Title: The 2020 U.S. Census Differential Privacy Method Introduces Disproportionate Discrepancies for Rural and Non-White Populations

Authors: J. Tom Mueller^a, Alexis R. Santos-Lozada^b

^aDepartment of Geography and Environmental Sustainability, University of Oklahoma, Norman OK, 73069

^bHuman Development and Family Studies, The Pennsylvania State University, University Park, 16802

Corresponding Author

J. Tom Mueller, PhD

Department of Geography and Environmental Sustainability

University of Oklahoma

562 Sarkeys Energy Center, Norman OK, 73069

E-mail: tom.mueller@ou.edu

Abstract

The recently finalized changes to the disclosure avoidance policies of the U.S. Census Bureau for the 2020 census, grounded in differential privacy, have faced increasing criticism from demographers and other social scientists. Scholars have found that estimates generated via Census-released test data are accurate for aggregate total population statistics of larger spatial units (e.g. counties), but introduce considerable discrepancies for estimates of subgroups. At present, the ramifications of this new approach remain unclear for rural populations. In this brief, we focus on rural populations and evaluate the ability of the finalized differential privacy algorithm to provide accurate population counts and growth rates from 2000 to 2010 across the rural-urban continuum for the total, non-Hispanic white, non-Hispanic Black, Hispanic or Latino/a, and non-Hispanic American Indian population. We find the method introduces significant discrepancies relative to the prior approach into counts and growth rate estimates at the county level for all groups except the total and non-Hispanic white population. Further, discrepancies increase dramatically as we move from urban to rural. Thus, the differential privacy method likely introduced significant discrepancies for rural and non-white populations into 2020 census tabulations.

Key words: Differential Privacy, Population Growth, Rural, Census, Decennial Counts

Introduction

The U.S. Census Bureau has adopted a new disclosure avoidance system for 2020 decennial tabulations (Mervis, 2019; Schneider, 2021; U.S. Census Bureau, 2021a). The system relies on differential privacy—an approach which uses a Census-designed algorithm known as TopDown to inject noise into aggregate population counts based on a global privacy-loss budget (also known as “epsilon”) and a series of specifications that remain invariant in comparison to the source file (Abowd & Schmutte, 2019). The goal of this policy is to prevent disclosure, or the re-identification of individuals within census data, while still providing public data accurate enough for science and public policy.

This policy represents a significant shift from traditional disclosure avoidance protocols, which have historically involved suppression, top and bottom coding, and limited amounts of data swapping (Ruggles, Fitch, Magnuson, & Schroeder, 2019; Zayatz, 2007). To say this policy has been met with controversy would be an understatement. The Census Bureau has provided researchers with test data for evaluation which applies versions of the new 2020 system to 2010 data, and multiple papers have found significant issues with each successive vintage of experimental data. In general, demographers find there is little issue with overall total population counts for larger geographic units such as counties and block-groups, but serious issues for population subgroups and small geographic areas (Hauer & Santos-Lozada, 2021; Kenny et al., 2021; Santos-Lozada, Howard, & Verdery, 2020). For example, researchers found that migration estimates derived from prior vintages of demonstration data were only usable for about half of US counties (Winkler et al., 2021), and that the algorithm produced inaccurate infant mortality rates for rural counties (Santos, 2021). As foreshadowed by Figure 1 below, our results do not suggest the Census Bureau has resolved this issue in their finalized differential privacy algorithm.

[Figure 1 here]

In this brief, we continue the effort of evaluating the new disclosure avoidance system and use the final release of census-provided differential privacy demonstration data to evaluate the differences in 2010 population estimates and county-level population growth rates from 2000 to 2010 between the traditional disclosure avoidance approach and the new differential privacy algorithm. In doing so, we evaluate how the algorithm used in the final census redistricting data is likely influencing discrepancies—meaning differences between statistics generated via traditional disclosure avoidance methods and the new differential privacy algorithm—in 2020 population counts across the rural-urban continuum.

