Effects of Differential Privacy Among Communities of Color Across Southern Arizona

Making Action Possible in Southern Arizona (MAP Dashboard)
White Paper #17
April 06, 2021

Prepared by

Jason R. Jurjevich, Ph.D., Associate Professor of Practice, School of Geography, Development & Environment, University of Arizona

Nicholas Chun, Ph.D. Student, School of Geography, Development & Environment, University of Arizona

Author contact information:

Jason R. Jurjevich: jjason@email.arizona.edu

Table of Contents

Acknowledgements

The authors thank David Van Riper, Alex Brasch, Beth Jarosz, and Paul Lask for their feedback and helpful insight on earlier versions of this report. Thanks also to Bishan Zhao for help with the web-based dashboard. Finally, we extend our appreciation to the Making Action Possible (MAP) coalition for their financial support of this research.

Executive Summary

Individual confidentiality is protected in all U.S. Census Bureau (USCB) data products under current federal law. Recently however, innovations in computational science, combined with widely available sources of public data, are making it easier for outside parties to potentially identify individuals (Garfinkel et al. 2018; Jarmin 2018). It is now more difficult for the Bureau to provide quality statistical information while simultaneously safeguarding individual confidentiality (Abowd 2018). As a result, the Bureau has explored different techniques to uphold data privacy standards.

Beginning with 2020 Census data products, the USCB is implementing differential privacy to protect respondent confidentiality. In most simple terms, differential privacy distorts the data by injecting "noise" into publicly available data, making it more difficult to identify individuals. Differential privacy attempts to strike a balance between privacy and accuracy. However, certain populations require more data distortion to guarantee the same level of privacy compared to larger populations. Communities of color, for example, are one group that will unfortunately shoulder the unequal costs of differential privacy. More noise will render less accurate data, likely leading to "wildly inaccurate numbers" for sub-county geographies (Wezerek and Van Riper 2020).

Knowing the extent to which differential privacy renders data unreliable and the implications for public policy remains largely underexplored. Our concern is that unequal adjustments across population subgroups and space will disproportionately affect the true representation of communities of color, 1) distorting and potentially silencing the stories of already marginalized communities, and 2) making it more difficult to achieve data-driven, equity-focused policy.

The USCB released demonstration data files so data users could assess the potential effects of differential privacy. Our analysis is based on the first vintage demonstration data, released in October 2019. The USCB made adjustments to this vintage, releasing a revised demonstration file in May 2020. However, based on our analysis, the 2020 vintage produced greater data distortion for populations of color at the census tract-level compared to the 2019 vintage.

In this white paper, we address two questions: 1) how reliable are differentially-private data for Arizonans of color? 2) to what extent does differential privacy introduce unequal data distortion among Arizonans of color at sub-county geographies?

Key Findings

How reliable are differentially-private data for Arizonans of color?

- Differential privacy error is severe for non-Hispanic populations of color and negligible for non-Hispanic White and Hispanic/Latino populations in Southern Arizona (Table ES1).
- Differential privacy error renders data for non-Hispanic populations of color unreliable across many census tracts.
- Differential privacy introduces data distortion that could make it difficult, and in some cases improbable, to realize data-driven and equity-focused governance.
- Differential privacy error introduces issues of data injustice because non-Hispanic populations of color are more likely to shoulder the disproportionate costs of differential privacy—more noise rendering less accurate data.

	Median Actual Population	Median Absolute Error	Relative Error
White	1,928	12	0.6%
Hispanic/Latino	1,002	27	2.7%
American Indian and Alaskan Native	30	8	26.7%
Asian	62	10	16.1%
Black/African American	76	11	14.5%
Native Hawaiian and Pacific Islander	3	3	100.0%
Other	4	3	75.0%
Two or More Races	58	14	24.1%

Table ES1. Median Population and Differential Privacy Error by Race/Ethnicity for Southern Arizona Census Tracts

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

To what extent does differential privacy introduce unequal data distortion among Arizonans of color across census tracts?

- Population size and race/ethnicity, not rural-urban status, are statistically significant factors of differential privacy error in Southern Arizona census tracts.
- More than half of all census tracts—55%—in Santa Cruz, Pima, Cochise, and Yuma Counties have at least one racial/ethnic group with severe data distortion.
- Our web-based map application and dashboard, available here, features an Error Index that identifies areas disproportionately affected by differential privacy. This allows users to explore the geography of differential privacy error by race/ethnicity across Southern Arizona census tracts (Figure ES1).

