Appendix 1: Differential Privacy

Kem C. Gardner
POLICY INSTITUTE
THE UNIVERSITY OF UTAH

UTAH STATE DATA CENTER

November 7, 2018

Karen Battle
Chief, Population Division
U.S. Census Bureau
4600 Silver Hill Road, Room 6h174
Washington, DC 20023
POP.2020.DataProducts@census.gov

Dear Dr. Battle,

I am writing in response to the Federal Register notice, Docket Number 180608532-8537-01, requesting feedback from data users on 2020 Census data products.

I am writing on behalf of the Utah State Data Center and the Demography Team at the Kem C. Gardner Policy Institute (Institute).

The Demography Team at the Institute is mandated by the state to create long-range population projections at the state and county level and annual population estimates at the state, county, and subcounty level to inform planning and investment purposes for state and local government. Additionally, we are responsible for creating population estimates for areas aiming to become incorporated cities and we produce population estimates utilized for tax allocation purposes if Census Bureau estimates are unavailable. The State Data Center program provides technical assistance to local entities working with Census Bureau programs and products.

At the Institute, the data included in the SF1 dataset is essential to informing our models, creating scenarios, and benchmarking our long-term projections. Our four metropolitan planning organizations and state Department of Transportation then disaggregate the projections to Subcounty geographies.

In response to questions 1, 2 and 7:

Having reliable and accurate data at small geographic levels is essential to properly inform all of this work, particularly in a state like Utah where the majority of the population lives in a handful of urban counties but the remaining 25% of the population is highly dispersed throughout large, rural areas. Reliable block-level data provides a more thorough understanding of the demographics of these dispersed populations, especially due to the fact that in some of the more rural counties one Census tract might encapsulate multiple towns and their neighboring unincorporated areas – essentially nullifying the ability to identify characteristics of these small geographies. While maintaining confidentiality of data is paramount, we hope that we will be able to continue learning about our smaller, more dispersed communities through *block-level* Census summary data.

Our first round of population projections was published in 2017 and our estimates became the codified resource for the state in the 2018 Legislative Session. Due to the newness of our process, there are some products we have not yet used but intend to use in the future. Across the production of all of our population estimates and projections, we utilize the following data *at the block-level* because it is the most reliable, micro-level data available in the state to inform our models and create our framework of understanding of a place. From SF1, tables numbered: P1, P12, P16, P17, P18, P19, P20, P21, P22, P23, P24, P25, P26, P27, P28, P29, P30, P31, P32, P33, P34, P38, P39, P40, P41, P42 and P43; H1, H3, H4, H5, H13, H16, H17, H18, H19. Although we have not done so yet, we fully intend to add information on race and ethnicity into our estimates *after* the 2020 Census, which would require utilization of tables numbered: P3, P4, P5, P6, P7, P8, P9, P10, P11, P15, P12A-I, P16A-I, PH17A-I, P18A-I, P37A-I,

Additional tables at the tract and county-level geography utilized in our estimate and projections processes include: PCT20, PCT20A-I, PCT21, PCT22, PCO1, PCO2, PCO3, PCO4, PCO5, PCO6, PCO7, PCO8, PCO9, PCO10.

<u>In response to question 3:</u>

If we lack these summary tables from the decennial Census, there is not another resource in the state that could provide the same information at the same granularity. Once again, the nature of where our populations live demand a thorough, standardized, dependable resource. Relying on individual counties or state data collection would not provide that type of resource. There is no state dataset that provides the same level of rigor and detail that we can obtain from the Census Bureau data, *especially at the block-level*.

Additionally, the movement of the estimate and projection work to the Institute from the State provided the opportunity for a different formulation approach. The methodologies for estimates and projections at the Institute have been created in-house, utilizing local knowledge and smaller local datasets to supplement the Census Bureau data. However, the Census Bureau data provides the gold-standard baseline which we build all work on for the ensuing decade. If we do not receive detailed, accurate decennial Census data, our work for the next 10 years will be impacted.

