Differential Privacy and the 2020 Decennial Census

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

Metropolitan Washington Council of Governments
Cooperative Forecasting and Data Subcommittee
July 14, 2020

Shape
your future
START HERE >

United States®
Census
2020

Confidentiality of Census Data

Title 13, Section 9 of the U.S. Code prohibits the Census Bureau from releasing identifiable data "furnished by any particular establishment or individual."

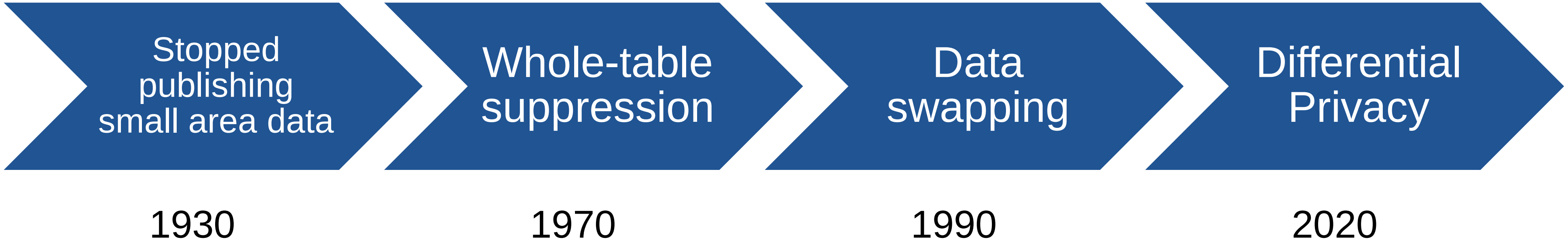
Census Bureau employees are sworn for life to safeguard respondents' information.

Penalties for violating these protections can include fines of up to $250,000, and/or imprisonment for up to five years!

The Census Bureau’s Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.



Stopped publishing small area data

1930

Whole-table suppression

1970

Data swapping

1990

Differential Privacy

2020

The Growing Privacy Threat

More Data and Faster Computers!

In today’s digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

Similarly, today’s computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents’ data.

The Census Bureau’s Decision

To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections, and will use differential privacy for the 2020 Census.

The Census Bureau has already successfully deployed differentially private solutions to protect other data products, including:

- Post-Secondary Employment Outcomes (PSEO)
- Veteran Employment Outcomes (VEO)
- OnTheMap for Emergency Management

Differential Privacy

All statistical techniques to protect privacy impose a tradeoff between the degree of privacy protection and the resulting accuracy of the data.

- **Swap rates, noise injection parameters, cell suppression thresholds, etc. determine this tradeoff.**

Differential privacy offers a number of important advantages over traditional statistical techniques to protect privacy.

- **Provides quantitative assessment of privacy risk.**
- **Infinitely tunable – parameter "dials" can be set anywhere from perfect privacy to perfect accuracy.**
- **Privacy guarantee is mathematically provable and future-proof.**
- **The precise calibration of statistical noise enables optimal data accuracy for any given level of privacy protection.**

Shape your future START HERE >

United States® Census 2020

Demonstrating Privacy, Assessing and Improving Accuracy

The DAS Team's priorities over Fall 2019 were:

- To scale up the DAS to run on a (nearly) fully-specified national histogram
- To demonstrate that the DAS can effectively protect privacy at scale
- To permit the evaluation and optimization of the DAS for accuracy and "fitness for use"

These initiatives were largely successful, but much more work needs to be done over the remainder of this year.

The engagement and efforts of our data users have been enormously helpful in helping to identify and prioritize this remaining work.

Committee on National Statistics Workshop

December 11-12, 2019

Evaluation of the Demonstration Data Products (DDP): 2010 Census data run through a preliminary version of the 2020 DAS

Data user assessments and findings on DAS implications for:

- **Redistricting and related legal use cases**
- **Identification of rural and special populations**
- **Geospatial analysis of social/demographic conditions**
- **Delivery of government services**
- **Business and private sector applications**
- **Denominators for rates and baselines for assessments**

What We've Learned

The October vintage of the DAS falls short on ensuring "fitness for use" for several priority use cases.

Particular areas of concern:

- Population counts for political geographies
- Population counts for American Indian and Alaska Native Tribes and Tribal Areas
- Systemic biases (e.g., urban vs. rural)
- Housing statistics and vacancy rates

These issues are substantially driven by post-processing of the noisy statistics within the DAS.

What We've Learned

- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise (tunable through selection of ε)
  - Post-processing error due to process of creating internally consistent, non-negative integer counts from the noisy measurements
- **Post-processing error tends to be much larger than DP error**
- **Improving post-processing is not constrained by DP**

Making population counts more accurate.

A set of accuracy metrics have been developed based on use cases and stakeholder feedback. The metrics will allow the public to see the improvements that are made to the Disclosure Avoidance System.

The selected metrics:

- Reflect input from external data users;
- Show differences between major DAS runs and publicly available 2010 tabulations
- Provide accuracy, bias, and outlier information for basic demographic tabulations
- Provide accuracy, bias, and outlier information for categories of use cases

These metrics will inform data users of accuracy improvements we are able to make while also informing their ongoing engagement throughout the remaining work.

https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

Privacy-Protected Microdata Files

To further assist with data users’ evaluations, we are also releasing “Privacy-Protected Microdata Files” (PPMFs), which are the underlying microdata files for the entire nation used to generate the Detailed Summary Metrics.

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.

https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

Upcoming Milestones

September 2020

- DSEP will set final list of invariants for the 2020 Census (beyond apportionment totals, which are already invariant)
- The Census Bureau has already announced that state population counts and block-level unit counts (Group Quarters and Housing Units) will be reported as enumerated.

March 2021

- DSEP will set the final privacy-loss budget for the 2020 Census and its allocation across 2020 Census data products.
- This decision will be informed by extensive assessment of data accuracy for priority use cases of decennial data, feedback from our stakeholders, and our legal obligations under Title 13.

June-July 2021

- PL94-171 Redistricting Data files will be released.
- Additional data products, including the Demographic and Housing Characteristics files and Demographic Profiles will follow later in 2021.

Shape your future START HERE >

United States Census 2020

Additional Resources

Disclosure Avoidance and the 2020 Census Website

https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Questions? Suggestions?

Email them to 2020DAS@census.gov

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

301-763-1960 (Office)
michael.b.hawes@census.gov