Differential Privacy Policy and Its Implications for DC's 2020 Census Data

by Travis Pate

September 2022

District of Columbia
Office of Planning

INTRODUCTION

Those who rely on Census data for small area planning may want to take note of a new privacy policy the U.S. Census Bureau is implementing starting with the 2020 Census. The policy, called differential privacy, is an innovative approach to protecting the public's personal information. Yet, there is a cost to such protections in terms of data accuracy on the ground. Differential privacy policy addresses concerns that bad actors could use the increasing amounts of personal data available through social media and administrative records to identify individuals and households in Census data. If this were to happen, this would be a major breach of trust in the Census Bureau, which is required to keep all personal identifiable information confidential. The Census Bureau has typically applied some form of privacy protection measures to their data products with the purpose of protecting personal information. To maintain confidentiality, differential privacy policy conducts a kind of 'shell game' that moves people's data around from one neighborhood block to another. Furthermore, household characteristics - such as age, race, and, ethnicity – are evaluated as separate pieces and reassembled in different households

This report measures the potential impact of the new differential privacy approach on the District's 2020 Census data by analyzing demonstration data that applied the differential privacy algorithms to 2010 Census data products (total population, race/ethnicity population, total housing, and housing status). This report uses demonstration data released on November 16, 2020 by the Census Bureau. The findings of this analysis show that the differential privacy approach has a varying degree of impact on demographic characteristics based on the size of population groups and the level of geography. Of major concern is that data could potentially be altered to the point of mispresenting small population groups and neighborhood sized geographies, which means that Census data often used for small area planning would be unreliable for the next decade. Note: At the time of writing, the Census Bureau announced that another round of demonstration data would be released which is a reversal of what had been previously stated. After reviewing the demonstration data released on April 28, 2021, the Census Bureau achieved greater accuracy overall while the concerns highlighted in this report remain. The conclusion of this report discusses the new data release further.

BACKGROUND

Differential Privacy History

Title 13 of the U.S. Code prohibits the Census Bureau from publishing any information in a manner that may be used to identify the information provided by any census or survey respondent. In order to produce official data products while meeting this legal requirement to protect confidentiality, the Census Bureau has historically relied upon the application of statistical disclosure avoidance methods to alter the published data sufficiently to mitigate the risk that individual respondent data could be reliably re-identified. From the 1990 Census through the 2010 Census, this process involved the introduction of "noise," or statistical uncertainty into the data via the swapping of entire households' records across geographies. Other procedures include data aggregation combined with data suppression based on table population thresholds, and the injection of synthetic data into tables. Leading up to the 2020 Census, the Census Bureau's Data Stewardship Executive Policy Committee (DSEP) determined that the data swapping methods used in past censuses are no longer sufficient to protect the confidentiality of census records. This is due to growing privacy threats posed by the amount of external data sources that may be used to attempt the re-identification of respondents, and improvements in the computing algorithms that can reconstruct individuals' records from aggregate data. In 2019, DSEP decided that the Census Bureau will use new, mathematically provable, disclosure avoidance techniques for noise injection based on differential privacy for all 2020 Census public data releases.

The Database Reconstruction Theorem, first introduced by Irit Dinur and Kobbi Nissim in 2003, demonstrates that the calculation of any statistic from a confidential data source reveals a tiny amount of private information about the confidential data. If you release too many statistics, at too high a degree of accuracy, then after a finite number of tabulations you will reveal all of the underlying confidential information used to create the tabular summaries. Differential privacy, first conceptualized by Cynthia Dwork (2006), provides a framework for quantifying this leakage of private information, and in doing so, enables its mitigation through the injection of precisely calibrated amounts of noise. Consequently, differential privacy as an approach to disclosure avoidance allows for quantifiable, future-proof privacy guarantees. These guarantees are set through the establishment of a privacy-loss budget (PLB) and its allocation to each tabular summary. Under this approach, any statistic, tabulation, or calculation to be performed against the confidential data will have a certain amount of noise added to it. For the 2020 Census, DSEP will establish a global PLB for all 2020 Census Data Products based on the findings of the demonstration data products and feedback from the State Data Center (SDC) network and other stakeholders. The redistricting data products (used for redrawing Ward boundaries in the District of Columbia and state legislative boundaries in other states) will not be affected by the new differential privacy policy.

