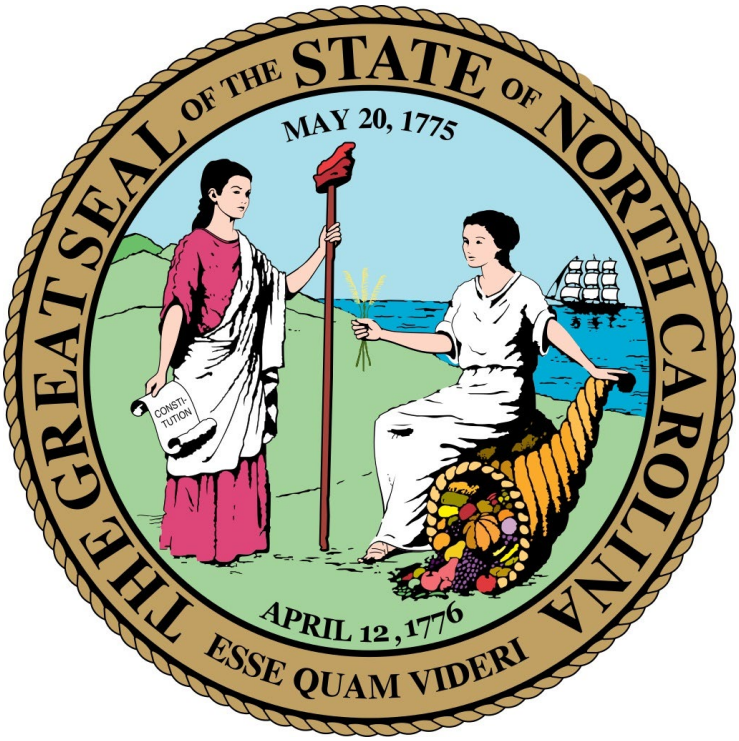
OSBM

Backgrounder

Impact of 2020 Census Differential Privacy Procedures for State Agencies

January 2020
Updated March 3, 2020



Prepared by:

Office of State Budget and Management

Impact of 2020 Census Differential Privacy Procedures for State Agencies

Summary

The US Census Bureau will make major changes to the data products available for the upcoming 2020 Decennial Census and future surveys like the American Community Survey. These changes are related to the Census Bureau's longstanding commitment to confidentiality of individual information and its awareness of the risks to breaches of privacy in an age of greater computing power and the proliferation of commercial and governmental data. Although we do not yet know the full impact, these changes may limit the amount of census data available to state agencies. These changes could impact the ability of agencies to plan, monitor, and distribute resources using decennial census data.

It is imperative that you review how your agency uses decennial census data directly or indirectly to carry on its mission and to understand how these changes may impact you. We provide a very brief overview of differential privacy (with links to more lengthy discussions and videos), the potential impacts to data availability, resources available to analyze the potential impacts, and information on how you can provide feedback to influence Census Bureau decisions on data products they will make available from the 2020 Decennial Census. While differential privacy will be used for the 2020 decennial census, there remain many decisions about how differential privacy will be implemented. We encourage you to provide feedback to the Census Bureau to inform those decisions.

Background

Title 13 (13U.S.C.§9(2007)) mandates the Census Bureau to not: "…use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports." For this reason, the Census Bureau releases data only in aggregated forms and uses disclosure avoidance techniques to limit the release of information in the aggregate data that might risk the disclosure of information about individual respondents. Over the years, the Census Bureau has used different techniques including:

- data suppression (not releasing tables for small populations and smaller geographic areas),
- data "swapping" (matching households in different geographic areas and swapping selected characteristics), and
- "blank and imputing" records (changing records to a blank value and then imputing values for all blank cells using statistical models).

In the presence of greater and more diffused computing power coupled with the access of other publicly and privately available data, these techniques are no longer sufficient for the Census Bureau's disclosure avoidance mandate. Beginning with the release of the 2020 Census data (in 2021) the Bureau will use a new disclosure avoidance procedure called differential privacy (DP). These same procedures will eventually be used in other Census Bureau products such as the American Community Survey.

What is Differential Privacy? [1]

All statistical tables leak some information, but when combined with other tables this information may risk the disclosure of an individual's information. The differential privacy (DP) process will inject artificial noise in the data to decrease the risk of exposing individual information. Unlike previous methods of disclosure avoidance, DP provides a statistical measure – epsilon – so that users can understand how "noisy" the data are. Epsilon is a mathematically derived variable that represents the trade-off between noise in the data and disclosure risk.

Although DP is the disclosure avoidance policy that will be used for the 2020 Census, the Bureau is still working on the ways in which this policy will be implemented. Below is a discussion of key things that the Census Bureau is currently determining.

1. Where should the global and local privacy loss budget be set?

Privacy loss budgets determine the risk of individual information disclosure the Census Bureau is willing to assume at different levels, and budgets can be set by determining a value for epsilon. At higher levels of epsilon, the data are more accurate but the risk of disclosing individual information is higher (thus, you have "privacy loss"). At lower levels of epsilon, there is more noise in the data (thus the data are less accurate) but your risk of disclosing individual information is smaller. The Census Bureau will have to set a privacy loss budget at the global level, but also have to set local privacy loss budgets for each table and geographic level, which must not exceed the global budget.

