TEXAS DEMOGRAPHIC CENTER

Evaluating the Impact of Differential Privacy Using the Census Bureau's 2010 Demonstration Data Products Released on June 8, 2021

The U.S. Census Bureau will adopt a new approach, known as *differential privacy,* in its release of the 2020 census products. This new approach, while deemed to be the best way to protect respondents' identity and privacy, will result in a significant trade-off in data accuracy, potentially making the public data less useful for many scientific, program administration, and policy applications.

To get users involved in the development of the new Disclosure Avoidance System (DAS), the Census Bureau re-constructed the 2010 census data with the implementation of the differential privacy and released them to the public as demonstration data. Users are encouraged to compare the demonstration data with the originally released 2010 data so that they can evaluate the potential impact on their work and provide feedback. Since the release of the baseline demonstration data in 2019, there have been 4 new releases including the most recent one on April 28th, 2021. Each new release has seen improvements in the overall accuracy, thanks to the feedback from users and the collaborative work by the Census Bureau with area experts.

The Texas Demographic Center has been reviewing and conducting analyses on the Census Bureau's 2010 Demonstration Data Products. We created an information sheet after the release of the baseline demonstration data in 2019 to inform our affiliates about DAS development. We highlighted a few key variables that would no longer be available in the 2020 census releases following differential privacy, and we communicated a few of our concerns based on our review of the demonstration data for Texas counties and places.

In this information sheet, we report results from some of our analyses on the DAS data released in April, 2021. We hope that the examples presented here will help our users evaluate the potential impact of differential privacy and they will take the opportunity to provide feedback to the Census Bureau.

Analysis of the 2021.6.8 Demonstration Data

Ahead of the scheduled publication of the 2020 Census P.L. 94-171 redistricting data by September 2021, the Census Bureau released the final set of DAS demonstration data on June 8th, 2021, created to meet the fitness-for-use accuracy targets for the redistricting and Voting Rights Act use cases.

Differential Privacy (DP) and Privacy Loss Budget (PLB)

In order to ensure respondent confidentiality in all data products as required by Title 13 and Title 26, the U.S. Census Bureau developed disclosure avoidance policies that have been implemented in different versions throughout the Bureau's history. More powerful computers, privacy threats, and the availability of personal online data require new safeguards be implemented to avoid disclosure of respondent data. The 2020 Census will use a new, powerful protection system known as differential privacy (DP) to "swap data" and "insert noise" into the data in order to maintain the confidentiality of respondents.

A key concept of the differential privacy is the adoption of the privacy loss budget (PLB), often expressed as epsilon, or e. When e equals 0, it indicates zero privacy loss at the expense of inaccurate and useless data. On the other hand, a higher PLB improves data accuracy but increases the risk of privacy disclosure.

Refer to Formal Privacy Methods for the 2020 Census for details and the formal definitions.

Table 1: Mean Absolute Percent Error of the DAS Demonstration Data for Population 18 and over, Texas House and Senate Districts

Population Group 18+	MAPE PLB 4.5 TX House	MAPE PLB 4.5 TX Senate	MAPE PLB 12.2 TX House	MAPE PLB 12.2 TX Senate
Total	0.1%	0.0%	0.0%	0.0%
Non-Hispanic White	0.0%	0.0%	0.1%	0.0%
Black Alone or in Comination	0.4%	0.1%	0.2%	0.1%
AIAN Alone or in Comination	0.9%	2.6%	0.4%	0.4%
Asian Alone or in Comination	7.7%	1.6%	1.0%	0.6%
PCI Alone or in Comination	8.2%	5.0%	10.2%	9.6%
Some Other Race Alone or in Comination	0.3%	0.1%	1.2%	0.2%
Two or More Race Alone or in Comination	3.5%	0.1%	1.2%	0.7%
Hispanic	2.8%	0.4%	2.3%	0.4%

Source: U.S. Census Bureau, 2010 Census Summary File 1 and 2021.4.28 Demonstration Data

Previous releases have adopted the same privacy-loss budget (PLB) of 4.5 (4.0 for persons and 0.5 for housing units) as the baseline demonstration data for comparison purposes. In the April 2021 release, a new set of data utilizing a PLB of 12.2 (10.3 for persons and 0.9 for housing units) was added as well as one using the original PLB. The higher PLB is anticipated to more closely approximate the final PLB used in the redistricting data.