Rural demography is persistently plagued with data issues (Isserman & Westervelt, 2006; Mueller et al., 2021). Data is often suppressed, unavailable, or unreliable. Thus, the injection of further noise into data already beleaguered by difficulties requires evaluation and caution. Having strong data on rural populations is essential, as this data is often used for important policy decisions, as well as in research focused on improving the lives of rural people—who face many unique difficulties related to healthcare, education, and employment (Bolin et al., 2015; Brown & Schafft, 2011; Crosby, Wendel, Vanderpool, & Casey, 2012).

Rural America is not a monolith. Although made up of a larger share of non-Hispanic white individuals than urban America, rural America is far more ethnic and racially diverse than is often assumed (Johnson & Lichter, 2008, 2010; Lee & Sharp, 2017; Lichter, 2012). It is imperative that data on these populations

is accurate, as non-white populations in the rural reaches of the United States face significant levels of structural and interpersonal discrimination, resulting in worse health outcomes, higher rates of poverty, and lower educational outcomes, among other hardships, than their white neighbors and urban counterparts (Brooks, Mueller, Thiede, 2020; Cosby et al., 2019; Gagnon & Mattingly, 2018; James & Cossman, 2017; Thiede, Kim, & Valasik, 2018; Tickamyer, Sherman, & Warlick, 2017). While rural America is not homogeneous in its ethnoracial makeup, there is significant regional clustering of populations. There exist well-established destinations for Hispanic and Latino/a immigrants; rural individuals racialized as Black are clustered in the South; and rural American Indian populations are heavily clustered on and around reservation land (Johnson & Lichter, 2010; Liebler & Ortyl, 2014). Thus, it is important to not only ask how effective differential privacy data is at estimating ethnoracial-specific growth patterns along the rural-urban continuum, but it is also essential to test for regional variation in accuracy.

Drawing on the outlined issues, we proceed through two primary research questions:

1. How does the finalized differential privacy algorithm impact growth estimates across the rural-urban continuum for the overall population and specific ethnoracial groups?
2. Does this vary regionally across the United States?

Methods

We pursue our research questions using the final release of the U.S. Census Bureau Differential Privacy Demonstration Products, released in August of 2021. All data were extracted via IPUMS-NHGIS (Manson, Schroeder, Van Riper, & Ruggles, 2021). Differential privacy algorithms are structured around a global privacy-loss budget, or the acceptable amount of information about a record leaked when the data is made public (Abowd & Schmutte, 2018). This acceptable level of privacy loss is conventionally known as epsilon (Abowd & Schmutte, 2018). Epsilon can be interpreted as follows: a smaller epsilon will result in more privacy, but less accuracy, whereas a larger epsilon allows for more privacy loss, but will result in more accurate public data. Epsilon can range from zero to infinity, with zero representing perfect privacy/no accuracy and infinity representing no privacy/perfect accuracy (Van Riper, Kugler, & Ruggles, 2020). In general, the impact of the noise introduced via differential privacy decreases as the population in question increases towards infinity.

The final Census Bureau Demonstration Product reflects the official epsilon used in the August release of redistricting data. In preparing this data, the Census used an epsilon of 19.6, 17.14 for persons and 2.47 for housing units (U.S. Census Bureau, 2021a). We employ this data in concert with the comparable public Summary File 1 (SF1) tables for 2010 and in relation to the 2000 Decennial Census Summary File 1 100% Data Tables for 2000. We use this data to establish a time-consistent county-level database and, as necessary, collapse counties in Alaska, Virginia, and Colorado which experienced boundary changes between 2000 and 2010 into time consistent geographic units.

We evaluate two outcomes in our analysis. We begin by comparing both the absolute and relative differences between the existing 2010 summary files (SF) and the 2010 differential privacy files (DP). We compare the absolute differences by subtracting SF counts from the DP counts and taking the absolute

value. Our relative comparison is a ratio of DP to SF counts (DP/SF). We then expand our analysis and turn to the 2000-2010 county-level growth rate for the overall population and key ethnoracial groups.