<u>In response to questions 4, 5, and 6:</u>

Our population projections are essential to local legislators, state departments and agencies, school districts, water districts, non-governmental organizations, and many others, because they are the most reliable resource to plan for the future. While other resources may exist, the projections and estimates created at the Institute provide a context-sensitive approach to understanding Utah's population which are based on the Census Bureau decennial data. Since Utah has been a high growth state for several decades, this is a critical undertaking.

As stated previously, our subcounty population estimates are used by state agencies such as the Utah Tax Commission for newly incorporated areas. Local governments such as Salt Lake City also request our detailed analyses of changes between censuses. Additionally, county governments ask us to employ decennial data to build information about small areas of interest which are not incorporated places or census-designated places. All of these products are unique to the Institute and reflect our strong capacity to retrieve and use block-level data. Similar products are not available from other entities.

Due to the breadth of reach of the aforementioned players, the entire state population is affected by the use of this data whether they realize it or not.

<u>In response to question 9:</u>

If it is possible to obtain single year of age and sex by race and ethnicity or single year of age and sex by ancestry group, both at the county level, that would help inform our work immensely.

Thank you for your attention to our comments. If you have any questions, I can be reached by email at mallory.bateman@utah.edu or by U.S. mail at: Mallory Bateman, Kem C. Gardner Policy Institute, 411 E South Temple, Salt Lake City, UT 84111.

Sincerely,

Mallory Bateman
Utah State Data Center Coordinator

Technical Memo

Kem C. Gardner
POLICY INSTITUTE
THE UNIVERSITY OF UTAH
DAVID ECCLES SCHOOL OF BUSINESS

February 10, 2020

To: Natalie Gochnour, Director
From: Pamela S. Perlich, Director of Demographic Research
Mike Hollingshaus, Demographer
Subject: Differential Privacy Overview

Background

The U.S. Census Bureau (CB) will implement a new procedure to protect individual privacy in the 2020 decennial census. Because this “differential privacy” (DP) procedure is in its final development stage, impacts are still uncertain. We do know that DP will infuse the 2020 census data with noise and reduce accuracy, especially for small populations. The procedure’s use has generated great concern, confusion and controversy among data users, including researchers and policymakers. A wide range of professional organizations whose work, resources, and political representation depend upon these data have organized to communicate their deep concerns about the method’s potential harms. This memorandum defines DP, assesses potential impacts, and recommends a Utah response.

Review Process

On October 29, 2019 the CB provided a demonstration file based on data from the 2010 census. They invited review and comment as they continue to “develop and fine-tune disclosure avoidance systems.”[1] We analyzed Utah data and found that policymakers, businesses, planners, and researchers should be concerned by potential significant data inaccuracies. We conclude, along with others, that the data will be less reliable for planning and research, especially for small populations. Since Utah has many small population counties, it could be heavily disadvantaged. Differential privacy is a big deal.

What is Differential Privacy?

Title 13 requires the CB to ensure individuals cannot be identified from published census data. The CB has used various methods to protect privacy throughout its history.[2] More available external datasets and advanced computing capabilities have introduced new threats. The most recent campaign to protect privacy has been branded “disclosure avoidance.”[3]

DP is a new method for avoiding disclosure in the age of high-powered computing and big data. It lets the CB quantify the risk that a person can be identified, and implements procedures to avoid crossing a predetermined risk threshold. This is accomplished by injecting the data with random “noise.”

In DP, the noise is a series of random numbers (positive or negative) that are added to the actual counts. The spread of those numbers is determined by the error “budget.” When the budget is bigger, the random numbers are further from zero, injecting more noise. The total budget is allocated to line-item budgets for individual data tables. State total populations are fixed, because of the constitutional mandate for representative apportionment. But, all other published data—even total households—will be intentionally published with noise.

