Rural Population Research is Constrained by Census Data Availability and Reliability

J. Tom Mueller and Alexis R. Santos-Lozada

Our understanding of the problems facing rural America, as well as solutions to those problems, is fundamentally tied to the ability of researchers to access accurate data. Data on rural people and places are often unreliable, suppressed, or not collected at all. Although this has been an issue for decades, it has recently gotten worse. With the 2020 Census, the U.S. Census Bureau has introduced a new approach for ensuring privacy called *differential privacy*—an approach that introduces more error into the population estimates for smaller population groups, such as rural, Latino/a, American Indian, and Black populations.[1]

KEY FINDINGS

- Rural data are heavily suppressed by the federal government.
- The data that are available are often unreliable for rural counties and certain racial/ethnic populations.
- New differential privacy policies make public Census data inaccurate for rural Latino/a, American Indian, and Black populations.

This brief describes longstanding limitations with federal rural data and then uses the U.S. Census Bureau's differential privacy demonstration data for 2010 to illustrate the discrepancies introduced by the new differential privacy approach.

Our Knowledge about Rural Populations is Limited due to Data Supression and Reliability Concerns

Researchers studying rural America have long lamented the difficulties surrounding rural data. Data on rural labor, economic concerns, mortality, and other dimensions are often suppressed from publicly available files due to privacy concerns. This has led to complex efforts to fill in missing data via computational methods or by linking datasets. Beyond issues of suppression, rural data are also beleaguered by very large standard errors in survey research. This means less accurate estimates for rural statistics derived from sources such as the American Community Survey—the rolling survey which replaced the long-form Census in 2005—because the precision of the estimate is much more noisy. These issues not only make understanding the state of rural America difficult, they also lead to less research on rural populations overall.

Census Bureau Differential Privacy Policies have Made Matters Worse

Although data on rural populations have always been more limited than data on urban populations, prior waves of the Decennial United States Censuses have produced population data on rural America generally deemed as reliable and accurate—at least at the county level. This long-term reliability has been due to the balance struck by the Census Bureau between data accuracy and disclosure avoidance (the federally mandated requirement for the U.S. Census Bureau to protect the confidentiality of respondents).[2] In the past, the Census Bureau has prevented disclosure through suppression, data swapping, and top- and bottom-coding.[3] Unfortunately, this changed with the 2020 Census and the introduction of differential privacy. Differential privacy injects "noise" into estimates, with more noise injected into estimates for smaller populations. This means that estimates are less accurate when population counts are small.[2,3]