To assess growth rates over the study period and discern how growth estimates vary between the SF and DP data, we divide the population counts from both versions of the 2010 counts by the 2000 decennial counts. To compare these two estimates, we rely on rate ratios wherein the DP growth rate is divided by the SF growth rate. If the growth rate ratio is greater than 1, then the DP growth rate is greater than the one produced with the SF data. Contrastingly, a growth rate ratio of less than 1 indicates that the growth rate estimated with DP data is less than the rate estimated using data from the SF. We evaluate estimates across the total population and four ethnoracial groups with a dominant, yet often regional, presence across rural America. These include non-Hispanic white, non-Hispanic Black, Hispanic or Latino/a, and non-Hispanic American Indian.

It should be made clear that the relative 2010 differences (e.g. count ratios) and growth rate ratios are exactly equivalent due to the fact that both the SF and DP count ratios are divided by the same divisor—the 2000 SF count. However, to highlight that the new disclosure avoidance technique not only impacts current population counts, but also impacts our understanding of regional and ethnoracial growth, we felt it was important to present and compare discrepancies in the context of growth rates. Growth rates are frequently used by demographers, journalists, and planners to understand what is happening across the country. For example, the recently released census dashboard prominently displays county-level growth rates based on the new 2020 DP for the whole country (U.S. Census Bureau, 2021b). Thus, although the bottom panel of Figure 1 and the top panel of Figure 2 are generally identical,¹ we felt it prudent to discuss our analysis in the framework of population growth due to the frequency with which growth, or a lack thereof, is invoked in rural public policy discussions.

To compare estimates across the rural-urban continuum, we generate each set of estimates for all categories of the Rural-Urban Continuum Codes (RUCC). The RUCC classification, developed by the USDA Economic Research Service, is a two-dimensional scheme for classifying U.S. counties (USDA, 2013). The scheme has nine levels differentiated by both population size and adjacency to metropolitan areas and the codes are as follows:

- 1 – Counties in metropolitan areas of 1 million or more residents;
- 2 – Counties in metropolitan areas with a population of 250,000 to 1 million;
- 3 – Counties in metropolitan areas with a population less than 250,000;
- 4 – Nonmetropolitan counties with an urban population of 20,000 or more which are adjacent to a metropolitan area;
- 5 – Nonmetropolitan counties with an urban population of 20,000 or more which are not adjacent to a metropolitan area;

¹ Due to population counts of zero, the calculation of ratios of relative discrepancies, growth rates, and subsequent growth rate ratios resulted in a number of inestimable rates and rate ratios due to the presence of zeroes in denominators. These observations have been treated as missing for this analysis and include 1 county for Hispanic or Latina/o, 33 counties for non-Hispanic Black, and 7 counties for non-Hispanic American Indian for relative differences in 2010 statistics. For growth rate ratios this includes 1 county for Hispanic or Latina/o, 74 counties for non-Hispanic Black, and 13 counties for non-Hispanic American Indian.

- 6 – Nonmetropolitan counties with an urban population of 2,500 to 19,999 which are adjacent to a metropolitan area;
- 7 – Nonmetropolitan counties with an urban population of 2,500 to 19,999 which are not adjacent to a metropolitan area;
- 8 – Nonmetropolitan counties with either a completely rural population or an urban population less than 2,500 which are adjacent to a metropolitan area;
- 9 – Nonmetropolitan counties with either a completely rural population or an urban population less than 2,500 which are not adjacent to a metropolitan area (USDA, 2013).²

In our analysis of growth, we rely on the RUCC classifications generated from the 2000 census (USDA, 2013). Anchoring rurality at the start of the study period is essential due to selection effects imposed by county reclassification over time due to growth (Brooks et al., 2020). Finally, we also evaluate growth rate ratios across regions. To do so, we rely on the nine census-defined regional divisions: Pacific, Mountain, West North Central, West South Central, East South Central, South Atlantic, Middle Atlantic and New England (U.S. Census Bureau, 1994).