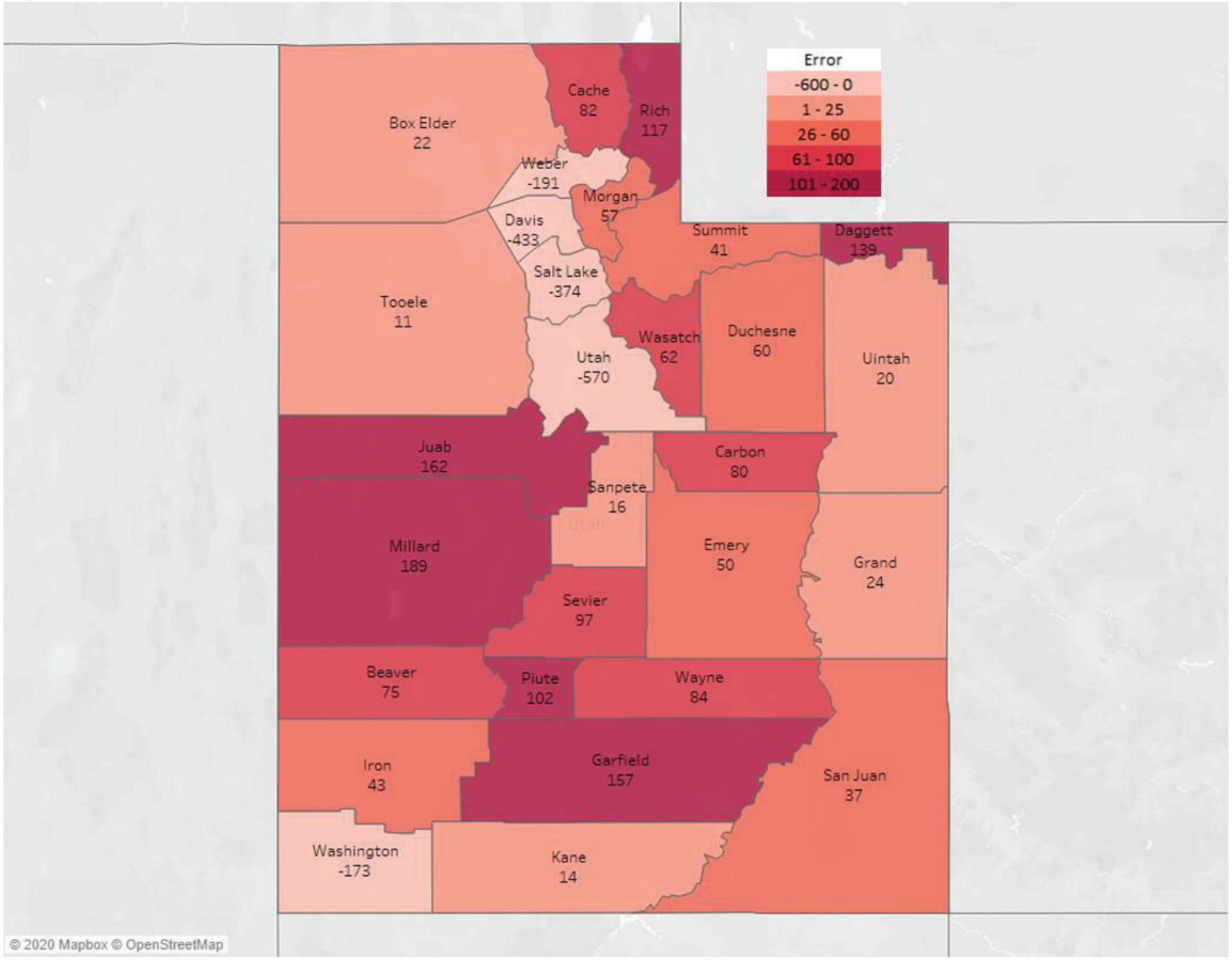
While DP enhances privacy, it degrades data quality. As noise increases, accuracy decreases. The error budgets will be published, so some statisticians can assess the quality of the data for their own large-scale research. However, most data users do not have the option of discarding the published decennial census data. Policy implementation, political representation, and funding formulas are tied to the enumeration. This will impact research and planning across the public and private sectors for the next decade.

How might Differential Privacy Affect Data Quality?
To illustrate DP's impact upon the published numbers, the CB applied the DP algorithm to 2010 census data.[4] Researchers at the National Historical Geographic Information System reformatted the data for comparison with published 2010 summary files, from which we extracted a Utah subset.[5] Geographic levels include the state, counties, cities and places, unified school districts, senate and house legislative districts, and tracts. We compared three metrics at the county level: total population, median age, and persons per household (PPH).[6]

Findings

Total Population. Figures 1 and 2 show the errors and percentage errors in total population for each of Utah's counties.

