Feedback on the April 2021 Census Demonstration Files

David Van Riper, Jonathan Schroeder, and Steven Ruggles
IPUMS-NHGIS
University of Minnesota
May 28, 2021

Introduction

Producing accurate, usable data while protecting respondent privacy are dual mandates of the U.S. Census Bureau. In 2018, the Census Bureau announced it would use a new disclosure avoidance technique based on differential privacy for the 2020 Decennial Census of Population and Housing. Instead of suppressing data and swapping sensitive records as the Bureau had previously done, the new approach injects noise into counts. Unfortunately, noise injection also makes the data less accurate and can hamper many use cases.

The Census Bureau has released five demonstration products that apply different versions of the new approach to 2010 census data. To assess the most recent demonstration products, we compare them with previous demonstration products and with the originally published 2010 summary data. The final two demonstration products were released on April 28, 2021. The e4 product has the same overall privacy budget (ε≈4.4) as previous demonstration products but reallocates the budget to different geographic units and modifies post-processing. The e12 budget is much larger (ε≈12.2), which would be expected to provide substantially greater accuracy. We understand that the e12 product uses the same parameters that the Census Bureau currently plans to use for the 2020 census.

Analysis of Total Population Counts

We first analyze large discrepancies between the five demonstration products and the originally published 2010 data, measured as the percentage of geographic units where population counts differ by more than 5%. We compare four different geographic classifications: tracts, block groups, places, and American Indian Areas/Alaska Native Areas/Hawaiian Home Lands (here abbreviated as American Indian). We divide the population size of each geographic classification into deciles based on the 2010 total population.

Figure 1 shows the percentage of each decile of each geographic unit that differed from the true 2010 population counts by over 5%. The five rows of charts represent the five demonstration products produced by the Census Bureau. The leftmost column of charts show the error for census tracts. The demonstration products perform well for all but the smallest tracts, but the most recent demonstration products are not as accurate as the earlier ones.

The second column of Figure 1 shows the percentage of census block groups with error in total population counts exceeding 5%. These errors are substantially greater than the tract errors, and are especially pronounced for the April 2021 e4 product, which performs substantially worse than any prior demonstration product. By contrast, the accuracy of census places, shown in the third column of graphs, has improved substantially in the most recent demonstration products.

Finally, the population counts for American Indian Areas/Alaska Native Areas/Hawaiian Home Lands, although improved relative to the earliest demonstration products, remain problematic, with substantial error for all but the largest units.

Even where these charts show *relatively* few errors, such as for places in the April 2021 e12 product, there remain many instances of unacceptably large error. For example, the census-designated place of Fire Island, NY, had a 2010 population of 292, but the e12 product reports it as 392, a +34% error. The village of Vandalia, MI, has had a population between 300 and 450 in every census since 1880, including 2010 when its population was 301, but the e12 product reports its population as 245, a -19% error.

Analysis of Black and Hispanic/Latino Population Counts

To understand how the disclosure avoidance measures affect the counts for population subgroups, we carried out the same analyses for the Black-alone population (Figure 2) and the Hispanic/Latino population (Figure 3). These figures show far more large discrepancies than the total population counts.

For the Black population shown in Figure 2, the new demonstration products do not represent a substantial improvement over prior releases, and the pervasive discrepancies are disturbing. For most block groups and places the discrepancy in the Black population exceeds 5% in every demonstration product, and even the e12 product--which ought to be the most reliable one--does not perform much better than earlier demonstration products. For the Black population the census tracts in the April 2021 e4 product perform no better than the previous releases, and even the e12 product is only a small improvement

The Hispanic/Latino population, shown in Figure 3, is even less accurate than the Black population. The discrepancies exceed 5% for the great majority of geographic units. Again, there is little or no improvement between the most recent data releases and the earlier ones, and even the E12 product indicates unacceptable levels of error.