Results

The absolute and relative variation between the 2010 SF and DP statistics is presented in Figure 2. Here we see that the total population counts are generally preserved by the differential privacy algorithm, a trend we will see throughout our results. However, when we turn to population subgroups, this accuracy depreciates considerably. At an absolute level, the largest discrepancies are observed for non-Hispanic white and Hispanic or Latino/a populations in the most urban counties. Inversely, we see that the relative accuracy of estimates decreases across the rural-urban continuum for all groups except for the non-Hispanic white population. Although there is clear rural-urban variation, it should also be noted that even in the most urban counties, all non-white subpopulations still had a notable level of relative variation between SF and DP statistics. In sum, when comparing SF and DP counts for 2010, we see an inverse relationship where the greatest absolute variation is in the most urban counties, but the greatest relative variation is found in the most rural counties.

(Figure 2)

As would be expected due to the equivalence between growth rate ratios and the relative counts in Figure 2, we find the same results for growth rate ratios in Figure 1. We observe little variation between SF and DP growth estimates for the total and non-Hispanic white population across the entire rural-urban continuum. However, as we move into other ethnoracial subgroups, there is major variation between DP and SF growth estimates and a large number of observations occur well outside 1.5 times the interquartile range of the upper and lower quartile—represented by dots. Further, this variation visibly increases as we

² A metropolitan county is defined by the Office of Management and Budget (OMB) as a county with either a core population of at least 50,000, or that is connected to a core metropolitan county by greater than 25% of commuting (OMB, 2010). Please see <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/documentation/> for a map of RUCC distribution across the United States.

move to greater degrees of rurality. The dispersion of estimates is similar for non-Hispanic Black and Hispanic or Latino/a growth estimates, and the most severe for non-Hispanic American Indians.

In the bottom panel of Figure 1, we remove values outside 1.5 times the interquartile range below lower quartile and 1.5 times the interquartile range above the upper quartile to allow for a closer examination of the variation around the preferred rate ratio value of 1. As can be seen, the rate ratios for total population and non-Hispanic white are remarkably clustered around 1.0 across the entire range of rurality. However, this is dramatically altered in the case of our three ethnoracial minority groups of interest. Particularly for the four most-rural RUCC classifications (RUCC 6 – 9), we see an interquartile range spanning much greater values than observed in the total population or non-Hispanic white context. It is important to note that although we do see heightened variation for ethnoracial minorities in rural counties, the *average* rate ratio is still between 0.92 and 1.02 in all cases (analysis not presented here). Thus, if we were to only evaluate the success of differential privacy via measures of central tendency, and not measures of relative discrepancies as we do here, we would likely come away with a false sense of success for the differential privacy algorithm.

When looking regionally, we see significant regional variation for all three ethnoracial minority groups (Figures 3 through 5). The way regions vary is in part reflective of the regional distribution of populations. Non-Hispanic Black populations had the greatest discrepancies in areas with a lower relative share of non-Hispanic Black residence including West North Central, West South Central, and Mountain. Comparatively, Hispanic or Latino/a populations were the most variable in West North Central and the very rural counties of Mountain. Finally, non-Hispanic American Indian populations had high levels of discrepancies in many regions, including West North Central, South Atlantic, West South Central, and East North Central. This regional variation shows that the differential privacy algorithm is failing in the areas where we likely most need it to succeed—in areas where ethnoracial minorities are least represented. In these areas, we need to ensure our population counts are correct for both ensuring due political process as well as proper demographic understanding.

[Figure 3 here]

[Figure 4 here]

[Figure 5 here]

Conclusions

In this brief, we have shown that the 2020 U.S. Census Bureau differential privacy algorithm fails to accurately capture county-level population totals and population growth across the rural-urban continuum for all groups except the total and non-Hispanic white population. This poses serious issues for future demographic research and public policy. Even if univariate distributions and measures of central tendency are preserved with this disclosure avoidance approach, bivariate and multivariate relationships will clearly be altered. Our results show data produced under the final global privacy-loss budget of 19.6 cannot accurately capture even a simple demographic question, that being county-level population growth for standard population subgroups. We have used the county in our analysis due to its importance to both policy and scholarship on rural America, but it is likely the discrepancies are much higher for non-white and rural populations at all smaller levels of geography where the absolute population base will be even

smaller. For example, analysts in Virginia recently argued that the block level data in their state “cannot be accepted as fact” (Hafner, 2021).