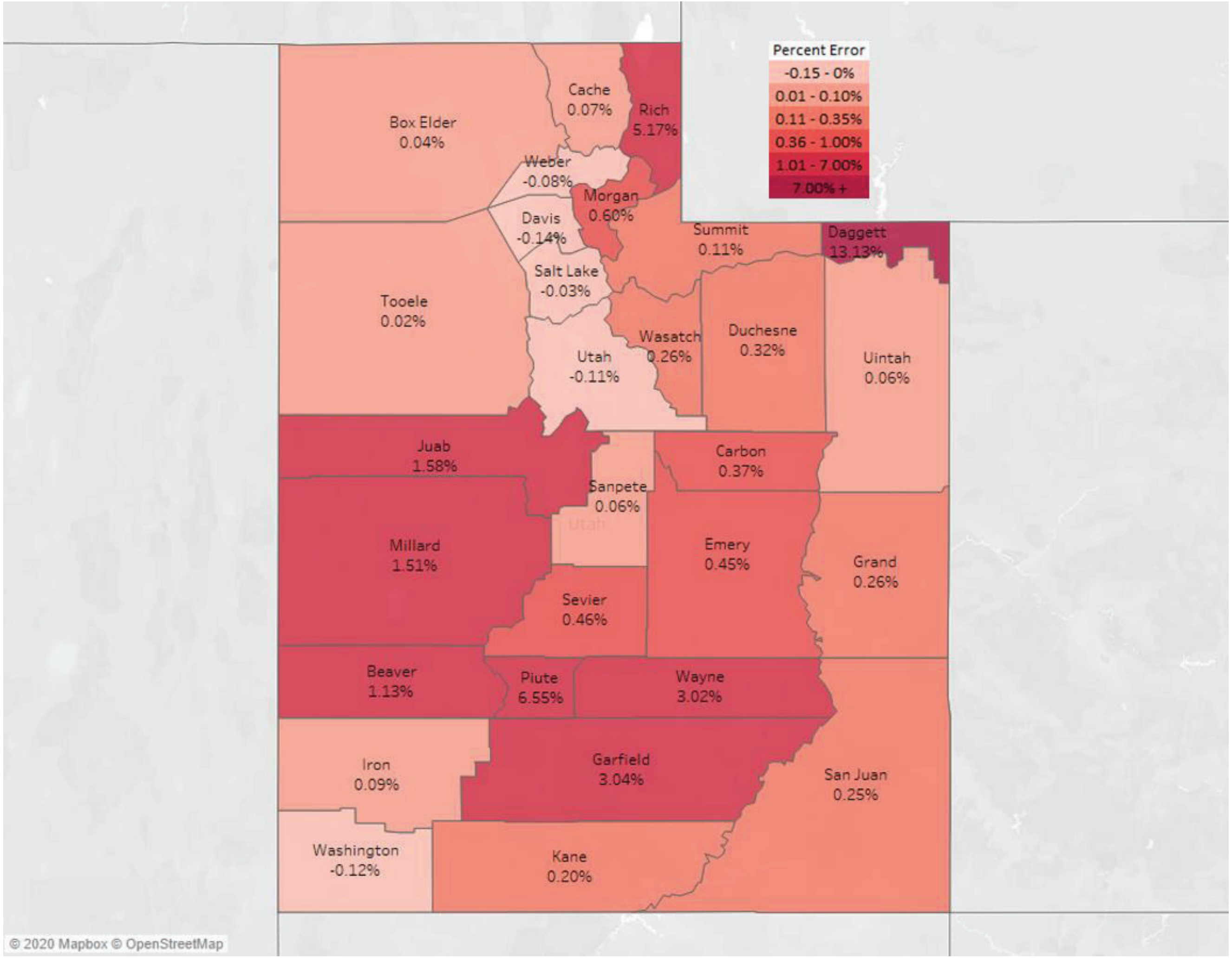
Figure 1. County Population Errors from Differential Privacy Demonstration, 2010



Source: U.S. Census Bureau and Kem C. Gardner Policy Institute.

