2020 Census Disclosure Avoidance Improvement Metrics

Jason Devine, Christine Borman, and Matthew Spence

U.S. Census Bureau

Committee on National Statistics - National Academies
Disclosure Avoidance Working Group
March 18, 2020

Pre-decisional Draft

Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

Shape your future START HERE >

United States Census 2020

Committee on National Statistics Workshop

December 11-12, 2019

The focus of the workshop was the evaluation of the 2010 Census Demonstration Products developed from 2010 Census data produced using a preliminary version of the 2020 Disclosure Avoidance System (DAS)

Included the PL94-171 tables and many of the tables planned for inclusion in the 2020 Demographic and Housing Characteristics File (DHC) data products

Data user assessments and findings highlighted implications for:

- **Redistricting and related legal use cases**
- **Distribution of funds**
- **Accuracy of counts for rural and small population groups**
- **Geospatial analysis of social/demographic conditions**
- **Delivery of government services**
- **Business and private sector applications**
- **Denominators for rates and baselines for assessments**

Shape your future
START HERE >
United States Census 2020

What We've Learned: Accuracy

- **The October vintage of the DAS falls short on ensuring "fitness for use" for several priority use cases.**
- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise
  - Post-processing error due to statistical inference creating non-negative integer counts from noisy measurements
- **Post-processing error tends to be much larger than differential privacy error**
- **Positive bias in small counts and negative bias in large counts is the result of**
  - The inclusion of invariants and structural zeros in the DAS
  - Post-processing error specifically introduced by the use of a Non-negative Least Squares (L2) optimization routine
- **Improving post-processing is independent from the use of differential privacy**
- **Current initiatives include incorporating legal and political geographies into the geographic spine, adopting a multi-phase approach to post-processing, and developing data metrics to allow stakeholders to assess improvements**
- **Focus of metrics is on the total error, both from differential privacy and post processing**

Shape your future START HERE >
United States Census 2020

Metrics: Making population counts more accurate

A set of metrics are being developed based on use cases and stakeholder feedback. The metrics will allow the public to see the improvements that are made leading up to the finalization of the TopDown Algorithm (TDA) for Group I data products.

The selected metrics will:

- **Be straightforward and easy to interpret**
- **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 and specific use cases**
- **Metrics can be added as the DAS is scaled up to include additional variables**

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

Identifying Key Metrics

1. Reviewed use cases
   - Census inputs for the Population Estimates Program were considered one use case
   - Use cases received through a Federal Register Notice
   - Input received during the Harvard Data Science Review Symposium
   - Use cases presented at the Committee on National Statistics workshop
   - Other use cases identified through email submissions and other outreach
2. Categorized use cases
3. Identified accuracy measures for each use case category

Shape your future START HERE >
United States Census 2020

Measures used for the Metrics

Metrics are based on the following measures:

Mean Absolute Error (MAE)
Root Mean Squared Error (RMSE)
Mean Absolute Percent Error (MAPE)
Coefficient of Variation (CV)
Mean Algebraic Percent Error (MALPE)
Mean Error (ME)
Percentage Point MAE and ME for Rates
Percent Difference Thresholds
Total Absolute Error of Shares
Counts of Outliers

Shape your future START HERE >
United States Census 2020

Measures Defined

Mean Absolute Error (MAE) = (Σ (| MDF – CEF|))/N

MAE takes the absolute value of the difference between the MDF and the CEF value for each evaluation geography, sums them, and divides by the number of evaluation geographies. Provides an easy to interpret measure of the numeric error.

Root Mean Squared Error (RMSE) = Sqrt[(Σ (MDF – CEF)^2)/N]

RMSE takes the squared difference between the MDF and the CEF value for each evaluation geography, sums them, and divides by the number of evaluation geographies, then takes the square root. This is the standard Census Bureau error measure for sample survey statistics. It equals the standard error, when there is no bias.

Mean Error (ME) = (Σ (MDF – CEF)/N

ME takes the value of the difference between the MDF and the CEF value for each evaluation geography, and sums them, and divides by the number of evaluation geographies. Provides an easy to interpret measure of the numeric error.

Mean Absolute Percent Error (MAPE) = ((Σ ((|MDF – CEF|)/ CEF))/N)*100

MAPE takes the absolute value of the difference between the MDF and the CEF value for each evaluation geography, divides that by each respective CEF value, sums them, divides by the number of evaluation geographies, and multiplies the result by 100. An easy to interpret relative measure of error.

