Block Discrepancies by State for the 2021-04-28 Demonstration Data Products

David Van Riper
IPUMS NHGIS
vanriper@umn.edu
2021-05-19

I created two spreadsheets that summarize block-level discrepancies by state between the 2010 Summary File 1 data and the two 2021-04-28 demonstration data products (one with a global privacy loss budget of 12.2 and the other with a global privacy loss budget of 4.5) released by the Census Bureau and processed by IPUMS NHGIS.

This document describes the contents of the spreadsheets.

Race_by_age_differences.xlsx

This spreadsheet summarizes block-level discrepancies for the following race/ethnicity categories:

- Total population (all races)
- White alone, non-Hispanic or Latino
- Black or African American alone or in combination
- Asian alone or in combination
- Hispanic or Latino

The race/ethnicity categories comprise the rows of the spreadsheet.

Discrepancies for the five race/ethnicity categories are also analyzed by age:

- Total population (all ages)
- Population under 18 years of age
- Population 18 years and older

The age categories and the demonstration product (e.g., epsilon 12.2 or epsilon 4.5) comprise the columns of the spreadsheet.

For each combination of race/ethnicity and age categories, I computed the following metrics by state:

- Mean absolute numeric error
- Mean absolute percent error
- Percent of blocks with an absolute percent error greater than 5
- Percent of blocks with an absolute percent error greater than 10

The *mean absolute numeric error* is computed via:

$$Mean\ absolute\ numeric\ error\ = \frac{\sum|dp\ -\ sf|}{n}$$

Where n is the number of blocks in a given state, and $\sum|dp\ -\ sf|$ is the sum of the absolute numeric differences between the counts from the demonstration product (*dp*) and 2010 Summary File 1 (*sf*).

The *mean absolute percent error* is computed via:

$$Mean\ absolute\ percent\ error\ = \frac{\sum\left(\frac{|dp\ -\ sf|}{(dp\ +\ sf)\ /\ 2}\right)}{n}$$

Where n is the number of blocks in a given state, and the numerator is the sum of the absolute percent errors for all blocks in the state. The absolute percent error for a given block is the absolute difference between the counts from the demonstration product (*dp*) and 2010 Summary File 1 (*sf*) divided by the average of the *dp* and *sf* counts.

I use the average of the *dp* and *sf* counts in the denominator to account for situations where the *sf* count is zero and the *dp* count is non-zero. If I only used the sf count in the denominator, then the percent error would be undefined. By using the average in the denominator, I guarantee that fraction will not be undefined.

If both the *dp* and *sf* counts are zero for a given block, I set the absolute percent error to zero. A count of zero in both should be interpreted as a zero percent error.

Miscellaneous_differences.xlsx

This spreadsheet summarizes block-level discrepancies that meet the following criteria:

- Blocks changed from greater than 50% Non-Hispanic White alone to less than 50% Non-Hispanic White alone
- Blocks with population age 0 to 17 but no population ages 18+
- Blocks with population in Summary File 1 but no population in DP file
- Blocks with population in households but no occupied housing units
- Blocks with occupied housing units but no population
- Blocks with more than 15 persons per household

For each criterion, I compute the count and percentage of blocks that meet it, by state. These criteria comprise the rows of the spreadsheet, and the demonstration data product epsilon value (e.g., epsilon 12.2 or epsilon 4.5) comprise the columns.

These following describes the criteria in more detail.

Blocks changed from greater than 50% Non-Hispanic White to less than 50% Non-Hispanic White

I flagged census blocks that were greater than 50% non-Hispanic White alone in the 2010 Summary File 1 dataset and less than 50% non-Hispanic White alone in the demonstration data. Essentially, these are census blocks that flip from majority non-Hispanic White in the 2010 Summary File minority non-Hispanic White in the demonstration data.

Blocks with population aged 0 to 17 but no population aged 18 and older

I flagged census blocks in the demonstration data that contained persons aged 0 to 17 but no persons aged 18 and older. These are census blocks populated solely by minors in the demonstration data.

Blocks with population in 2010 Summary File 1 but zero population in demonstration data

I flagged census blocks that contained persons in the 2010 Summary File 1 but zero persons in the demonstration data.

Blocks with population in households but no occupied housing units

I flagged census blocks that contained population in households (I subtracted the group quarter population from the total block population) but had no occupied housing units. These blocks are improbable based on census rules – a block with at least one person in a household should have at least one occupied housing unit.

Blocks with occupied housing units but no population

I flagged census blocks that contained occupied housing units but had zero persons. Logically, if a census block contains at least one occupied housing unit, it must have at least one person residing in the block.

Blocks with more than 15 people per household

I flagged census blocks that had more than 15 people per household. I divided the number of people in households by the number of occupied housing units. If the quotient was greater than 15, I flagged the block.