Concerns about the New Differential Privacy

According to Census Bureau Chief Scientist John Abowd (2017 & 2018), "all data publication inherently involves some inferential disclosure." Abowd maintains that this is "the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared." It is possible—even likely—that scientists, planners, and the public will soon lose the free access we have enjoyed for the past six decades to reliable public Census Bureau data describing American social and economic change. Abowd also believed that the differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau. In addition, Abowd suggested that by imposing an unrealistic privacy standard, the Census Bureau may be forced to lock up data that are indispensable for basic research and policy analysis and that such data are essential for testing theories of past change, understanding present conditions, and making projections into the future.

Data user communities across the country have also voiced grave concerns about the Census Bureau's differential privacy policy. Based on discussions and letters from SDCs across the country, the planned data distortion will last for the entire decade and carry far reaching implications. The SDC network members stated that while they understand the Census Bureau's objective to balance privacy with data usability and availability, the repercussions and loss of data should be carefully weighed and evaluated relative to the Bureau's responsibilities for privacy protection under Title 13.

Closer to home, the University of Virginia's Weldon Cooper Center for Public Services wrote a letter to the governor of Virginia in 2020 expressing their concerns about the possible impact of the Census Bureau's differential policy. In summary, some of the stated concerns of the Center are as follows:

- Funding equity across localities will be severely impaired. While federal dollars to each state will be equitable because the state population will reflect the actual census count, money going to each community and program will not, as their population totals will be distorted. The targeted population of each funding program could artificially become smaller or larger, undermining program effectiveness and resources.
- Many federal, state, and local statistics will produce inconsistent, unreasonable results, as they rely on the census count as a benchmark. Health, education, and criminal justice, for example, heavily rely on age-, gender-, race-

specific census data to derive statistically sound rates that may be compared over time. The noise injection will make such rates incomprehensible and comparisons across geography and time meaningless.

- Government services will be significantly impacted. Systems such as housing, transportation, and emergency management need accurate census data for planning, budgeting, and program delivery.

The District of Columbia's State Data Center (SDC), like other SDCs, is concerned about the potential impact of the differential privacy approach on the District's 2020 Census data. The Census Bureau has maintained an open dialog with its partners in regards to the evolving nature of the differential privacy policy. Three demonstration data products have been released for public use (with one more yet to be released before the policy is finalized). This report shows that the new differential privacy policy needs to be greatly improved from its current state before data users can be confident in the Census Bureau's 2020 Census data products.

Summary Findings for the District

The following sections will provide examples of how 2010 Census data in the District were altered by the differential privacy approach. The analysis focused on a few key indicators that were provided in the demonstration data – total population, race/ethnicity, and housing indicators. Age groups and other socioeconomic indicators were not provided in the most recent data set released by the Census Bureau, but presumably the findings would apply in similar ways. At the citywide and Ward geographic levels, the differential privacy algorithm made minor changes to large population groups and housing data, while making more significant changes to smaller population groups and housing indicators. At the census tract geographic level (comparable to neighborhoods), there are signs of major redistribution and distortion of smaller population groups as a result of the differential privacy policy. Table 1 provides a summary of these impacts.