2. Which variables will remain invariant and at what geographic levels?

In previous disclosure avoidance procedures, the Bureau determined that certain items should not be changed. These items were called "invariants." In 2000 and 2010, total population, total voting age population, total housing units, total occupied housing units, and group quarters were invariant at all geographic levels.

[1] This is a very brief overview of differential privacy. For more thorough discussions of DP see the Census Bureau page about disclosure avoidance; and a presentation by David Van Riper of IPUMS. Additionally, for information about the trade-offs and concerns of stakeholders see: Danah Boyd "Balancing Data Utility and Confidentiality in the 2020 US Census."

3. What products and tables should be made from the 2020 census?

Given the decisions on privacy budgets and invariants, the Bureau will need to determine what products and tables can be released for the 2020 Census. So far, the Bureau has determined it will produce these products for the 2020 Census:

- Apportionment Product,
- Redistricting File (PL 94-171),
- Demographic Profile,
- Demographic and Housing Characteristics File (replaces Summary File 1), and
- Congressional District Demographic and Housing Characteristics File (replaces Congressional District Summary File).

The Bureau has stated that they are not able to produce the following products at this time, but are continuing their research to determine if they can be produced for 2020:[2]

- Detailed race/Hispanic origin tables and some household tables from Summary File 1,
- Summary File 2,
- American Indian and Alaska Native Summary File, and
- Public Use Microdata Sample (PUMS) File.

There are no plans to produce these products:

- Summary of Population and Housing Characteristics (CPH-1), and
- Population and Housing Unit Counts Report Series (CPH-2).

All these decisions that have yet to be made will result in trade-offs between data accuracy, data privacy, and data usefulness. For this reason, the North Carolina State Demographer and the Governor's Census Liaison are reaching out to data users in state agencies so that you can:

(1) Prepare for the potential impacts to your own data use and policy analyses, and
(2) Review your own Census data needs, prepare use case scenarios, and send feedback to the Census Bureau.

Preparing Your Agency for the Impacts of Differential Privacy

The Census Bureau continues to provide information and resources so that demographic data users are prepared for these changes and to ensure that census data remains relevant and useful. We have attached a 2020 to 2010 Table Crosswalk file prepared by the Bureau. This will allow you to see which tables and variables present in the 2010 data products are currently proposed for inclusion in the 2020 products (and in which tables). Please review these crosswalks closely with an eye on those items that you use directly or indirectly for your own policy analyses, funding formulas, or reporting. We, and our colleagues across the country,

[2] See a full discussion here.

have already expressed our concerns about the lack of some household characteristic information, e.g. household size, families, some household types.

In addition, the Census Bureau has made available a 2010 demonstration product that can be used to analyze the potential impact of differential privacy. The demonstration applies DP to the 2010 Decennial Census data and tables from the proposed PL 94-171 and the Demographic and Housing Characteristic products have been produced. Comparing the originally produced 2010 Census data to what that data would look under differential privacy will allow you to understand the potential impacts for your data use. (Note that there are some limitations in this comparison because the original 2010 Decennial Census data included disclosure avoidance procedures such as data swapping, blank and imputing, and data suppression.)

Two other institutions have re-formatted these demonstration products so that users can more readily analyze these data. You can access them at IPUMS or Cornell (each have tried to make it easier for users to access, but have formatted the information in different ways).

OSBM Analysis of 2020 Demonstration Data

OSBM has prepared a brief analysis comparing the 2020 Demonstration Data product (which uses 2010 decennial census data and incorporates DP) with the original summary file data from the 2010 Census. The tables below summarize the differences for various geographic areas. Generally, the larger the population of a geographic area, the smaller the difference between the original and DP data.

Table 1 shows very small mean absolute differences in State House and Senate district totals. There or no house districts with an absolute percent difference larger than 1%, and no Senate districts with absolute percent differences larger than half a percent. American Indian Areas are geographic areas defined by the Census Bureau in consultation with State and Federally recognized tribes. These areas range in population size from 2,113 (Waccamaw Siouan) to 490,899 (Lumbee).[3] The mean absolute percent difference for these eight areas is less than 2%.

Table 1. Summary Comparisons: Demonstration (DP) Data vs. 2010 Original Counts - State House and Senate Districts and American Indian Areas

Type of Area	Number of Areas	Mean Absolute Percent Difference	Mean Absolute Difference	Minimum Absolute % Difference	Maximum Absolute % Difference
Am. Indian Area	8	1.8	223	0.1	6.2
State House	120	0.2	182	0.0	1.0
State Senate	50	0.1	171	0.0	0.4

[3] Total population includes all population living within these areas and not just American Indian populations.

For counties (Table 2), the differences are minimal – mean absolute percent difference for all counties is 0.2% (a mean absolute difference of 80). Again, mean absolute percent differences increase as the size of the county decreases, but even for the smallest sized counties the maximum absolute percent difference is less than 2%.

