DAS Noisy Measurements

John M. Abowd and Michael B. Hawes

NAC/CSAC Differential Privacy Working Groups
September 17, 2021

Shape
your future
START HERE >

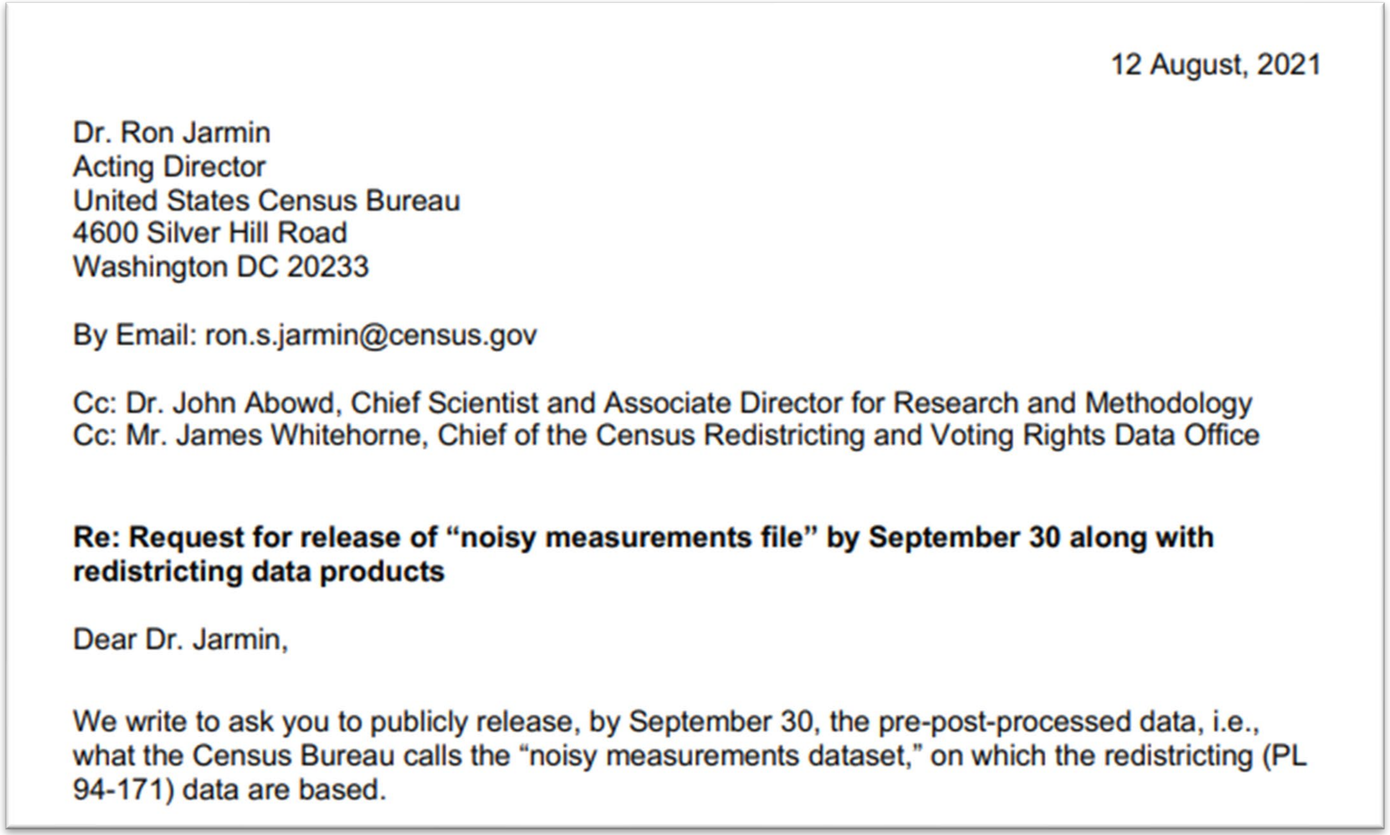
United States®
Census
2020

The Request

The Census Bureau has received a request to release the DAS *noisy measurements* for the 2020 Census P.L. 94-171 redistricting data.

The noisy measurements are tabulations taken directly from the confidential Census Edited File to which differentially private noise has been added.

These noisy measurements serve as the primary input used by the TopDown Algorithm (TDA), which post-processes these measurements to produce internally consistent microdata for tabulation and publication.



