2010 Demonstration Privacy-Protected Microdata Files: Production Settings 2021-08-12

Over the past several months, the Census Bureau actively tuned the parameters of the 2020 Census Disclosure Avoidance System (DAS) to ensure fitness-for-use of the P.L. 94-171 Redistricting data product and the redistricting and Voting Rights Act use cases. Over the past two years, our development and tuning of the DAS benefited substantially from feedback from our federal advisory committees, stakeholder groups, and data users, as well as from the continuing support of the Committee on National Statistics' (CNSTAT) expert group.

To enable this invaluable feedback, we released a series of demonstration data products using 2010 Census data for evaluation. In June 2021, based on user feedback, the Data Stewardship Executive Policy (DSEP) Committee chose the parameters for production of the 2020 Census redistricting data. This set of demonstration data reflects those production parameters and settings.

Included in this release are Detailed Summary Metrics (DSM) and Privacy-Protected Microdata Files (PPMFs) at the chosen Privacy-loss Budget of 17.14 for person files and 2.47 for housing units.

Detailed Summary Metrics

The Detailed Summary Metrics we release for these Disclosure Avoidance System (DAS) data runs allow our data users to assess improvements and their impact on fitness-for-use in a variety of ways. They provide a variety of accuracy measures for a range of use cases that our data users have identified. Taken together, the detailed summary metrics provide a comprehensive snapshot of the overall fitness-for-use of the resulting data. That said, we recognize that our data users assess accuracy and fitness-for-use for diverse use cases in very different ways, so we are also releasing Privacy-Protected Microdata Files for users to perform more specific analyses that reflect their particular use cases.

Privacy-Protected Microdata Files

Privacy-Protected Microdata Files (PPMFs) 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 protection of respondent privacy through the application of differentially private statistical noise. The microdata included in the PPMFs do not include any actual census responses. They are simply the microdata format, generated by

the DAS, and used by the Census Bureau's tabulation production system to produce privacy-protected tables.

While these PPMFs are untabulated microdata records, the IPUMS National Historic Geographic Information System (NHGIS) will be tabulating, formatting and posting data tables for direct comparison to published 2010 Census tabulations. 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 processing work.

Privacy-Loss Budgets

The Census Bureau released the first set of demonstration data products for data user evaluation in October 2019. Since then, we released additional sets (in May 2020, September 2020, November 2020, and April 2021) to allow our data users to review and assess improvements to the DAS algorithms. Until the April 2021 release, we maintained the conservative PLB set for the initial demonstration data product (4.0 for the persons file, 0.5 for the housing units file). While that decision to hold the PLB constant across the earlier demonstration runs meant that the resulting data would have substantially more noise (error) than was expected in the final 2020 Census data products, holding the PLB constant enabled us and our data users to home in on the elements of the algorithm that were causing systemic distortions that needed to be addressed. We acknowledge that this has unfortunately led some of our data users to expect comparable amounts of noise in the final 2020 Census data.
The April 28, 2021, demonstration data featured a higher PLB of 10.3 for the persons file and 1.9 for the housing units file.

On June 9, 2021, the Data Stewardship Executive Policy Committee chose the production settings for the PLB of the redistricting data of 17.14 for the person file and 2.47 for housing units.

This higher PLB tunes the resulting data for greater accuracy and ensures that they meet the accuracy targets that we have established for redistricting, Voting Rights Act enforcement, and other priority uses of the redistricting data.

The files included in this release are:

- Detailed Summary Metrics (released June 8, 2021)
- Person-level file (ε= 17.14)
- Unit-level file (ε=2.47)

For More Information, see: Developing the DAS: Progress Metrics and Data Runs Web Page

Improvements and Tuning Reflected in This Release

The chosen global privacy-loss budget is exponentially higher than the privacy-loss budget used in the April 2021 demonstration data. In making its decisions, DSEP gave significant consideration to the feedback we received from our data users who analyzed the April 2021 demonstration data. That feedback, and steps taken to address those comments, include the following:

- Stakeholders identified a regression in the accuracy of data for tribal geographies and other off-spine geographies. The DAS team made changes to the 'optimized spine' to address these concerns; those changes were integrated into the spine that was approved by DSEP.