Table 2. Summary Comparisons: Demonstration (DP) Data vs. 2010 Original Counts – Counties by County Population Size in 2010

Population Size	Number of Counties	Mean Absolute Percent Difference	Mean Absolute Difference	Minimum Absolute % Difference	Maximum Absolute % Difference
< 25,000	26	0.5	70	0.0	1.6
25,000 to 49,999	21	0.2	67	0.0	0.5
50,000 to 99,999	26	0.1	51	0.0	0.2
100,000+	27	0.1	126	0.0	0.1
Total	**100**	**0.2**	**80**	**0.0**	**1.6**

For municipalities (defined as incorporated places in census jargon), the analyses show the same patterns regarding population differences – the largest absolute percent differences are for places of less than 500 people (Table 3). Note, 60% of municipalities in North Carolina had populations in 2010 of less than 2,500. An analysis of the mean percent differences (not reported here) show that DP (at least as applied in the demonstration products) leads to increases in numbers for smaller populated places (as percent differences are positive) and decreases the numbers for the largest places (as percent differences are negative). Overall, for the 552 municipalities in 2010, the mean absolute percent difference was 8.8%, with Grandfather Village having the largest absolute percent difference of 220% (going from an original count of 25 people to 80 people in the demonstration data). Using these data, there is a net increase of 12,284 in the populations of all places less than 10,000 people, and a net decrease of 33,061 in the population of all places of 10,000 people or more. These sorts of deviations in census counts could lead to erroneous conclusions regarding the population trends for smaller communities (assuming that the same rules regarding invariants used in the demonstration products remain for the 2020 Census data release).

Table 3. Summary Comparisons: Demonstration (DP) Data vs. 2010 Original Counts –Places by Place Population Size in 2010

Population Size	Number of Places	Mean Absolute Percent Difference	Mean Absolute Difference	Minimum Absolute % Difference	Maximum Absolute % Difference
<500	126	25.0	53	0.0	220.0
500 to 1,000	93	9.0	63	0.2	38.9
1,000 to 2,499	111	4.4	68	0.1	21.8
5,000 to 9,999	90	2.5	88	0.0	9.7
10,000 to 24,999	50	1.7	117	0.0	5.1
2,500 to 4,999	48	1.1	167	0.1	2.7
25,000 to 49,999	18	1.2	392	0.1	2.9
50,000+	16	0.7	1,136	0.0	1.6
Total	**552**	**8.8**	**122**	**0.0**	**220**

Additionally, we analyzed housing occupancy status. In the original summary file data, there was only one municipality with 100% housing unit occupancy. In contrast, there were 104 municipalities with 100% housing occupancy in the demonstration data (19% of municipalities), and all of these municipalities had populations of less than 10,000.

Based upon their own analyses and the feedback of data users, the Census Bureau is adjusting the ways that the disclosure avoidance system will be implemented for the 2020 census data. They are committed to limiting the amount of noise injected into population counts for political entities such as municipalities and counties as well as for key variables used in population estimates (such as housing unit occupancy, average household size, age groups, race/ethnicity). In addition, they are considering increasing the privacy loss budget from that used in the demonstration product.

Opportunities for Feedback

Although the Census Bureau has already decided that differential privacy will be the disclosure avoidance policy used for the 2020 decennial census, there remain many decisions regarding its implementation, as outlined above. We encourage you to review the table crosswalk and to use the 2010 demonstration product to analyze the potential impacts to your agency. We encourage you to submit feedback to the Census Bureau by emailing to: dcmd.2010.demonstration.data.products@census.gov.

The most effective feedback will be that which demonstrates how you use decennial census data and how the new procedures may or may not limit your ability to provide policy analyses or allocate resources. In addition, the most helpful suggestions should acknowledge the

delicate balance that the Census Bureau must make between accuracy and privacy such as: “we can accept less accurate block level data if that improves the accuracy of block group or tract level data” or “we can accept accuracy levels of no less than w/x/y/z percent for county/place/tract/block level data.” If you do send feedback, please copy me at: State.Demographer@osbm.nc.gov. This would allow OSBM staff to include that information in our interactions with Census Bureau personnel.

Useful Links

U.S. Census Bureau 2020 Demonstration Data Products https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html

Disclosure Avoidance and the 2020 Census
https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

IPUMS Discussions, Resources, and Data Regarding Differential Privacy
https://ipums.org/changes-to-census-bureau-data-products

Data & Society Paper on Differential Privacy https://datasociety.net/output/balancing-data-utility-and-confidentiality-in-the-2020-us-census/

Michael E. Cline, PhD
State Demographer
Demographic and Economic Analysis Section
North Carolina Office of State Budget & Management
Michael.cline@osbm.nc.gov
984-236-0686

Bob Coats
Governor’s Census Liaison
Demographic and Economic Analysis Section
North Carolina Office of State Budget & Management
Bob.Coats@osbm.nc.gov
984-236-0687