Errors range from an undercount of 570 in Utah County to an overcount of 189 in Millard County. Percentage errors range from -0.14% in Davis County to 13.13% in Daggett County. Larger counties tend to have underestimates, and smaller counties overestimates. This pattern has been identified by other researchers, and we discuss why it occurs next.

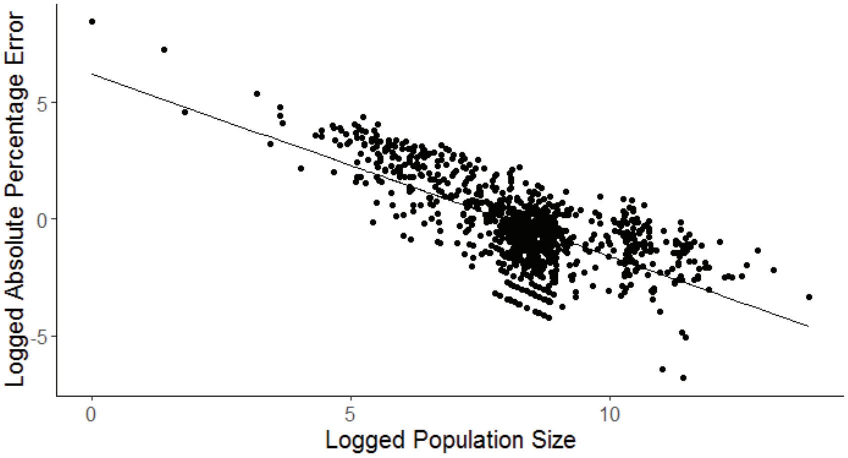
Figure 2. County Population Percentage Errors from Differential Privacy Demonstration, 2010



Source: U.S. Census Bureau and Kem C. Gardner Policy Institute.

Figure 3 shows the absolute percentage errors (APEs) in total population for all geographies (from counties to tracts) as a function of population size. A trend line shows a clear relationship, with smaller populations having larger APEs. The implications are clear. DP techniques will disproportionately harm planning for smaller populations (e.g., tracts, minority groups, less populated municipalities).

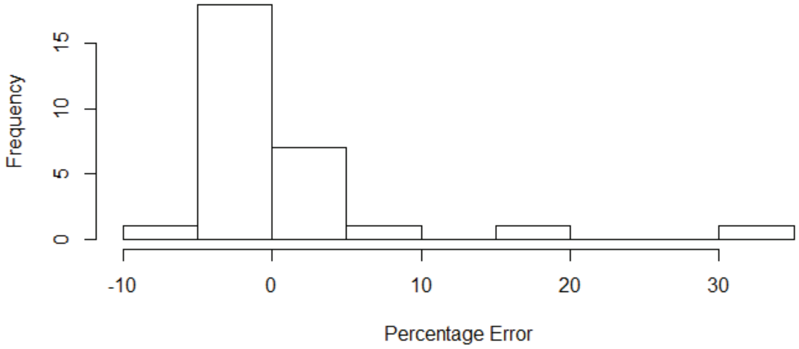
Figure 3. County Logged Population Percentage Errors Predicted by Logged Population Size, from Differential Privacy Demonstration, 2010



Source: U.S. Census Bureau and Kem C. Gardner Policy Institute.

Median Age. Accurate age data are crucial determinants of multiple planning priorities, such as public health and education. Percentage errors in median age range from -7.24% in Daggett County to 31.00% in Wayne County (where the DP estimate is 48.6 compared to the published 37.1). Age data are foundational to demographic research. Figure 4 shows the median age percentage error distribution.

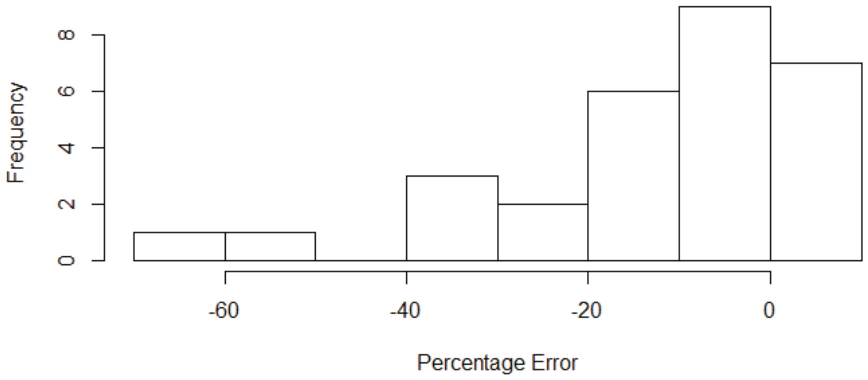
Figure 4. Histogram of County Median Age Percentage Errors from Differential Privacy Demonstration, 2010



Source: U.S. Census Bureau and Kem C. Gardner Policy Institute.