Coefficient of Variation (CV) = (RMSE/((Σ CEF)/N))*100

CV is the relative error version counterpart of RMSE. An easy to interpret relative measure of error.

Mean Algebraic Percent Error (MALPE) = ((Σ((MDF – CEF)/CEF))/N)*100

MALPE takes the difference between the MDF and the CEF value for each evaluation geography, divides that by each respective census value, sums them, divides by the number of evaluation geographies, and multiplies the result by 100. Identifies systematic bias.

Total Absolute Error of Shares = Σ|((MDF/ΣMDF) – (CEF/ΣCEF))|

This measure finds the proportion of each MDF value to the total MDF value for the summary geography and subtracts the proportion of the CEF value to the total CEF value for the summary geography. The absolute value of these proportional differences across evaluation geographies is then summed to the summary geography level. Provides a measure of the distributional error in the MDF shares.

Percent Difference Thresholds = Number of percent differences above a certain threshold

Unlike the other measures, Percent Difference Thresholds is a numeric value that relies upon on a set threshold (e.g., 5 and 10 percent). The end measure simply represents a count of how many evaluation geographies in the summary area exceeded a particular threshold in their absolute percent difference of the estimate. Provides an intuitive measure of the distribution of differences.

Shape your future
START HERE >
United States Census 2020

Metrics Geography

Data users are particularly concerned about data fitness for use for counties, political entities such as incorporated places and minor civil divisions (MCDs), American Indian/Alaska Native/Native Hawaiian (AIANNH) Areas, Census Tracts, and for limited uses cases, blocks

Metrics will be provided mainly for counties, places, and tracts, and in some cases block groups and block

In most cases, measures will be provided by geography type by size categories

Geographic metrics will isolate the resident population of the United States from the resident population of Puerto Rico

Shape your future START HERE >
United States Census 2020

Use Case Categories

Zero-Sum Total: Uses that rely on the accuracy of the distribution in addition to the overall accuracy because a fixed amount of something is being distributed across categories. For these uses, the accuracy needs may be greater for the distribution than for the actual estimates.

Zero-Sum Category: Same as zero sum total except use cases rely on estimates for some subset of the total.

Variable-Sum Total: Similar to zero sum use cases except that the total of what is being distributed can vary. For these types of uses, the accuracy of the estimate is more important than the accuracy of the distribution.

Variable-Sum Category: Same as variable sum total but for a subset of the population.

Single Year of Age Accuracy: These use cases require accuracy for single years of age rather than age groups.

Accuracy of Rates: These uses cases rely on a measure of the size of a subgroup(s) within the total population.

Percent Threshold: Use case depends on the subset of the population crossing a percent threshold.

Average Numeric Threshold: Use case depends on the subset of the population crossing a numeric threshold.

Shape your future START HERE >
United States Census 2020

Example Use Cases

- Total Population – state and local planning, distribution of tax revenue, and as the denominator for several important statistics, including mortality rates
- Total Housing Units, Occupancy Rates, and Vacancy Rates – state and local planning – including for transportation planning specifically, and to help researchers understand aspects of the population
- Average Household Size - used by states and local governments to create their own population estimates.
- Total Population by Single Year of Age for those Under 20 and Other Ages – Basis for statistics related to teens, to predict generational change for school district planning, school enrollment statistics, and catchment ratios.
- Total Population for those 65 Years and Older and Other Ages – Needed for state and local planning, such as accessing emergency preparedness, planning for nursing homes, and ensuring services are reaching this at-risk sub-population.
- Median Age and Total Population by Age in 5-Year Age Groupings for those between 20 and 65 Years Old by Sex – Important for state and local planning related to the "aging" of a community and used to derive health, education, and criminal justice statistics.
- Total Population by Major Race and Hispanic Origin Groups Alone – used to enforce the Civil Rights Act of 1964 and the Voting Rights Act of 1965
- Total Population for American Indian and Alaska Native Race Group – used in federal funding allocation formulas for the Tribal Transportation Programs and Indian Housing Block Grant funding.
- Identification of Small Subpopulations – used for outreach and programs designed for specific populations
- Group Quarters Subpopulations – used for demographic research, and for prison populations, for state-reapportionment.

Shape your future
START HERE >
United States Census 2020

Going Forward

A document describing all of the proposed metrics, along with metrics calculated for the 2010 Demonstration products will be provided to the Committee on National Statistics Workshop and other stakeholders for review and comment

The agreed on metrics will be calculated and provided to the public to show improvement, or lack of improvement, with each future major DAS run

Shape your future START HERE >
United States Census 2020