Additional Errors and Inconsistencies

Errors in other characteristics are equally problematic. The errors in counts of the number of children in the population of administrative units would wreak havoc on educational planning. For example, the e12 data product has a +3.4% overcount of children in Hoboken City School District, NJ (population 50,005); a +7.2% overcount in Naches Valley School District, WA (pop 8,078); and a +12.0% overcount in Mendocino Unified School District, CA (pop 5,665). Not only are there errors in the number of children in school districts, it appears that those errors include systematic biases. Among the smallest decile of school districts, 63% have an overcount of children, with a mean percent error of +6.1%. This is an example of a pervasive systematic bias found throughout these datasets: where counts are very small, they tend to be biased upwards.

In addition to the many large relative errors (mostly, though not all, in smaller counts), there are also numerous cases of very large absolute errors. In the e12 product, the total population of the Los Angeles School District is 5,950 above the actual count. Because of the large population of the district, this overcount is only a 0.1% increase in its total population, but an overcount of nearly 6,000 is still not a “small” inaccuracy, and importantly, it is not a *uniform* overcount among all groups. Most of the increase is due to an overcount in *children* by 4,790, a more substantial 0.4% increase, and in a group of particular importance for a *school district*. Conversely, the E12 count of children in the neighboring Long Beach School District is low by 1,536, resulting in a -1.2% error in a very large school district (total population 510,940). Such inaccuracy is sure to have adverse impacts on these districts’ ability to serve their students and their families effectively.

The data also include numerous logical inconsistencies. For example, there are many cases in which the number of households exceeds the population size. Among county subdivisions, the largest such discrepancy occurs in Republican City township, NE (a minor civil division). The township actually had 131 residents and 67 households, but according to the E12 data product, it had 140 residents and 180 households. This is impossible, of course, but it also represents a +269% error in the household count as well as a +24 percentage-point error in the township’s housing occupancy rate (from 14% to 38%). Lewis town, VT, had no residents in 2010, but the E12 product assigns it 8 households (with no population). There are also many cases in which the number of adults in the population is implausibly low, including *91,047 blocks* where the E12 product codes *all* residents as children, with no adults present.

Conclusion

The new demonstration products are limited to the content of PL94-171, and therefore do not permit analysis beyond a small number of very simple characteristics. Earlier demonstration products revealed major problems in other characteristics--such as age distributions--and we are unable to assess whether these errors have been addressed.

Given the limited time available and the limited content provided in the new demonstration products, we were unable to conduct more than a basic analysis. Nevertheless, that basic analysis yields profoundly disturbing results.

There were minimal improvements in the performance of the new demonstration files relative to the previous ones. We were disappointed to discover that the E12 file is not substantially more accurate on most measures than the e4 files. We were also dismayed to learn that the new datasets were virtually as bad as the previous ones with respect to the accuracy of counts for minority populations. The Census Bureau describes the e12 product as highly accurate. We find that although the e12 product has mitigated some egregious errors for the total population, major discrepancies remain for minority populations. This level of error will severely compromise demographic and policy analyses.

The demonstration files include troubling cases with extreme error in total or adult population counts, even if these are comparatively rare. Small localities can sometimes have their population doubled or halved by the disclosure avoidance noise. For example, the e12 product doubles the population of Saltaire village, NY, from 37 to 75, and it triples the population of Islandia city, FL, from 18 to 57.

For those who might object that these examples constitute cherry picking, we note that each "cherry" represents a community that deserves good data. The planned system would enter every community into a bad data lottery where the losers suffer for 10 years with material losses of federal funding. Litigation by undercounted communities is inevitable, and in these cases the Census Bureau will probably be forced to release the true counts.

Based on our analysis of the new demonstration products, we conclude that they are not ready for public release. We found pervasive biases and inconsistencies, high levels of inaccuracy in the counts of minority populations, and isolated large errors in the population counts for particular communities. Accordingly, the disclosure avoidance measures used in the e12 data product make the data unfit for many research and administrative purposes.