Accompanying the release of the demonstration data, the Census Bureau also published a detailed summary metrics table that includes a variety of accuracy measures, most of which are at the national level. In this analysis, we focused our attention on Texas and compared both sets of the demonstration data with the corresponding 2010 census data products at different geographic levels, on select characteristics.

Texas Senate and House Districts

Using the block information from the 2010 SF1 data and the DAS demonstration data, we updated the Texas Senate and House Districts with the most recent re-districting information. When comparing the population aged 18 and over for Texas Senate districts, the differences in total population were minimal, with all of the 31 districts less than 0.3 percent in both versions. For the House districts, we found that 8 of the 150 districts had absolute differences greater than half a percent but still less than one percent in the PLB 4.5 data, none exceeding 0.4% in difference.

Table 2: Counties with the Biggest Percent Difference between the 2010 Summary File 1 and the DAS Demonstration Data

County	SF1 Population	Demo Data Population	Abs Percent Difference
PLB 4.5			
Loving	82	96	17.1%
King	286	271	5.2%
Kenedy	416	435	4.6%
Terrell	984	963	2.1%
Motley	1,210	1,191	1.6%
PLB 12.2			
Loving	82	77	6.1%
King	286	290	1.4%
Sterling	1,143	1,135	0.7%
Motley	1,210	1,217	0.6%
Irion	1,599	1,607	0.5%

Source: U.S. Census Bureau, 2010 Census Summary File 1 and 2021.4.28 Demonstration Data

Table 3: Differences in Population by Place Population Size between 2010 Summary File 1 and the DAS Demonstration Data

Population	Number of Places	MAPE PLB 4.5	MAPE PLB 12.2
1-500	513	6.0%	2.6%
501-1,000	260	1.6%	0.7%
1,001-5,000	591	0.9%	0.5%
5,001-10,000	143	0.6%	0.3%
10,001 plus	241	0.4%	0.2%

Source: U.S. Census Bureau, 2010 Census Summary File 1 and 2021.4.28 Demonstration Data

When comparing the population age 18 and over by race/ethnicity for each district, however, we began to see larger differences (Table 1). The overall measure of accuracy, the mean absolute percent error (MAPE), remains minimal for Non-Hispanic Whites and Blacks. But other race/ethnic groups saw larger differences. Notably, MAPE for the Pacific Islander population was 10.2% for the House Districts and 9.6% for the Senate Districts in the PLB 12.2 version.

Counties

We also compared total population of Texas counties. Overall, MAPE for the demonstration data is 0.24% at PLB 4.5 and 0.09% at PLB 12.2. Table 2 lists five counties with the biggest absolute percent differences for the two versions of the demonstration data. As expected, all of the counties listed have relatively small population in 2010 of less than 2,000. While the highest percent difference is 17.1% for PLB 4.5, it is only 6.1% for PLB 12.2. In fact, all counties but two in Texas have absolute difference less than 1% for PLB 12.2.

Places

Table 3 shows findings for Texas places, including the Census Designated Places (CDPs), grouped by size of the population and the calculated mean absolute percent difference (MAPE) for each group. Similarly, the biggest differences were found among places with populations of less than 1,000 at PLB 4.5. It is noteworthy that the 773 places with this population size comprise more than 40 percent of all Texas places. The 12.2 version saw a significant improvement in MAPE, especially for places with smaller population.

Block Groups

To examine block level differences, we utilized the tables created by IPUMS/NHGIS. We extracted information relevant to Texas. Table 4.1 shows selected measures for Texas blocks and Table 4.2 lists measures regarding the race/ethnic compositions of the blocks in Texas.

