What Are Synthetic Data?

Synthetic data can mean many different things depending upon the way they are used. Sometimes, as in computer programming, the term means data that are completely simulated for testing purposes. Other times, as in statistics, the term means combining data, often from multiple sources, to produce estimates for more granular populations than any one source can support. An example of this usage is the U.S. Census Bureau's Small Area Income and Poverty Estimates. In data confidentiality applications, synthetic data are modeled statistical outputs released in a format that closely resembles the confidential data format. Synthetic data can be disaggregated to the individual- or business-record level, or aggregated into tabular format.

> ***… the term means combining data, often from multiple sources, to produce estimates for more granular populations than any one source can support.***

What decisions about the use of synthetic data in the American Community Survey (ACS) have been made?

The Census Bureau hasn't made any decisions yet about the use of fully synthetic data in the ACS. During this exploratory research phase, we welcome feedback from our data users about whether this tool should be considered as a strategy to help us mitigate the data quality issues we have begun to see from lower response rates to surveys, and to provide more accurate data when survey respondents are not representative of the broader community. We do this while ensuring your responses to our surveys are kept confidential.

We are also invested in ensuring our data users feel confident in the synthetic data if we do determine that this tool is valuable for the ACS. We know that making synthetic data that will satisfy every user case is impossible. That's why we're experimenting with allowing data users to validate the synthetic output against internal data. You can learn more about that process in the presentation available at <https://acsdatacommunity.prb.org/p/conferences>.

How has the Census Bureau researched and used synthetic data in its products?

Of the many public-use datasets and online tools created by the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD), its LEHD Origin-Destination Employment Statistics (LODES), and OnTheMap web application have partially synthetic data on where workers live. In addition, the LEHD's Post-Secondary Employment Outcomes data product uses differential privacy techniques to protect individual confidentiality.

United States® Census Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

@uscensusbureau

How has the Census Bureau used or experimented with synthetic data in the ACS in the past?

Over the last decade, our ACS team has increasingly received feedback about the high margins of error for data on small communities in our five-year estimates. Despite the comparatively large sample size for the ACS in comparison to other surveys, the small number of households randomly selected for interviews may result in a high margin of error. Small sample sizes also can mean our imputation models do not reflect the unique characteristics of respondents in low-level geographies.

Among many other research experiments, we have been looking into how synthetic data might help us improve our small area estimates. The "Using Administrative Data to Improve ACS Small Area Estimates" slide deck available at <https://acsdatacommunity.prb.org/cfs-file/__key/widgetcontainerfiles/3fc3f82483d14ec485ef92e206116d49-s-AAAAAAAAAAAAAAAAAAAAA-page-0conferences/Session-3-_2800_Berman_2900_.pdf> is just one example of how synthetic data can help us provide more accurate information on historically undercounted communities like people who identify as American Indian and Alaskan Native, young children, or people living in poverty.

In addition to using synthetic small area data to improve the accuracy of group quarters data, we use synthetic data in the ACS to protect their confidentiality. You can learn more about both uses in "Changes to ACS Group Quarters Small Area Estimation" at <www.census.gov/programs-surveys/acs/technical-documentation/user-notes/2011-01.html> and the "Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau" at <www.census.gov/content/dam/Census/library/working-papers/2019/adrm/5%20Legacy%20Techniques(tagged)%20CED-DA%20Report%20Series.pdf>.

We've used similar models to give researchers access to more granular data on poverty through our Survey of Income and Program Participation (SIPP) Synthetic Beta (SSB) product available at <www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html>. However, more research is necessary for us to understand whether these strategies would be valuable for the ACS.

What research is the Census Bureau currently exploring concerning synthetic data use on the ACS?

Synthetic data methods are well-known to the statistical community as they have been used in other surveys, such as the SIPP, and within statistical software packages (e.g., synthpop within R). The Census Bureau is researching a new fully synthetic data product to explore whether this method would allow us to produce more accurate data—correcting for known sources of error and potentially allowing for more tabulations at lower levels of geography—for our users while maintaining our respondents' privacy. We began conversing with data users about this work in 2019 to solicit feedback and engagement on what we expect will be a multiyear process to explore and research alternatives for the ACS' future in an era of declining survey response rates and increasing reliance on data and statistics that reflect an increasingly diverse country. We will continue to provide public updates in multiple forums as well as in blog posts available at <www.census.gov/newsroom/blogs/research-matters/2020/08/acs-disclosure-avoidance-and-release-plans.html>.

How are synthetic data and privacy/confidentiality connected?

All Census Bureau surveys have to balance the competing requirements of releasing statistics and protecting privacy. Modern privacy theory makes clear that retaining accuracy and privacy in our statistical products requires a trade-off. While sampling in surveys may increase privacy, the interaction between formal privacy methods and surveys is an active research area. Synthetic data are one tool we can use to continue to provide granular data at low levels of geography without sacrificing the privacy of our respondents.