Our findings raise serious questions about the validity of the differential privacy approach, especially when considering that the final epsilon of 19.6 is far greater than conventional differential privacy wisdom would support. For example, one of the coinventors of differential privacy, Frank McSherry, stated that an epsilon greater than 1.0 represents a serious privacy compromise and an epsilon of 14 was relatively pointless (Greenberg, 2017). Thus, it appears that the final algorithm may ultimately be introducing an unacceptable level of noise to the data for small populations, while not even generating the confidentiality the Census Bureau argues it provides.

The only ethnoracial subgroup the algorithm appears to be currently capturing accurately is non-Hispanic whites, thus creating a situation where white populations, in rural and urban areas, are afforded a greater level of accuracy in the 2020 census than other ethnoracial groups. This racial inequality is complemented by a rural inequality, wherein estimates for rural non-white populations are extremely prone to discrepancies between the old and new disclosure avoidance techniques. When turning to more complex questions such as intersectional hardship, poverty, migration, or unemployment—which all involve groups with small population representation—the new disclosure avoidance approach is likely to fail. To be clear, this is not an issue that only effects rural or non-white populations, but likely impacts all small-n groups in the United States. This is a fundamental issue of applying a differential privacy approach—which is far more effective when n approaches infinity—to a form of data such as the United States Census where accurate small- n population counts are essential.

Even in the prior method of disclosure avoidance—which relied on suppression, top and bottom coding, and limited amounts of data swapping—geographic areas with very small populations were more likely to need protection and disclosure avoidance. However, the new approach represents a significant deviation from prior practice and appears to introduce a far greater level of unequally distributed discrepancies. Small populations in the United States do not have any less of a right to accurate representation within the Census, especially at a spatial level as politically salient as the county. As the U.S. Census Bureau acknowledges, the balance between accuracy and disclosure avoidance is a tightrope (Abowd & Schmutte, 2019). However, as it currently stands it appears we are sacrificing important levels of accuracy for fears of disclosure.

References

- Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1), 171–202. <https://doi.org/10.1257/aer.20170627>
- Bolin, J. N., Bellamy, G. R., Ferdinand, A. O., Vuong, A. M., Kash, B. A., Schulze, A., & Helduser, J. W. (2015). Rural Healthy People 2020: New Decade, Same Challenges. *Journal of Rural Health*, 31(3), 326–333. <https://doi.org/10.1111/jrh.12116>
- Brooks, M. M., Mueller, J. T., & Thiede, B. C. (2020). County Reclassifications and Rural-Urban Mortality Disparities in the United States (1970-2018). *American Journal of Public Health*, 110(12), 1814–1816. <https://doi.org/10.2105/AJPH.2020.305895>
- Brown, D. L., & Schafft, K. A. (2011). *Rural People and Communities in the 21st Century: Resilience and Transformation*. Cambridge, UK: Polity Press. https://doi.org/10.1111/ruso.12057_2
- Cosby, A. G., McDoom-Echebiri, M. M., James, W., Khandekar, H., Brown, W., & Hanna, H. L. (2019).