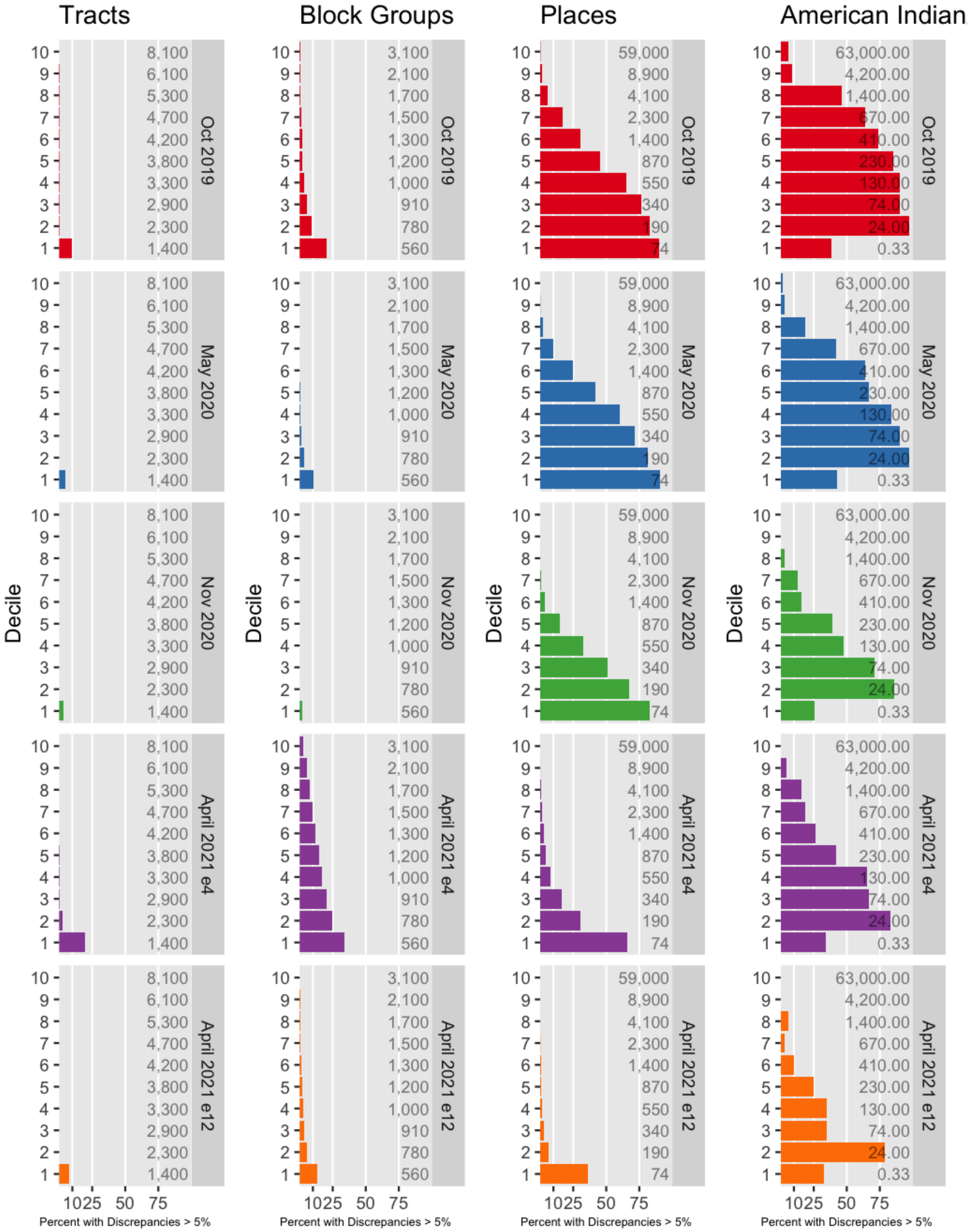

Figure 1. Percentage of units with a discrepancy between the demonstration data and 2010 Summary File 1 products greater than 5% for total population counts.