Table 1 - Summary of Impacts of Differential Privacy Policy on 2010 Census Data by Geography

	Total Population	Race/Ethnicity	Total Housing Units	Household Status (Occupied/Vacant)
Citywide	No Change	Significant	No Change	Insignificant
Ward	Insignificant	Significant	No Change	Insignificant
Census Tract	Insignificant	Significant	No Change	Significant

Source: U.S. Census Bureau, 2010 Demonstration Privacy-Protected Microdata Files v. 2020-11-16

Summary Impact: Citywide

The differential privacy policy did not result in significant changes to total population or total housing units counts at the citywide geography. Table 2 shows the changes applied to 2010 Census data by the differential privacy approach for population counts by racial group. The Census Bureau increased or decreased the population of racial groups from a range of -184 to +277 individuals. For larger population groups, such as Black/African Americans, this resulted in a small change compared to their overall population (a 0.1% change for this group). However, these changes affected smaller population groups to a much larger degree, such as American Indian and Alaskan Natives which resulted in a 9.8% gain in population and Native Hawaiian and Other Pacific Islanders which resulted in a 91.7% gain in population.

Table 2 - Comparison of the 2010 Census data and Differential Privacy data for population groups at the District level

	Total Population	White	Black or African American	American Indian and Alaska Native	Asian	Native Hawaiian and Other Pacific Islander	Some Other Race	Two or More Races	Hispanic or Latino
2010 Census	601,723	231,471	305,125	2,079	21,056	302	24,374	17,316	54,749
Diff. Privacy Policy Results	601,723	231,462	304,967	2,283	20,872	579	24,239	17,321	54,902
Change	0	-9	-158	204	-184	277	-135	5	153
Percent Change	0	0	-0.1	9.8	-0.9	91.7	-0.6	0	0.3

Source: U.S. Census Bureau, 2010 Demonstration Privacy-Protected Microdata Files v. 2020-11-16

Table 1 indicates there was no change in total citywide housing unit count. Housing unit status did change by an insignificant amount where occupied housing units and vacant housing units were swapped in equal amounts. Occupied units increased by 13 units, and vacant units decreased by 13 units citywide.

Of note here is the change to smaller population groups even at the citywide geography. While increasing the population counts in order to provide protection for the privacy of these small groups may seem like a useful solution, changing the counts by approximately 10% to over 90% highlights how the approach can misrepresent those population counts.

Summary Impact: Wards

Changes to total population and to occupied and vacant housing at the Ward level were largely insignificant. Table 3 shows small changes in population counts and no change to housing unit counts. The largest impact of the differential privacy policy on housing indicators can be seen in the changes to vacant housing. Due to there being a low number of vacant homes in Wards (an average of 3,750 units in each Ward), the differential privacy policy resulted in a 6.6% increase in Ward 4 and a 5.3% decrease in Ward 8.

Table 3 Change in Total Population and Housing counts from the 2010 Census Data to the Differential Privacy data across Wards

	Total Population	Total Housing Units	Occupied Housing Units	Vacant Housing Units
Average numeric change	0	0	1.6	-1.6
Largest numeric increase	67	0	206	160
Largest numeric decrease	-47	0	-160	-206
Average percent change	0%	0%	0.0%	0.6%
Largest percent increase	0.1%	0%	0.8%	6.6%
Largest percent decrease	-0.1%	0%	-0.5%	-5.3%

Source: U.S. Census Bureau, 2010 Demonstration Privacy-Protected Microdata Files v. 2020-11-16

Table 4 shows the changes in the population counts of racial and ethnic groups across Ward boundaries. These changes had less of an impact on larger groups such as White, where population size varies from around 1,300 in Ward 7 to 65,000 in Ward 3, or Black/African American where population varies from 3,900 in Ward 3 to 67,500 in Ward 7 (Census 2010 totals). The differential privacy policy had a greater effect on smaller populations that number in the hundreds or less in each Ward, including Native Hawaiian and Other Pacific Islander and American Indian and Alaska Native.