Persons per Household. One particular area of concern is the average household size, measured by PPH. This value is critical in the large and small-scale economic and population models utilized by entities such as the Kem C. Gardner Policy Institute, Wasatch Front Regional Council (WFRC), and Mountainland Associations of Government (MAG). It is used to convert between population and household counts. Figure 5 is a histogram of the percentage errors in this metric for Utah's counties. They range from -66.21% in Rich County to 0.07% in Davis County.

Figure 5. Histogram of County Persons per Household Percentage Errors from Differential Privacy Demonstration, 2010



Source: U.S. Census Bureau and Kem C. Gardner Policy Institute.

In Rich County, the DP estimate for PPH is 0.95, compared to the published 2.81. This estimate is impossible and completely unacceptable for planning and analysis. By definition, average household size must be no smaller than 1.0. Note that these are at the county level. Many planning models use PPH at even smaller geographies such as tracts, and these are subject to even larger error.

Potential Explanations

After initial data collection and processing, DP introduces new measurement errors in two ways. The first is the direct error, purposefully introduced into the model by the DP algorithm to protect privacy. The second is indirect error, which are computational errors introduced during the post-processing of published statistics. Data quality deteriorates as errors propagate with each additional computation.[7]

A common finding in population research is that direct error produces larger percentage errors for smaller populations. For example, each age category for Wayne County includes few people, so percentage errors are large. Direct error also tends to make smaller populations larger, and larger populations smaller. This pattern results from the requirement that counts cannot be negative, so decrements tend to come from larger populations and increments from smaller.

A PPH less than one for Rich County is explained by indirect error, and results from the noise being generated independently for households and population. The CB does not directly add noise into PPH; rather, it is obtained by simply dividing the published household population count by the household count. For Rich County, the household count just happened to be larger, by random chance. Even when estimates are feasible, they will have larger errors due to the additional computation.

These DP procedures have been developed and evaluated by statisticians who assert that, by statistical standards, data accuracy will not be compromised. Analysis by the data user community highlights significant data errors that have consequential planning, policy, funding, and representation implications. Data users and statisticians view accuracy from different perspectives.

Cautions

The demonstration file provided by the CB is only a prototype. They continue to develop and test procedures. So our findings should be cautiously interpreted. The CB will revise the algorithm based upon public feedback, especially from their partners at the Federal-State Cooperative for Population Estimates (FSCPE), the State Data Centers (SDC), and the

Census Information Centers (CIC). They have invited other entities with technical census data expertise to submit feedback on how their commonly used statistics are impacted.[8] So there is still time to influence the outcome.

Additionally, results of DP are based upon a random process. If a second demonstration file were generated using the same DP algorithm, results would differ. One file might underestimate a county's population, with another overestimating. This unpredictability is why the DP algorithm is effective for avoiding privacy disclosure.

How might Differential Privacy Affect Data Users?

It is alarming that DP could potentially degrade the census 2020 data accuracy to the point that they are unusable for quality research and planning. Both academic and applied researchers have expressed informed criticism. Multiple CB network partners identified serious incongruences that will hamper informed planning.[9] The academic research community has criticized the method and questioned its necessity.[10] We have not yet identified concerns among business groups who rely upon the data for planning. But, they may exist, or likely will soon.

These new data inaccuracies could harm the work we do at the Gardner Institute on behalf of the state. In particular, the PPH measure is critical for estimates and projections methods that convert between households and people. These are utilized in our housing unit method at the tract level.[11] Our household projections rely upon a closely related measure called headship rates to convert population into households.[12] Real estate, land use, and commuting models used by organizations such as WFRC and MAG often rely upon accurate estimates of the PPH metric.