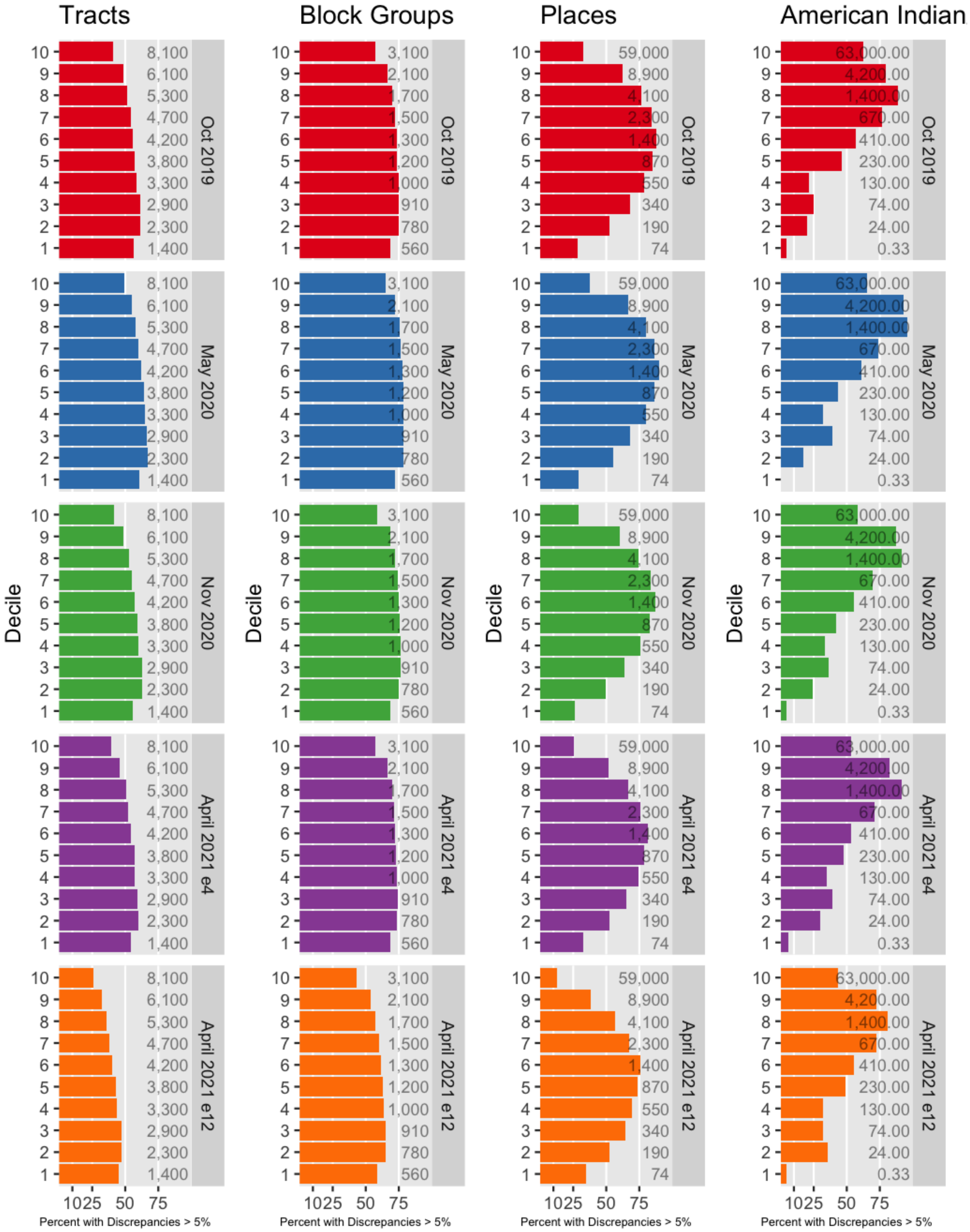

Figure 2. Percentage of units with a discrepancy between the demonstration data and 2010 Summary File 1 greater than 5% for Black-alone population counts.