Table 4 Change in Race/Ethnicity population counts from the 2010 Census data to the Differential Privacy data by Ward								
	Ward 1	Ward 2	Ward 3	Ward 4	Ward 5	Ward 6	Ward 7	Ward 8
White	112	-26	141	-229	-21	-41	-26	81
Black or African American	-59	113	-131	77	-52	199	-239	-66
Asian	98	-260	8	11	98	-181	50	-8
Native Hawaiian and Other Pacific Islander	-8	33	61	28	63	20	42	38
American Indian and Alaska Native	-96	102	11	-34	-83	-5	175	134
Some Other Race	-69	48	-75	157	-93	-30	-43	-30
Two or More Races	89	-36	19	-57	111	23	17	-161
Hispanic or Latino	-9	34	-55	-15	135	24	133	-94
Source: U.S. Census Bureau, 2010 Demonstration Privacy-Protected Microdata Files v. 2020-11-16								

Table 5 demonstrates how the changes to small populations result in major differences at the Ward level due to differential privacy policy. The American Indian and Alaska Native group had their population drop by 24.4% in Ward 1 and had a 91.8% increase in population in Ward 8 where they have 400 and 150 total population respectively. Native Hawaiian and Other Pacific Islanders had a dramatic 323.1% increase in population in Ward 7, this is due to there only being 13 people in this group in Ward 7 according to Census 2010 data.

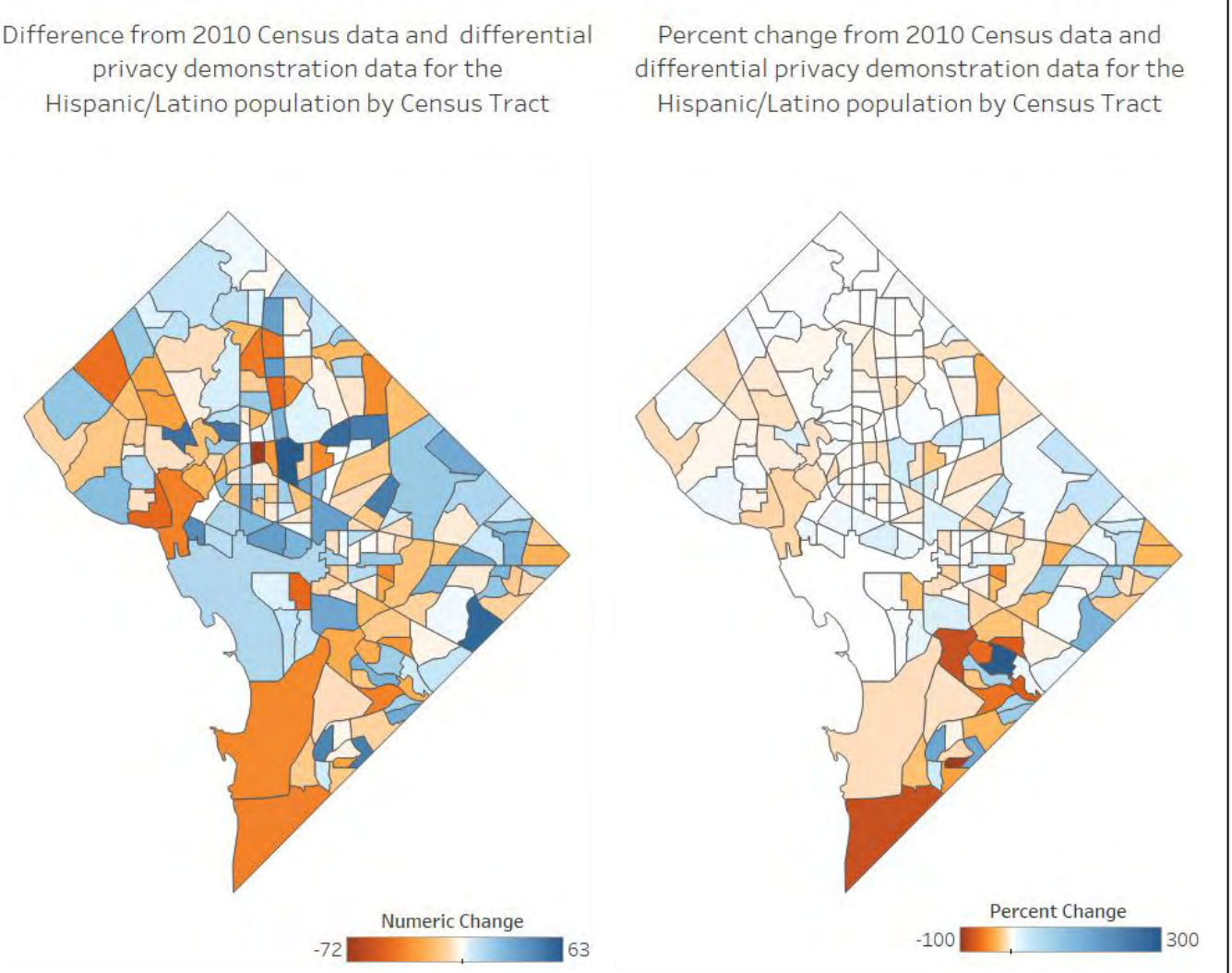
Table 5 Percent Change of Race/Ethnicity population counts from the 2010 Census data to the Differential Privacy data by Ward								
	Ward 1	Ward 2	Ward 3	Ward 4	Ward 5	Ward 6	Ward 7	Ward 8
White	0.3	0.0	0.2	-1.2	-0.2	-0.1	-2.0	3.1
Black or African American	-0.2	1.1	-3.4	0.2	-0.1	0.6	-0.4	-0.1
Asian	3.1	-3.7	0.2	0.9	9.5	-5.7	36.8	-3.1
Native Hawaiian and Other Pacific Islander	-17.0	50.8	225.9	47.5	196.9	51.3	323.1	190.0
American Indian and Alaska Native	-24.4	46.4	6.5	-10.2	-29.1	-1.6	79.5	91.8
Some Other Race	-0.9	1.7	-6.3	1.9	-4.2	-2.9	-5.8	-8.6
Two or More Races	2.9	-1.5	0.8	-1.9	5.5	1.1	1.4	-13.9
Hispanic or Latino	-0.1	0.4	-0.9	-0.1	2.9	0.6	8.0	-7.2
Source: U.S. Census Bureau, 2010 Demonstration Privacy-Protected Microdata Files v. 2020-11-16								