These new data limitations will affect many other arenas, including (but not limited to):

- Federal funding may be ill-distributed.
- The geographic distribution of sales tax receipts may be more difficult to predict.
- Businesses, local and state governments, and nonprofits will have less reliable data for determining demand, location choice, and marketing strategies.
- Planning for higher and primary education will be affected, particularly since it relies so heavily upon accurate data for population age characteristics.
- Planning for water, energy, and other utility demand will be less accurate, especially at smaller areas.
- The drawing of State and House Legislative boundaries will be affected.
- Patterns of family living such as marriage, divorce, and cohabitation will be harder to track.
- Race/ethnic estimates will be inaccurate, especially in Utah where some groups have small population sizes, impacting policies such as Equal Employment Opportunity.
- Economic indicators such as the labor force participation rate may be incorrectly estimated, indirectly affecting other measures such as the unemployment rate.
- The Utah Department of Health will have additional errors injected into their estimates of death and disease rates, making public health research less reliable.
- Many research instruments, such as polls and surveys, are designed and weighted according to census data; their findings will be less accurate.

What can be done?

Despite considerable pushback from key data users, the CB appears poised to implement DP for the first time with the decennial 2020 census summary products. Data users should be aware that the data will be less reliable for small populations and geographies.

However, the CB is also seeking public feedback on the method, including potential accuracy concerns identified from the demonstration products. Feedback should be submitted by summer of 2020 at dcmd.2010.demonstration.data.products@census.gov.[13] Some states have provided feedback, and *we are in the process of preparing our report*. The CB particularly seeks feedback on the summary file tables and statistics deemed most critical for planning. This will help them allocate the error budgets to minimize bias where needed, and potentially revise the algorithm. The CB is already aware of some issues, including the impossible PPH estimates, and is working on improving their algorithms to avoid serious errors.[14] But, so many post-processing statistics can be calculated that it seems unlikely they will avoid all pitfalls.

While other data sources are available, many are private and expensive. Furthermore, vendors often rely upon accurate census counts to ensure their own data integrity. Ripple effects will be keenly felt throughout the public and private spheres. The Federal Statistical Research Data Centers (RDCs) will still be available for approved work projects, and the University of Utah now houses one of these institutions. But, security restrictions make RDCs difficult to access.[15]

Recommendations

These findings portend large data inaccuracies if the CB implements DP as currently specified. In our opinion, the consequences are sufficiently serious to warrant further action. We recommend the following steps:

1) Complete our formal analysis and submit it to the CB.
2) Adapt this document as needed for presentation to other stakeholders, especially the Governor's Office and Legislature, two entities that may want to consider sending formal comments.
3) Strategically connect with other data users and policy organizations, and encourage them to connect with their national organizations and amplify their concerns.
    - For example, on January 31 we participated in a conference call (by invitation) with Jerry Howe from the Utah Office of Legislative Research and General Counsel. He has performed some preliminary analysis showing redistricting implications.
    - Various business and nonprofit associations might help inform stakeholders and work with the CB to minimize losses. The National Conference of State Legislatures recently published a blog on this topic, which could serve as a useful prototype.[16]

Summary

DP is a privacy protection process the CB intends to implement for all publicly published 2020 decennial census data. These particular privacy restrictions will degrade data accuracy, impacting the quality of population research and planning in both private and public spheres. Several applied and academic research entities are challenging the CB, and providing suggestions for how to minimize the loss in data quality through public feedback options. Most likely, census data users will have to deal with less reliable data. We should take further action to minimize detrimental impacts to Utah business, policy, and research. Finally, it is quite possible the CB will lose some of their hard-earned respect as a quality data distributor.

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

[2] U.S. Census Bureau. (2019). A history of census privacy protections. https://www2.census.gov/library/visualizations/2019/communications/history-privacy-protection.pdf

[3] U.S. Census Bureau. Disclosure avoidance and the 2020 census. https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html.

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

[5] National Historical Geographic Information System (NHGIS). Differentially private 2010 census data. Minneapolis: Integrated Public Use Microdata Series (IPUMS). Retrieved January 15, 2020, from https://www.nhgis.org/differentially-private-2010-census-data

[6] We have assembled a detailed dataset with many additional variables.