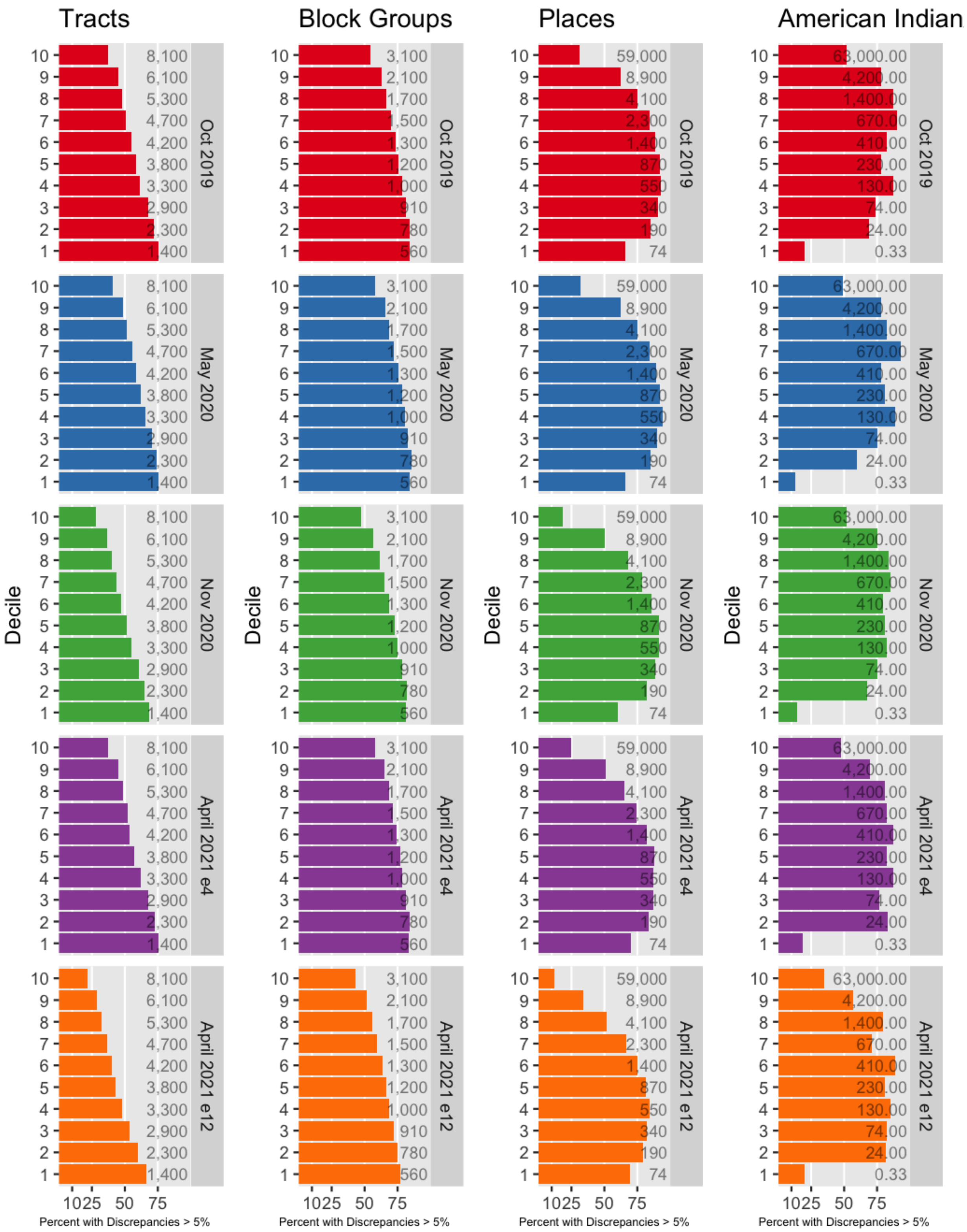

Figure 3. Percentage of units with a discrepancy between the demonstration data and 2010 Summary File 1 greater than 5% for Hispanic population counts.