

Changes Coming to 2020 Census Data

Julia Milton, Consortium of Social Science Associations (COSSA)

For the data released as part of the 2020 Census, the Census Bureau will adopt a standard called “differential privacy” to fulfill its legal obligation to protect individuals from reidentification. Differential privacy relies on an algorithm that injects precise amounts of random noise into the data until it reaches a desired threshold of obfuscation. It allows the Census Bureau to be

that differential privacy swings the pendulum too far away from usability. They argue that if the quality of the data produced under differential privacy is seen as unacceptably low by data users, they will turn elsewhere to meet their needs—to private, less transparent, potentially less rigorous databases produced by corporations who lack the Census Bureau’s commitment to protecting individuals’ privacy.

However, while users of Census data are indeed concerned about the

disruption differential privacy will have, many agree with the Census Bureau’s assessment that the potential harm successful reconstruction of confi-

dential Census records would have on public trust would be far greater. Further, given the Census Bureau’s determination that the status quo is not an option, differential privacy will be the disclosure avoidance standard for the 2020 Census. The questions remaining are how to implement it in a way that best mitigates data users’ concerns.

The Impact of Differential Privacy on 2020 Data Products

To give data users a better understanding of how differential privacy will affect the data released from the 2020 Census, the Census Bureau released a set of demonstration data products (bit.ly/35Ga0Fp) that apply its differential privacy algorithm to the 2010 Census; it allows researchers and data users to see the impact of the new disclosure avoidance system for themselves. At a December 2019 Committee on National Statistics (CNSTAT) workshop, data users reported on their experiences working with the 2010 demonstration products. Recordings and materials from the workshop are available at bit.ly/38W2F6v.


While many data users reported that the demonstration products were fairly close to the original 2010 data files when analyzing larger geographic entities and populations, several areas of major concern were identified:

- Results become much less accurate at the smallest levels of analysis (e.g., sparsely populated geographic areas or very small minority populations) due to the noise injected by the differential privacy algorithm. Less populous areas tend to gain population and more populous areas tend to lose.
- Analyses using the Census’s primary hierarchical geographic units (nation/state/county/tract group/tract/block group/block) return more accurate results than those that rely on other units of geography (such as county subdivisions or places).
- Concerns remain about how the new data can be used to analyze trends, including what kind of bridge estimates would be produced to make the data comparable over time and whether the Census Bureau plans to devote resources to ensuring that colleagues in other statistical units can perform legally-mandated longitudinal analyses using the new data.
- The Census Bureau has not yet determined how it will produce estimates of uncertainty. Depending on the methodology used, the uncertainty measurements could affect the overall privacy-loss budget, forcing additional tradeoffs to data accuracy elsewhere.

According to the Census Bureau, some types of errors can be addressed without affecting the privacy-loss budget. These include many impossible or implausible results identified during the workshop (e.g., results indicating less than one person per household). Such errors are artifacts of the processing techniques employed *after* the algorithm is run. However, other issues are a result of the noise intentionally introduced by the dif-

ferential privacy algorithm itself and would require accuracy tradeoffs elsewhere to address.

How to Get Involved

The Census Bureau plans to continue making improvements to its 2020 Data Products. Stakeholders can email dcmd2010.demonstraton.data.products@census.gov to share concerns and report problems or issues. Data users are asked to share their source code to ensure Census staff are able to reproduce and resolve the errors encountered. In addition, the Census Bureau plans to continue working through CNSTAT to collect feedback by establishing working groups and holding additional meetings. Information on these follow-on activities has not yet been released, but interested stakeholders should email the CNSTAT study director, Daniel Cork (dcork@nas.org). 

References

- Abowd, John M. 2018. “The U.S. Census Bureau Adopts Differential Privacy.” KDD ‘18 Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK (August 2018): 2867, DOI: 10.1145/3219819.3226070.
- Abowd, John M., and Ian M. Schmutte. 2019. “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices.” *American Economic Review*, 109(1): 171-202.
- Census Bureau, “Disclosure Avoidance and the 2020 Census.” www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html.
- Rodríguez, Rolando A. and Amy D. Lauger. 2019. “Innovating Data Privacy for the American Community Survey.” 2019 ACS Data Users Conference, Washington, DC (March 2019).
- Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. “Differential Privacy and Census Data: Implications for Social and Economics Research.” *AEA Papers and Proceedings*, 109: 403-08.
- Wood, Alexandra, Micah Altman, Aaron Bembeneck, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O’Brien, Thomas Steinke, and Salil Vadhan. Nissim, Kobbi, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembeneck, Mark Bun, Marco Gaboardi, David O’Brien, and Salil Vadhan. 2018. “Differential Privacy: A Primer for a Non-technical Audience.” *Vanderbilt Journal of Entertainment and Technology Law*, 21(1): 209-276.

United States Census 2020

more precise in deciding how much risk it is willing to take to produce useable data and to be more transparent about those tradeoffs. Unlike the Census Bureau’s traditional disclosure avoidance techniques, differential privacy will allow the Census Bureau to be more transparent about the algorithms and their parameters.

To implement differential privacy, the Census Bureau must make a number of value-based policy decisions that will affect how well-protected the information is from the threat of reidentification and how useful it ends up being for data users. The most important of these is the “privacy-loss budget,” which defines the maximum amount of reidentification risk the Census Bureau is willing to allow, in total and across its individual tabulations and geographies.

Concerns about Differential Privacy

Critics of the Census Bureau’s decision to adopt differential privacy argue that it is overzealously interpreting the legal standard it must meet to protect its data and