[7] Alonso, W. (1968). Predicting best with imperfect data. *Journal of the American Institute of Planners, 34*(4), 248–255. https://doi.org/10.1080/01944366808977813

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

[9] The Steering Committees of Census Information Centers, Federal State Cooperative for Population Estimates, and State Data Centers. (November 27, 2019). *Joint letter to U.S. Census Bureau Director Steven Dillingham regarding differential privacy.*

[10] See, for example, Mervis, J. (2019). Researchers object to census privacy measure. *Science*, 363(6423), 114–114. https://doi.org/10.1126/science.363.6423.114. Also, Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. (2019). Differential privacy and census data: implications for social and economic research. *AEA Papers and Proceedings, 109*, 403–408. https://doi.org/10.1257/pandp.20191107

[11] The Census Bureau uses a similar method for their own small-area estimates. See U.S. Census Bureau (2019). Methodology for the subcounty total resident population estimates (Vintage 2018): April 1, 2010 to July 1, 2018. https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2018/2018-su-meth.pdf

[12] These are currently done at the county-level for multiple age groups, and we therefore anticipate large errors. The Census Bureau does not project households, but the Harvard Joint Center for Housing Studies uses a similar method as us to convert Census Bureau population projections into households for the U.S. See McCue, D. (2018). *JCHS Updated Household Growth Projections: 2018-2028 and 2028-2038 (No. W14-1).* Joint Center for Housing Studies of Harvard University. http://www.jchs.harvard.edu/sites/default/files/w14-1_mccue_0.pdf

[13] There does not appear to be an exact date planned for the public comment period to close. The guideline "summer of 2020" was provided in a recent presentation: Hawes, M., Senior Advisor for data Access and Privacy Research and Methodology Directorate at the U.S. Census Bureau. (January 21, 2020). Title 13, differential privacy, and the 2020 decennial census. AUBER webinar. https://auber.org/member/auber-webinars/

[14] Hawes, M., Senior Advisor for Data Access and Privacy Research and Methodology Directorate at the U.S. Census Bureau. (January 21, 2020). Title 13, differential privacy, and the 2020 decennial census. AUBER webinar. https://auber.org/member/auber-webinars/

[15] Researchers must obtain Census Bureau Special Sworn Status, and all research must be approved and conducted in a secure location. See U.S. Census Bureau. (2019). Federal Statistical Research Data Centers: Secure research environment. https://www.census.gov/about/adrm/fsrdc/about/secure_rdc.html

[16] National Conference of State Legislatures (2020). Differential privacy for census data explained. https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx

Kem C. Gardner
POLICY INSTITUTE
THE UNIVERSITY OF UTAH

DAVID ECCLES SCHOOL OF BUSINESS

April 24, 2020

U.S. Census Bureau
2020 Census Disclosure Avoidance System

Cc: Natalie Gochnour, Director, Kem C. Gardner Policy Institute

Dear Census Bureau Planners:

The Kem C. Gardner Policy Institute serves as the demographic team for the State of Utah. We are home to the Utah State Data Center and represent Utah in the Federal State Cooperative for Population Estimates and the Federal State Cooperative for Population Projections. We also produce independent demographic estimates and projections for Utah: https://gardner.utah.edu/demographics/

We bench all of our demographic estimates and projections on decennial census data. It is crucial to our demographic analyses and planning efforts in Utah.

We have serious concerns that your implementation of differential privacy algorithms jeopardizes the accuracy of Census 2020 data, especially for small population groups and entities.

We are hopeful that you might reconsider using the proposed differential privacy algorithms and implement a different solution. If this is not possible, we request that you do the least damage as possible to the accuracy of this essential data.

Attached are:

1) Detailed listings of data from the decennial census that we use in our demographic work, and
2) An analysis of data from the differential privacy algorithm test with explanations of our significant concerns.

We are hopeful that the Census Bureau will maintain the quality and accuracy standards that have historically been the standard for decennial census data.

Sincerely,

Pamela S. Perlich

Pamela S. Perlich, PhD
Director, Demographic Research