12 August, 2021

Dr. Ron Jarmin
Acting Director
United States Census Bureau
4600 Silver Hill Road
Washington DC 20233

By Email: ron.s.jarmin@census.gov

Cc: Dr. John Abowd, Chief Scientist and Associate Director for Research and Methodology
Cc: Mr. James Whitehorne, Chief of the Census Redistricting and Voting Rights Data Office

Re: Request for release of "noisy measurements file" by September 30 along with redistricting data products

Dear Dr. Jarmin,

We write to ask you to publicly release, by September 30, the pre-post-processed data, i.e., what the Census Bureau calls the "noisy measurements dataset," on which the redistricting (PL 94-171) data are based.

Signed by 59 prominent academics and privacy experts

Volume of Measurements

The TDA took over 16.6 Billion noisy measurements to produce the microdata supporting tabulation of the 6 tables (P1-P5 and H1) included in the 2020 Census P.L. 94-171 Redistricting Data Summary File.

Mitigating Disclosure Risk

One of the objectives of the 2020 Census Disclosure Avoidance System is to reduce the risk of re-identification of census respondents and the inadvertent disclosure of their census responses.

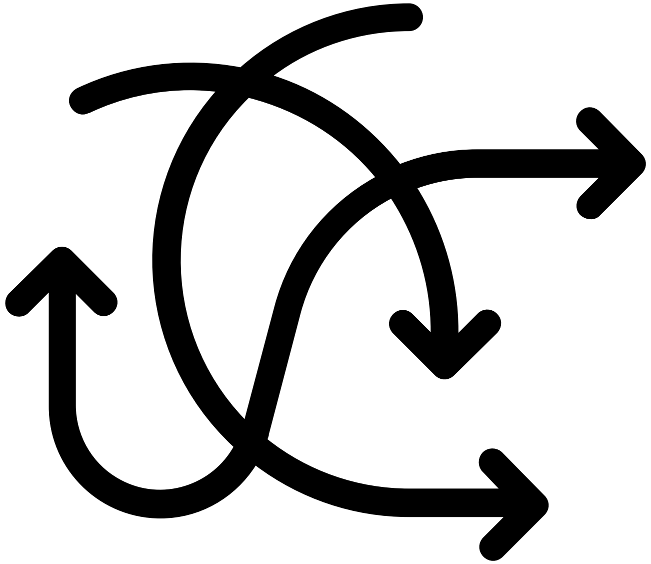
This is accomplished *generally* by establishing a privacy-loss budget (which protects against a broad spectrum of possible privacy attacks) and *specifically* by protecting against reconstruction-abetted re-identification (a known vector of attack). Differential privacy also protects against attacks that the Census Bureau has not pre-specified: inferences on individuals using any possible test an attacker could formulate, either currently known or unknown.

16.6 Billion Noisy Measurements

3.4 Billion Published Block-level Statistics

Releasing the balance (13.2 Billion measurements) not directly reflected in the published tabulations could increase the risk of reconstruction-abetted re-identification. This can only be assessed once the full suite of tabular data products have been produced (redistricting plus demographic and housing characteristics).

Ease of Use



The sheer volume of noisy measurements poses obvious challenges for dissemination and use.

Moreover, there are many different noisy measurements for each published statistic *(e.g., independent measurements of total population from the total population query, and from the sums of the query answers of each of the other queries taken).*

Estimation of any official statistic in tables P1-P5 and H1 uses a weighted average of many different noisy measurements, with the weights inversely proportional to the variance of the query. Estimation of a margin of error starts with this calculation as well.

Data Dependent Error

There are two sources of error in the published statistics:

Differentially private noise

- Unbiased
- Known distribution
- Reflected in the noisy measurements

Post-processing

- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

Data Dependent Error

Accounting for this error would require some combination of differentially private measurement against the CEF and parametric bootstrapping using the DAS code base.

There are two sources of error in the published statistics:

Differentially private noise

- Unbiased
- Known distribution
- Reflected in the noisy measurements

Post-processing

- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

Complications

The 2020 Census (P.L. 94-171) Redistricting Data Summary Files are the official population estimates used in redistricting and statutory allocation formulas.

1. **There cannot be two sets of official population estimates**
2. **How would calculations based on the noisy measurements be distinguished from the official redistricting data?**

The Census Bureau welcomes the advisory committees' recommendations on how we should address this issue.

Options

The Census Bureau has 4 principal options:

1. Release all the noisy measurements
2. Release weighted averages of the noisy measurements reflected in the published tabulations (DAS TopDown runs with the non-negativity constraints turned off)
3. Provide restricted access to all the noisy measurements in the FSRDC network
4. Not release the noisy measurements

The Census Bureau welcomes the advisory committees' recommendations on how we should proceed.

Future Discussion Topic

At a future meeting of the working groups, the Census Bureau would also like to discuss the possible development and release of total error-based measures of error/uncertainty.