Using demonstration data produced by the Census Bureau, Figure 1 presents the ratio between population counts produced by applying the new differential privacy approach to the 2010 Census (DP) and the official 2010 Census Summary File counts (SF). A ratio of 1.0 means the population counts are equal. Values lower or greater than 1.0 indicate a discrepancy. Figure 1 illustrates that county-level population estimates for the total and non-Hispanic (NH) White populations are relatively accurate for both metropolitan and nonmetropolitan counties (i.e., the ratio is very close to 1.0). However, accuracy is significantly diminished in the case of non-White populations, particularly in nonmetropolitan counties, with ratios ranging from 0.06 to 7.0 for nonmetro NH Blacks, 0.54 to 7 for nonmetro Latino/s, and 0.08 to 8.0 for nonmetro American Indians in counties where these populations were present.
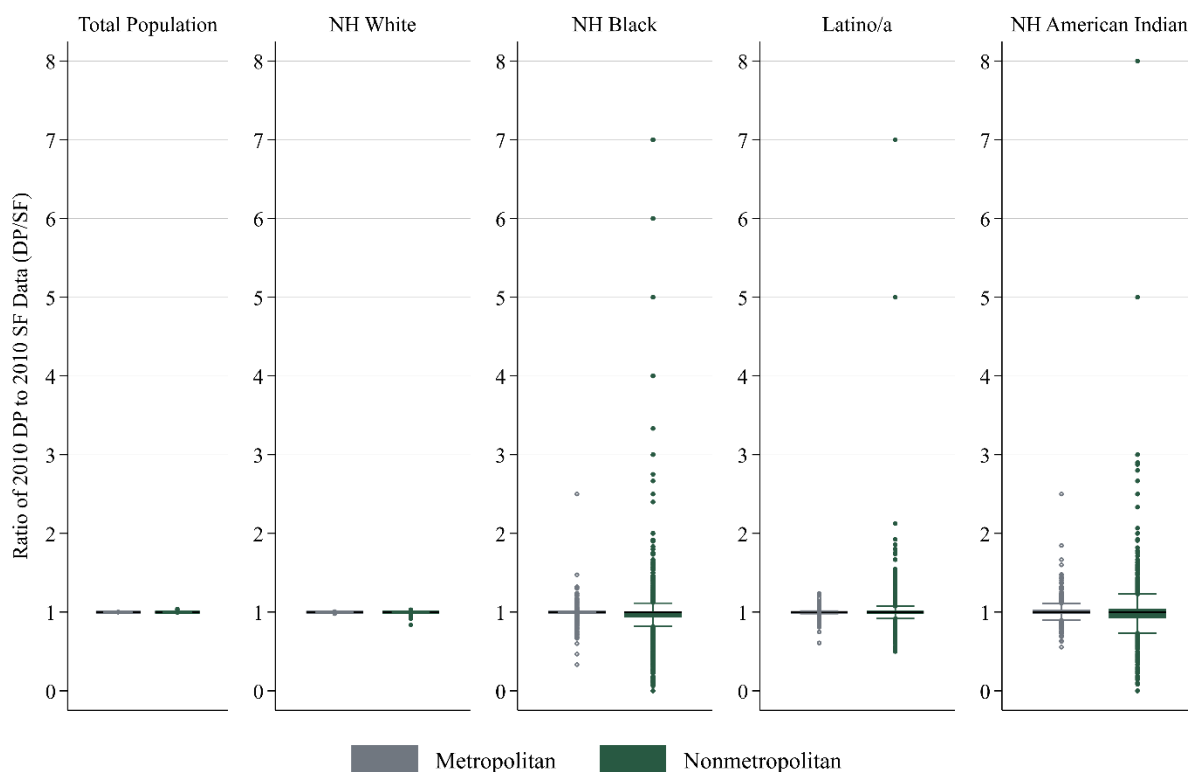
Figure 1. Discrepancies in County-Level 2010 Population Counts due to Differential Privacy
Data Source: DP is Census Differential Privacy Demonstration Data for 2010 Census; SF is 2010 Census Summary Files. All data extracted from IPUMS-NHGIS.[4]

This Problem is Solvable but Requires Political Intervention

Although the Census Bureau is often viewed as an apolitical entity, this has never really been the case. As a result, Congress is reluctant to push back against Census policies and statutory interpretations. At present, the Census Bureau is operating under an interpretation of federal statutes that is overly-cautious and that hinders both scientific research and accessing the data necessary to improve rural wellbeing. Congress should develop a clear set of guidelines for what constitutes disclosure risk and that better balance the tradeoffs between data accuracy and privacy. In addition to providing clear direction to the Census Bureau to ensure that federal data are usable for rural population research, Congress should also direct resources to facilitate more robust data collection efforts in rural areas, including investing in rural oversamples in existing federally funded surveys, such as the Current Population Survey, Health and Retirement Study, and National Health Interview Survey.

Data and Methods

The analysis presented here comes from a recently published article by the authors.[1] Data were extracted from IPUMS-NHGIS and analyzed numerically and visually using Stata V16.0.

References

1. Mueller, J. T., & Santos-Lozada, A. R. (2022). The 2020 US Census Differential Privacy Method Introduces Disproportionate Discrepancies for Rural and Non-White Populations. *Population Research and Policy Review*, 1-14.
2. 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.
3. Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. (2019, May). Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings* (Vol. 109, pp. 403-08).
4. Manson, S., Schroader, J., Van Riper, D., Kugler, T. & Ruggles, S. (2021). IPUMS National Historical Geographic Information System: Version 15.0 [Database].

About the Author

J. Tom Mueller (tom.mueller@ou.edu) is a Research Assistant Professor of Geography and Environmental Sustainability at the University of Oklahoma. **Alexis Santos-Lozada** is an Assistant Professor of Human Development and Family Studies at The Pennsylvania State University.

Suggested Citation: Mueller, J. Tom and Santos-Lozada, Alexis R. 2022. "Rural Population Research is Heavily Constrained by Federal Data Availability and Reliability." Brief # 2022-8. Rural Population Research Network.