The next section highlights a summary of the noise in the data that are being generated for small populations.

Summary Impact: Census Tracts

The differential privacy approach would have an impact on neighborhood-sized geographies and would impact small area and place-based planning. Therefore, planners should be aware that there are limitations in the data. Figure 1 shows how the differential privacy approach changed population counts for the Hispanic/Latino population from the 2010 Census by census tract. The Hispanic/Latino population made up 9.1% of the District's population in 2010. The map on the left shows how the differential privacy algorithm added to and subtracted from the original 2010 Census counts – in other words, this is a visualization of the differential privacy 'noise' being applied to the data. For the most part, the noise seems random with the differential privacy policy applied evenly across the District. However, the map on the right shows that where smaller Hispanic/Latino population sizes exist (south and southeast sections of the map), there are more dramatic effects to the population in terms of percent change. In several census tracts, all or nearly all of the Hispanic/Latino population was transferred to another area, while other census tracts acted as receiver tracts which saw as much as a gain of 300% in Hispanic/Latino population.

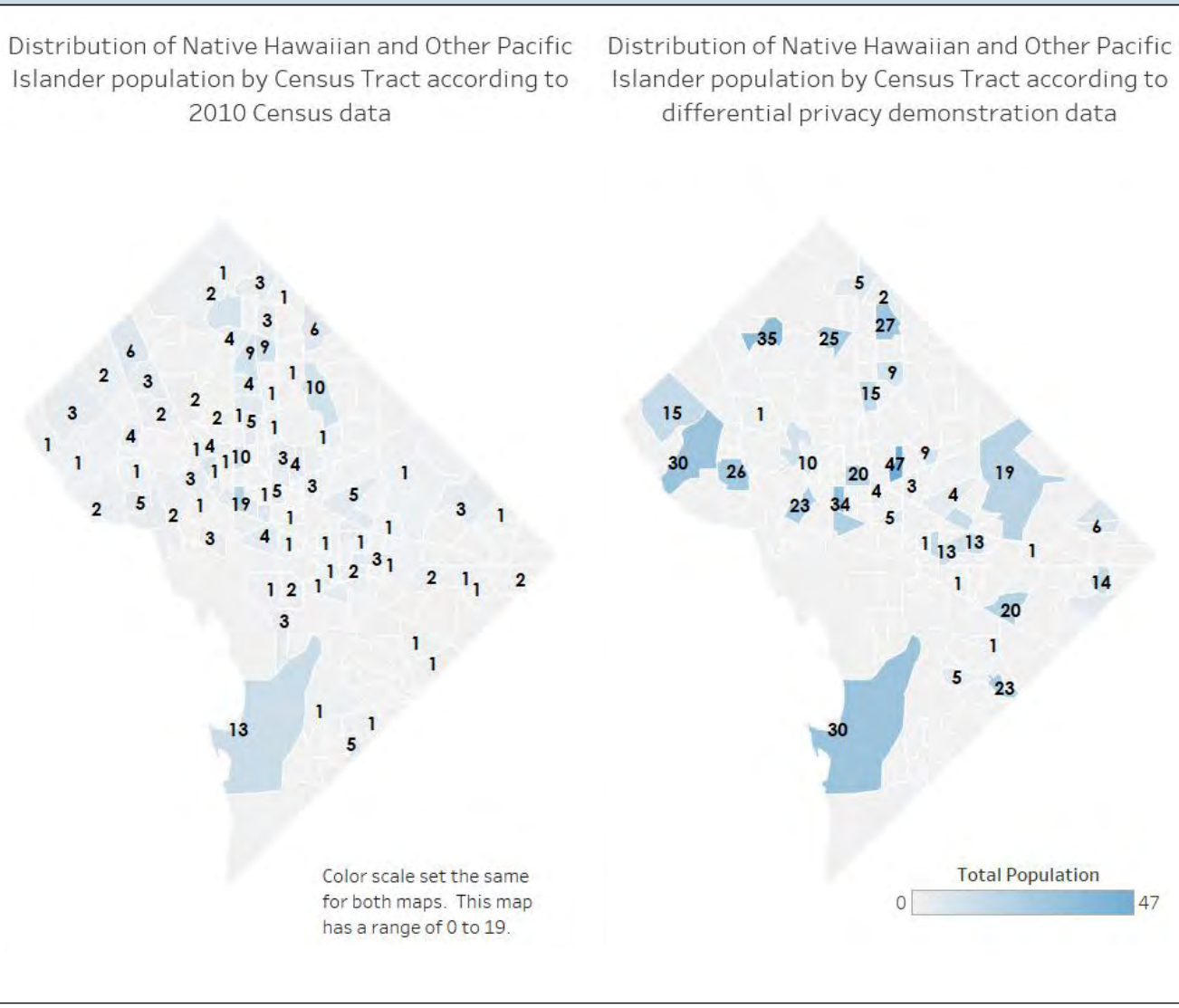
Figure 1. Numeric Change in Hispanic/Latino population (left) and percent change (right) by applying the differential privacy policy to 2010 Census data by census tract.



The differential privacy approach affects groups with small citywide populations to a greater degree, with significant changes in the distribution of the population within census tracts. Figure 2 shows the example of Native Hawaiian and Other Pacific Islanders which had 302 residents in the District according to the 2010 Census. The map on the left shows where the Native Hawaiian and Other Pacific Islanders population lived according to the 2010 Census data, and the map on the right shows how those households' data were aggregated into other areas by the differential privacy algorithm. There were 73 census tracts, out of 179 total for the District of Columbia, where all of the Native Hawaiian and Other Pacific Islanders population was removed and located to another tract. Additionally, the overall Native Hawaiian and Other Pacific Islanders population was increased by 277 for the District, and those new population data were added to the aggregated tracts as well.

