Disclosure Review Board (DRB) approval number: CBDRB-FY26-0032).

(Alabama, North Dakota, and Pennsylvania), most re-identified records have at least one similar synthetic record in the state, regardless of whether re-identified records are included in the training data. The proportion of re-identified records having a similar synthetic record differs by no more than 1.1 percentage points between the two versions of the data in each state. Most re-identified records do not have a similar synthetic record in the same PUMA, and the proportion having at least one similar synthetic record differs by no more than 2.5 percentage points between the two versions of the data in each state.[2] This shows that in data generated using the current synthesis method, the presence of a record that appears re-identifiable is not strong evidence of a similar record's presence in the internal data. All of this indicates that synthetic data is a viable path forward, supporting the conclusion of a 2024 JASON research study [3].

REFERENCES

[1] Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata." *Journal of Official Statistics, 21*, 441-462.

[2] U.S. Census Bureau. (2025). *Public Use Microdata Sample (PUMS) Accuracy of the Data (2024)*. Retrieved from https://www2.census.gov/programs-surveys/acs/tech_docs/pums/accuracy/2024AccuracyPUMS.pdf.

[3] JASON. (2024). *Synthetic data*. Unpublished report.

[2] The Census Bureau has reviewed this information product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. 7510312, DRB approval number: CBDRB-FY26-0033).