

## 2010 Demonstration Privacy-Protected Microdata File 2020-05-27

Over the past several months, the Census Bureau has been making a number of improvements to the 2020 Census Disclosure Avoidance System (DAS) to address the concerns raised by the data user community at the December 2019 Committee on National Statistics workshop. Throughout this process, we have received numerous requests for additional tools to help evaluate this ongoing progress. We are pleased to announce that in response to this feedback, and with the support of the [Committee on National Statistics' \(CNSTAT\) expert group](#), we have devised a solution to produce updated demonstration data sets.

The detailed summary metrics we released for these DAS data runs, and that we will continue to release as we make future improvements to the algorithm, allow our data users to assess these improvements and their impact on fitness-for-use in a variety of ways. That said, we recognize that our data users assess accuracy and fitness-for-use for diverse use cases in very different ways.

### Privacy-Protected Microdata Files

To assist with these assessments, we are now releasing new [“Privacy-Protected Microdata Files”](#) (PPMFs), which are the underlying microdata files for the entire nation used to generate the Detailed Summary Metrics. It is important to note that while the data in the PPMFs look like individual records, all of the data are privacy-protected. The microdata records generated by the DAS ensure respondent privacy through the application of differentially private statistical noise. The microdata included in the PPMF do not include any actual census responses. They are simply the microdata format used by the Census Bureau’s production system to produce privacy-protected tables that the DAS generated.

While these PPMFs are untabulated microdata records, members of the [Committee on National Statistics' expert group](#) are tabulating, formatting and posting data tables after upcoming design sprints. This partnership allows the census staff who would otherwise perform the time-intensive tabulation, data review and release process in-house to continue their focus on other important data collection and processing work.

### Algorithm Improvements Reflected in This Release

The 2010 Demonstration Privacy-Protected Microdata File 2020-05-27 is the same microdata that were used to generate the [Detailed Summary Metrics 2020-05-27](#). These microdata



census.gov  
2020census.gov  
@uscensusbureau

represent a run of the DAS that incorporated improvements to the algorithm finalized in the March 2020 development sprint. The most notable change involved how the DAS TopDown Algorithm (TDA) converts the formally private noisy tabulations taken from the confidential data into the non-negative integer counts that will be published, an operation that we call “postprocessing.”

Previously, the TDA conducted the postprocessing of all of the statistics for a particular geographic level at the same time. Unfortunately, as we saw in the demonstration data, the TDA had difficulty accurately performing this optimization when there were large quantities of statistics with zeros or very small values processed at the same time. The result was distortions in the data that effectively moved individuals from high- to low-density populations (e.g., from cities to rural areas, or from larger race groups to smaller race groups).

During the March sprint, we implemented a change to the algorithm design to address and mitigate this issue. Now, the TDA conducts the postprocessing in a series of passes through all the geographic levels.

At the national level, the state level, and finally at each lower level of geography, the first pass of the algorithm solely determines the population count for each unit within that geographic level (e.g., for all census tracts within a county).

Once those total population counts are determined, the second pass of the algorithm processes just the statistics necessary to produce the redistricting data (also known as the Public Law 94-171 data file), constraining those statistics to the sum of the population counts determined in the first pass.

The third pass through the algorithm then processes the core statistics necessary to support population by age, sex, and broad race/ethnicity categories for the demographic analyses that underlie the Population Estimates Program. Third-pass statistics are constrained to the sum of the statistics produced for the redistricting data.

A final pass through TDA processes the remainder of the statistics necessary for the Demographic and Housing Characteristics files and the Demographic Profiles, constraining these values to the sum of the ones produced in the third pass.

### [Microdata for Person and Housing Units](#)

This release of the PPMF 2020-05-27 only includes person-level microdata supporting the P.L. 94-171 Redistricting Data Summary File data and population tables in the proposed Demographic and Housing Characteristics file. We are not releasing the corresponding housing

unit-level microdata that support selected housing and household tables in the proposed Demographic and Housing Characteristics file. This is because no unit-level microdata were generated in the Disclosure Avoidance System (DAS) run that generated these microdata and that were assessed in our Detailed Summary Metrics 2020-05-27 release. In that run, the DAS team incorporated algorithm improvements to address error in household vacancy rates for the H1 table in the redistricting products, but the team did not yet complete the additional development work necessary to support the generation of the remaining housing and household tabulations included in the proposed DHC file.

Future PPMF releases will include both person and housing-unit microdata files.

### Privacy-Loss Budget

The DAS run that generated the [Detailed Summary Metrics 2020-05-27](#) and the PPMF 2020-05-27 uses the same global privacy-loss budget as was used for the 2010 Demonstration Data Products that were released in October 2019. By keeping the privacy-loss budget constant, we can isolate impacts attributable entirely to the DAS algorithm and post-processing development, and compare “apples to apples.” The person-level tables received a privacy-loss budget of  $\epsilon=4$ , and the housing unit-level tables received a privacy-loss budget of  $\epsilon=2$ .

For More Information, See: [Developing the DAS: Demonstration Data and Progress Metrics \(census.gov\)](#)