Some of the population counts were aggregated into tracts where there were no Native Hawaiian and Other Pacific Islanders population in 2010. Changes of this magnitude could have major impacts on community planning efforts. This example focuses on the Native Hawaiian and Other Pacific Islanders population, but this sort of major population redistribution under the differential privacy approach would likely be applied to other similar sized populations as well.

Figure 2. Distribution of Native Hawaiian and Other Pacific Islander population by census tract according to the 2010 Census (left) and the differential privacy demonstration data (right).



Changes to the overall population for individual census tracts were small; the largest gain for a single tract was an additional 18 people and the largest loss was 17 people to a single track. Total housing unit counts were not changed at the census tract geography. While total population and housing unit count were not significantly affected by the privacy policy changes at the census tract geography, occupied and vacant housing unit counts were impacted in similar ways to the racial/ethnicity group data. The largest increase to occupied housing units was 69 units while the largest decrease to occupied housing units was 80 units. The reverse is true

for vacant housing units, where there was an increase of 80 units and the largest decrease was 69 units. Vacant housing exists in low numbers in some census tracts and therefore underwent similar shifts from one tract to another as the Hispanic/Latino population group.

CONCLUSION

This report shows how the differential privacy policy approach proposed by the Census Bureau affects Census data at varying geographic levels and population sizes. The largest issue of concern is that smaller population groups' data are being altered in ways that leads to underrepresentation or overrepresentation in neighborhoods in particular but also at the Ward and citywide level. The District of Columbia SDC understands and supports the US Census Bureau's longstanding commitment to providing accurate statistical information while maintaining confidentiality as outlined in Title 13. The District of Columbia SDC also understands the challenges presented by the proliferation of information and technology and support the Census Bureau's efforts to plan for the future. However, the results of analyzing the differential privacy demonstration data show that the current state of the policy is in need of substantial improvement.

When the District of Columbia SDC analyzed the November 16, 2020 round of demonstration data, the Census Bureau had stated that these data would be the last round of data that would be made available to the public. The situation has since changed, and the Census Bureau announced there would be another round of demonstration data product released for review. In that announcement, the Census Bureau acknowledged that part of their process included factors that would lead to more noise (and error) than should be expected. The latest demonstration data should represent the expected accuracy of the 2020 Census data products, according to the Census Bureau. The District of Columbia reviewed the new demonstration data released on April 28, 2021 and while there were improvements in accuracy at the citywide and Ward geographies, the census tract level issues outlined in this report remain. When the final 2020 Census data products are released, it will be necessary to scrutinize the results for any uncharacteristic changes. It will be up to those with local knowledge of the District's population to recognize where the data may be less accurate. In the event that discrepancies are found, the District of Columbia SDC will issue a disclaimer to all data users outlining the impact of the differential privacy policy.

References

Cynthia Divork. (2006). *Differential Privacy.* 33rd International Colloquium on Automata, Languages, and Programming, part II (ICALP).

Irit Dinur and Kobbi Nissim (2003). *Revealing information while preserving privacy*. In PODS, pages 202-210.

John Abowd (2017). *Research Data Centers, Reproducible Science, and Confidentiality Protection: The Role of the 21st Century Statistical Agency.* https://www2.census.gov/cac/sac/meetings/2017-09/role-statistical-agency.pdf

John Abowd (2018). *Protecting the confidentiality of American Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau.* U.S. Census Bureau Chief Scientist for Research and Methodology.

University of Virginia, Weldon Cooper Center for Public Services, letter to governor of Virginia (2020).

David Van Riper, Tracy Kugler, and Jonathan Schroeder. *IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data*, version 20201116 & version 20210428_12-2 [Database]. Minneapolis, MN: IPUMS. 2021. https://www.nhgis.org/privacy-protected-2010-census-demonstration-data