Protecting the Confidentiality of the 2020 Census Redistricting Data

Stronger Protections for the Digital Age

Throughout our history, the U.S. Census Bureau has enhanced and strengthened how we protect the confidentiality of the information that we collect when individuals respond to surveys and censuses.

By law, we must ensure that we don't release information that could identify your information in the statistics we publish, like your name, address, sex, age, race, ethnicity, household composition, or other information provided by you, or on your behalf.

But today's computers have become so powerful that it is easier than ever to take the information we do release and reconstruct it or pair it with information gathered from other sources, like commercial data brokers who collect and sell information based on things like your purchases and financial transactions.

Once those datasets are matched, it's possible to identify people and their information within a set of data—or within a city block, small town, or rural area.

When we simulated this kind of attack on 2010 Census data, we found that we could reconstruct nearly the whole 2010 dataset, and that 52 million people could be identified that way, and thus their individual data could be disclosed. And 52 million is a best-case scenario, based on using only a small number of outside data sources from the time of the 2010 Census.

A worst-case scenario puts the number much higher, at 179 million people re-identified, because the quantity and quality of these outside data sources is stronger a decade later.

The Census Bureau can no longer rely on the protections used in 2010 if we are to meet our obligations to protect respondent confidentiality under 13 U.S. Code §§ 8(b) & 9. Protecting against new technology-enabled re-identification attacks, while maintaining the high quality of the decennial census data products, requires the implementation of a disclosure avoidance mechanism that is better able to protect against these new, sophisticated vectors of attack.

Differential Privacy: A More Precise Way to Protect Data and Preserve Accuracy

For the 2020 Census, we're using a mathematical framework called "differential privacy" to protect your information in published data. As with protections we've used in the past, differential privacy works by adding statistical noise, or "fuzziness," to the data, but in a calibrated way, using a mathematical algorithm. That allows us to assure that enough noise is added to protect your information, but not so much as to damage the statistical validity of our publications. We call this the 2020 Census Disclosure Avoidance System.

We specifically tuned the algorithm that calibrates noise for the redistricting data (P.L. 94-171), known as the "TopDown Algorithm," to meet fitness-for-use accuracy targets for the redistricting and Voting Rights Act use cases. Learn more about the performance of the 2020 Census Disclosure Avoidance System as measured against published 2010 Census data at <https://content.govdelivery.com/accounts/USCENSUS/bulletins/2e5e8a6>.

United States® Census 2020

D-FS-GP-EN-049
August 2021

Guidance for Working With the Data

One of the key factors in the vulnerability of earlier censuses was the reporting of exact counts for all geographies, including census blocks, the base census geographic area, with an average of about 49 people in occupied blocks. That precision was essentially the key to reconstruction of data and re-identification of the people behind census statistics.

The 2020 Census reports exact counts for:

- The total population at the state level.
- The number and type of occupied group quarters facilities at the block level.
- The number of housing units, whether occupied or not, at the block level.

These exact counts are referred to as "invariants." To someone trying to identify the people behind census statistics, invariants are like pre-filled letters in a crossword puzzle or numbers in a Sudoku puzzle. The more invariants in a dataset, the easier it is to use that that information to try to match it with other data sources. Those matches can reveal additional information that help identify the people represented in census statistics.

Because the TopDown Algorithm applies noise to results and not input data, fewer invariants are specified. This is significant because the noise can be calibrated to reduce the effects of noise-related distortion in a way not possible with methods used for the 2010 Census, which ran the tabulations after the noise was applied.

This change requires data users to consider and heed the following when using the data:

- **Block-level data are noisy and should be aggregated before use.** As with the disclosure avoidance methods used for the 2010 Census, block-level data are noisy and should be aggregated before use. Because the amount of noise that the TopDown Algorithm adds to statistics does not vary directly by population size or geographic area, block-level data are most affected by disclosure avoidance procedures. For example, it is equally likely that five people could be added to an area with a population of 10,000 or 100. But as data are aggregated together—across blocks or across demographic groups—the accuracy of the resulting data increases. Census Bureau researchers found that for block groups, a minimum total population between 450 and 499 is sufficient to provide reliable characteristics of various demographic groups, whereas a minimum total population between 200 and 249 provides reliable characteristics for places and minor civil divisions.[1]
- **Counts are consistent within tables, across tables, and across geographies.** For example, rows within a table sum up to the parent row and universe. The total population count in Table P1 is consistent with the total population count in Table P2. In addition, block-level tables sum to their corresponding block-group-level tables, block-group-level tables sum up to their tract-level tables, and so forth.
- **Data should not be divided across tables.** For example, values from Table P2 should not be divided by values from Table H1 to obtain the average number of people per household. The separation of the people universe from the housing universe introduces some inconsistencies, particularly at low levels of geography (tract and smaller), such as more households than people. Calculations at higher levels of geography will be reliable, but users who want more accurate statistics on people per household should wait for the release of the Detailed Demographic and Housing Characteristics Files.

[1] Tommy Wright and Kyle Irimata, "Empirical Study of Two Aspects of The Topdown Algorithm Output For Redistricting: Reliability & Variability, (August 5, 2021 Update)," <www.census.gov/library/working-papers/2021/adrm/SSS2021-02.html>.

United States® Census 2020

D-FS-GP-EN-049
August 2021

- **Data may be subtracted across tables to obtain new counts without a substantial decrease in accuracy.** For example, you can subtract Table P3 (voting-age population by race) from Table P1 (total population by race) to obtain the population under the age of 18 by race. However, subtracting data across tables at the block level may yield improbable results, such as a large number of children under 18 years relative to the number of adults. Subtracting aggregations of blocks or subtracting at larger geographic levels should attenuate this issue.

- **Noise introduced through the Disclosure Avoidance System is not the only source of variability, or error, in 2020 Census data.** While the Census Bureau makes every effort to count everyone once, only once, and in the right place, even the best efforts at complete enumeration may miss some people and erroneously count others. Noise introduced by disclosure avoidance may compound underlying errors or may offset those errors. However, as the population in the geographic area gets larger, disclosure avoidance averages out. In most cases, these other sources of variability in census data are more significant than the variability due to confidentiality protection.[2]

---

[2] "2020 Census Data Quality," <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/data-quality.html#evaluating>.

D-FS-GP-EN-049
August 2021

United States® Census 2020