Our analysis suggests that the newly released demonstration data, especially the version with the larger PLB of 12.2, saw improvements in the overall measures at different geography levels. It is clear,

Table 4.1: Selected Accuracy Measures for the Blocks in Texas

Measure	PLB 12.2 Number	PLB 12.2 Percent	PLB 4.5 Number	PLB 4.5 Percent
Blocks changed from greater than 50% Non-Hispanic White alone to less than 50% Non-Hispanic White alone	25,404	2.8	38,979	4.3
Blocks with population age 0 to 17 but no population ages 18+	6,250	0.7	11,391	1.3
Blocks with population in Summary File 1 but no population in DP file	3,003	0.3	8,244	0.9
Blocks with population in households but no occupied housing units	51,816	5.7	58,122	6.4
Blocks with occupied housing units but no population	4,296	0.5	8,278	0.9
Blocks with more than 15 persons per household	55,309	6.1	62,874	6.9

Source: IPUMS NHGIS, University of Minnesota, www.nhgis.org

Table 4.2: Select Accuracy Measures on Race/Ethnic Composition of the Blocks in Texas

Race	PLB 12.2 Total population	18 years and older	PLB 12.2 Under 18 years	PLB 4.5 Total population	18 years and older	PLB 4.5 Under 18 years
Mean absolute numeric error						
Total population	2	1	1	3	2	2
Non-Hispanic White alone	1	1	0	2	1	1
Black alone or in combination	1	0	0	1	1	1
Hispanic	1	1	1	2	1	1
Asian alone or in combination	0	0	0	1	0	0
Mean absolute percent error						
Total population	12.5	11.8	24.5	19.1	19.4	33.6
Non-Hispanic White alone	19.0	17.8	23.9	27.0	26.8	33.8
Black alone or in combination	31.4	24.3	25.2	34.4	28.4	28.1
Hispanic	30.8	27.2	29.4	36.6	34.4	36.4
Asian alone or in combination	25.1	20.0	16.2	25.9	21.7	16.9
Number of blocks with more than 10% absolute percent error						
Total population	24	23	31	33	32	38
Non-Hispanic White alone	24	22	23	32	31	29
Black alone or in combination	24	19	18	25	21	19
Hispanic	30	27	27	35	32	31
Asian alone or in combination	17	14	10	17	14	10
Number of blocks with more than 5% absolute percent error						
Total population	33	31	36	41	40	41
Non-Hispanic White alone	30	28	26	37	36	30
Black alone or in combination	25	20	18	27	22	19
Hispanic	34	31	29	38	35	33
Asian alone or in combination	17	14	10	17	15	10

Source: IPUMS NHGIS, University of Minnesota, www.nhgis.org

perhaps by design, that differential privacy will disproportionally impact data accuracy of geographies and groups with smaller population size.

Blocks

To examine block level differences, we utilized the tables created by IPUMS/NHGIS. We extracted information relevant to Texas. Table 4.1 shows selected measures for Texas blocks and Table 4.2 lists measures regarding the race/ethnic compositions of the blocks in Texas.

Our analysis suggests that the newly released demonstration data, especially the version with the larger PLB of 12.2, saw improvements in the overall measures at different geography levels. It is clear,

Opportunities for Feedback

Although we presented a variety of measures, this information sheet was not intended as a comprehensive analysis that covers all aspects of the topic. Census data users are encouraged to review the 2020 Census Data Products Planning Crosswalk to determine if variables pertinent to their analyses and research are impacted by differential privacy. Additionally, organizations are encouraged to conduct their own analyses using the 2010 Demonstration Data and compare them to the previously published 2010 Census data. Lastly, we encourage organizations to engage with the Census Bureau and provide input on what their analyses reveal and any other concerns they may have. Feedback for the new release is due by May 28th,

2021 and can be submitted to the Census Bureau by emailing: 2020DAS@census.gov.

The Texas Demographic Center offers assistance to users in navigating information on the DAS system and analyzing the related demonstration data. Questions and data requests can be sent to: tdc@utsa.edu

Useful Links:

2020 Disclosure Avoidance System Updates

https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html

U.S. Census Bureau 2010 Demonstration Data Products:

https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html

Disclosure Avoidance Webinar Series

https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series.html

IPUMS NHGIS Privacy-Protected Census Demonstration Data

https://nhgis.org/privacy-protected-demonstration-data#v20210428

About the Texas Demographic Center

The Texas Demographic Center produces, interprets, and disseminates demographic information to facilitate data driven decision making