- Stakeholders identified several measures of bias in the summary metrics that they indicated were areas of concern. In particular, stakeholders addressed concerns about both geographic bias (i.e., the accuracy of population counts being different at larger and smaller geographies) and characteristic bias (counts of racially or ethnically diverse geographies being different than more racially or ethnically homogenous areas). The DAS team made changes to the post-processing system parameters to address these concerns; those changes were integrated into the parameters that were approved by DSEP.

- Data users identified a need for more accuracy in race and ethnicity statistics at many levels of geography. The DAS team addressed those concerns by allocating additional privacy-loss budget to the race and ethnicity queries at various levels of geography; those changes were integrated into the global privacy-loss budget and privacy-loss budget allocations that were approved by DSEP.

- Data users identified a need for more accuracy at the place, Minor Civil Division, and tract levels. The DAS team addressed these concerns both through changes to the optimized geographic spine and through allocation of privacy-loss budget; those changes were integrated into the privacy-loss budget allocations and system parameters that were approved by DSEP.

- Data users identified a need for more accurate statistics on occupancy rates at the block group and higher levels of geography. The DAS team addressed those concerns by allocating additional privacy-loss budget to the housing unit data; that change was integrated into the global privacy-loss budget and privacy-loss budget allocations that were approved by DSEP.

These improvements – as well as other adjustments to the system – were then verified against a broad suite of accuracy measures to ensure that they successfully addressed the feedback we received. We are not able to satisfy all stakeholder feedback. For example, some data users recommended nearly perfect accuracy in block-level data, which we are unable to achieve because it would undermine the ability to implement a functional disclosure avoidance system. We are both legally and ethically bound to protect the privacy of the data provided by and on behalf of our respondents.

Query Strategy

The DAS TopDown Algorithm (TDA) operates by taking a series of measurements (queries) of the tabulations that support the redistricting data product, adding a small amount of uncertainty (noise) to each of those queries to protect privacy, then converting the results of those queries back into individual-level records for the entire population. These queries can be structured in a number of different ways, with implications for the relative accuracy of different sets of cross-tabulations by demographic characteristics.

The query strategy used for the demonstration data set used the following queries for the person-level data:

Statistical Table	Number of queries
TOTAL POPULATION	1
CENRACE *(all 63 allowed combinations of the OMB-designated race categories*	63
HISPANIC *(Hispanic, not Hispanic*	2
VOTINGAGE *(≥18 years, <18 years of age*	2
HHINSTLEVELS *(institutional vs. non-institutional group quarters types*	2
HHGQ *(household and group quarters types*	8
HISPANIC*CENRACE	126
VOTINGAGE*CENRACE	126
VOTINGAGE*HISPANIC	4
VOTINGAGE*HISPANIC*CENRACE	252
DETAILED (HHGQ x VOTING_AGE x HISPANIC x CENRACE	2,016

PLB Allocation

The relative accuracy of different tabulations similarly depends on the share of the PLB allocated to each of the queries performed by the algorithm. Queries for smaller tabulations or cross-tabulations, like total population counts or voting age population counts, can be very accurate for any geographic level even with minimal allocation of PLB. Queries for cross-tabulations with a large number of categories (e.g., VOTINGAGE*HISPANIC*CENRACE, with 252 different combinations) require larger allocations of PLB to achieve comparable levels of accuracy.

PLB allocation by query for the production settings demonstration data was finely tuned at different levels of geography to meet the accuracy targets discussed above. In general, however, PLB was allocated proportionally by the size of the query, with the DETAILED query (HHGQ x VOTING_AGE x HISPANIC x CENRACE) receiving the largest share of PLB.

Additional allocations of PLB were made to particular queries at specific geographic levels to further enhance the accuracy of certain statistics. For example, extra PLB was allocated to the total population query at the Block Group level to improve population counts for many “off-spine” geographic entities like places.