- Growth and persistence of place-based mortality in the United States: the rural mortality penalty. *American journal of public health*, 109(1), 155-162.
- Crosby, R. A., Wendel, M. L., Vanderpool, R. C., & Casey, B. R. (2012). *Rural Populations and Health: Determinants, Disparities, and Solutions*. John Wiley & Sons.
- Gagnon, D. J., & Mattingly, M. J. (2018). Racial/Ethnic test score gaps and the urban continuum. *Journal of Research in Rural Education (Online)*, 33(2), 1-16.
- Greenberg, A. (2017). How one of Apple's key privacy safeguards falls short. *Wired*. Published September 15, 2017. Accessed at <https://www.wired.com/story/apple-differential-privacy-shortcomings/>, Accessed on August 19, 2021.
- Hafner, K. (2021). Virginia's new census data is distorted at local levels, analysts say: 'It can't be accepted as fact'. *The Virginia-Pilot*. August 18, 2021.
- Hauer, M. E., & Santos-Lozada, A. R. (2021). Differential Privacy in the 2020 Census will distort COVID-19 rates. *Socius*.
- Isserman, A. M., & Westervelt, J. D. (2006). 1.5 million missing numbers: Overcoming employment suppression in County Business Patterns data. *International Regional Science Review*, 29(3), 311–335. <https://doi.org/10.1177/0160017606290359>
- James, W., & Cossman, J. S. (2017). Long-term trends in Black and White mortality in the rural United States: Evidence of a race-specific rural mortality penalty. *The Journal of Rural Health*, 33(1), 21–31.
- Johnson, K. M., & Lichter, D. T. (2008). Natural increase: A new source of population growth in emerging hispanic destinations in the United States. *Population and Development Review*, 34(2), 327–346. <https://doi.org/10.1111/j.1728-4457.2008.00222.x>
- Johnson, K. M., & Lichter, D. T. (2010). Growing diversity among America's children and youth: Spatial and temporal dimensions. *Population and Development Review*, 36(1), 151–176. <https://doi.org/10.1111/j.1728-4457.2010.00322.x>
- Kenny, C. T., Kuriwaki, S., McCartan, C., Roseman, E., Simko, T., Imai, K. (2021). The impact of the U.S. Census Disclosure Avoidance System on redistricting and voting rights analysis. *Working Paper*. <https://alarm-redist.github.io/posts/2021-05-28-census-das/Harvard-DAS-Evaluation.pdf>
- Lee, B. A., & Sharp, G. (2017). Ethnoracial diversity across the rural-urban continuum. *The ANNALS of the American Academy of Political and Social Science*, 672(1), 26-45.
- Lichter, D. T. (2012). Immigration and the new racial diversity in rural America. *Rural sociology*, 77(1), 3-35.
- Liebler, C. A., & Ortyl, T. (2014). More Than One Million New American Indians in 2000: Who Are They? *Demography*, 51(3), 1101–1130. <https://doi.org/10.1007/s13524-014-0288-7>
- Manson, S., Schroeder, J., Van Riper, D., Kugler, T. & Ruggles, S. (2021). IPUMS National Historical Geographic Information System: Version 15.0 [Database]. <https://doi.org/http://doi.org/10.18128/D050.V14.0>
- Mervis, J. (2019). Can a set of equations keep U.S. census data private? *Science*. <https://doi.org/10.1126/science.aaw5470>
- Mueller, J. T., McConnell, K., Burow, P. B., Pofahl, K., Merdjanoff, A. A., & Farrell, J. (2021). Impacts of the COVID-19 pandemic on rural America. *Proceedings of the National Academy of Sciences of the United States of America*, 118(1), 1–6. <https://doi.org/10.1073/pnas.2019378118>
- Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. (2019). Differential Privacy and Census Data: Implications for Social and Economic Research. *AEA Papers and Proceedings*, 109, 403–408. <https://doi.org/10.1257/pandp.20191107>
- Santos-Lozada, A. R. (2021). Changes in census data will affect our understanding of infant health. *Socius*.
- Santos-Lozada, A. R., Howard, J. T., & Verdery, A. M. (2020). How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences*, 1–8. <https://doi.org/10.1073/pnas.2003714117>
- Schneider, M. (2021, May 27). Census Bureau's use of "synthetic data" worries researchers. *The*

- Washington Post*. Retrieved from https://www.washingtonpost.com/national/census-bureaus-use-of-synthetic-data-worries-researchers/2021/05/27/ac061d04-bf02-11eb-922a-c40c9774bc48_story.html
- Thiede, B., Kim, H., & Valasik, M. (2018). The spatial concentration of America's rural poor population: A postrecession update. *Rural Sociology*, 83(1), 109-144.
- Tickamyer, A., Warlick, J., & Sherman, J. (Eds.). (2017). *Rural poverty in the United States*. Columbia University Press.
- U.S. Census Bureau. (1994). *Geographic Area Reference Manual (GARM)*. U.S. Department of Commerce, Economics and Statistics Administration.
- U.S. Census Bureau. (2021a). Census Bureau sets key parameters to protect privacy in 2020 census results. Census Bureau release number CB21-CN.42. Accessed at <https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html>, Access on August 19, 2021.
- U.S. Census Bureau. (2021b). 2020 Population and Housing State Data. Accessed at <https://www.census.gov/library/visualizations/interactive/2020-population-and-housing-state-data.html>, Accessed on August 19, 2021.
- USDA. (2013). Rural-Urban Continuum Codes. Retrieved February 2, 2020, from <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx>
- Van Riper, David; Kugler, Tracy and Ruggles (2020). Disclosure Avoidance in the Census Bureau's 2010 Demonstration Data Product. In *Privacy in Statistical Databases* (Eds. Josep Domingo-Ferrer and Krishnamurty Muralidhar).
- Winkler, R. L., Butler, J. L., Curtis, K. J., Egan-Robertson, D. (2021). Differential privacy and the accuracy of county-level net migration estimates. *Population Research and Policy Review*.
- Zayatz, L. (2007). Disclosure avoidance practices and research at the U.S. Census Bureau: an update. *Journal of Official Statistics*, 23(2), 253–265.

Figures

Figure 1. County-level growth rate ratio for ethnoracial groups in the United States comparing rates produced using the demonstration product and the summary file. Top panel includes all outliers and bottom panel suppresses all outer values—values greater than 1.5 times the interquartile range above or below the first and third quartile, respectively. This includes 622 for total population, 479 for non-Hispanic white, 814 for non-Hispanic Black, 687 for Hispanic or Latino/a, and 542 for non-Hispanic American Indian. Large numbers of exclusions are due to an extremely tight interquartile range. Counties are divided by RUCC classifications from 1 – most urban to 9 – most rural, see text for complete RUCC definitions.

Figure 2. Visual display of the absolute variation between summary file and differential privacy estimates (top) as well as the relative variation displayed as a ratio (bottom) by RUCC classifications from 1 – most urban to 9 – most rural, see text for complete RUCC definitions.

Figure 4. County-level growth rate ratio for Hispanic or Latino/a populations based on US Census regional divisions. Counties are divided by RUCC classifications from 1 – most urban to 9 – most rural, see text for complete RUCC definitions.

Growth Rate Ratio (DP/SF)

Declarations

Funding: Alexis R. Santos-Lozada is funded by the Social Science Research Institute and PRI at the Pennsylvania State University. PRI is supported by a grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (P2CHD041025). Santos-Lozada is also funded by a Diversity Supplement from the National Institute on Aging through the Interdisciplinary Network on Rural Population Health and Aging (R24-AG065159 and R24-AG065159-03S1).

Conflict of Interest/Competing Interests: The authors have no conflict to disclose.

Availability of data and materials: The data utilized in our study is available through the National Historical Geographic Information System (NHGIS, <https://www.nhgis.org/>).

Code Availability: Our code is available upon request by interested parties. Requests for our code should be sent to the corresponding author.

Authors' contribution: Mueller and Santos-Lozada had full access to all data and take full responsibility for the integrity of the data and the accuracy of the data analysis. Mueller and Santos conceptualized and designed the study, acquired, and analyzed the data. Both authors contributed to the data interpretation. Mueller drafted the original manuscript and the revised versions. Santos contributed to all versions of the manuscript.

Ethics approval: Because these data are country-level aggregate counts, this study is not considered to be research involving human subjects as defined by U.S. regulation (45 CFR 46.102(d)).

Consent to Participate: Not applicable

Consent for Publication: All authors have read and approved this manuscript for publication.

Acknowledgments

The authors are thankful to the IPUMS and NHGIS for making the materials required for this publication accessible through their platforms.