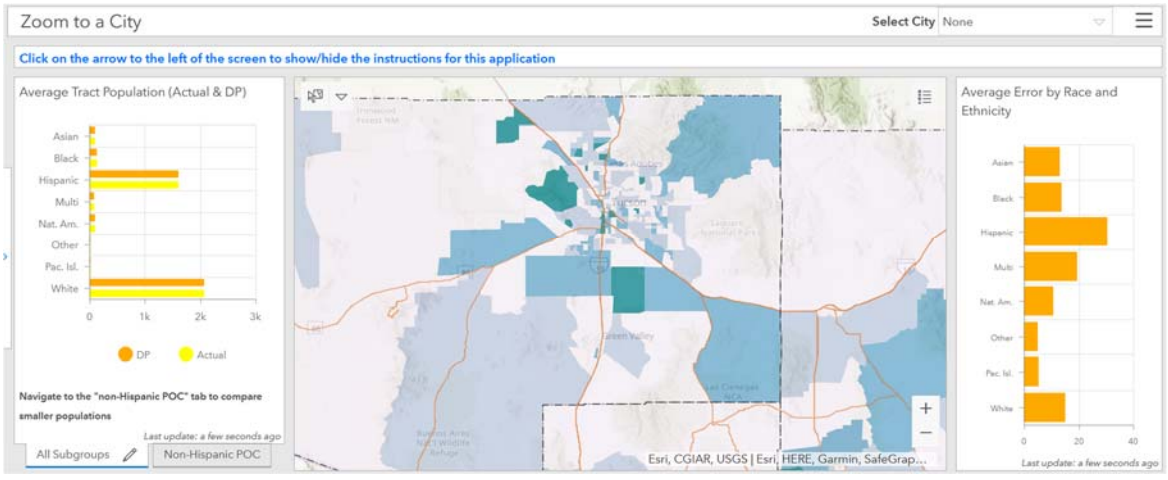

Figure ES1. Southern Arizona Differential Privacy Dashboard

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Map created by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

On-the-Ground Example

Differential privacy error will significantly affect the accuracy of 2020 Census data, especially for population subgroups, sub-county geographies, and less-populated areas. For example, consider the differential privacy error for the 2010 Black/African American population in Census Tract 25.01, located near the Santa Cruz River Park in South Tucson. The reported Census 2010 population is 243 individuals, compared to a differentially-private value of 381 (Table ES2). This results in an absolute error of 138 individuals and a relative error of 57% (i.e., the relative error is more than half of the actual value). Put another way, the differential privacy error is so large that it masks the actual population *decline* of 39 Black/African American Tucsonans during 2000 to 2010; instead, differential privacy yields a population *increase* of 99 people (Table ES2).

Differential privacy error is significant. It will almost certainly limit our ability to understand local neighborhood dynamics, make it more difficult to accurately measure disparate health and economic outcomes, and materially affect the Making Action Possible (MAP) data indicators. Our findings illustrate that less-accurate census data will almost certainly make it more difficult to achieve data-driven, equity-focused policy in Southern Arizona.

	Population			Population Change 2000 - 2010	
Black/African American	2000 Actual	2010 Actual	2010 DP	Actual	DP
Census Tract 25.01	282	243	381	-39	99

Table ES2. 2000 and 2010 (Actual and Differentially-Private) Black/African American Population, Census Tract 25.01

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through IPUMS (2019).

In February 2021, the USCB announced plans for the final privacy-loss budget for 2020 Census data products. Unlike the privacy-loss budget used in the demonstration data, the final privacy-loss budget for 2020 Census data will yield less noise and more accurate data. This decision, we believe, is a step in the right direction and is good news for data users. This also means that the data distortion in 2020 census data might be less severe than what we report here. In the end, our high-level findings remain relevant for public policy, research, and governance for two principal reasons: 1) the well-intentioned adjustments of differential privacy disproportionately distort and potentially silence the stories of marginalized communities and, 2) the data distortion could make it difficult, and in some cases improbable, to realize data-driven and equity-focused governance, especially for socially disadvantaged groups and communities of color.

Introduction: Effects of Differential Privacy Across Southern Arizona

The United States Census Bureau (USCB) is facing a critical challenge that until recently has garnered little public attention. Innovations in computational science, combined with widely available sources of public data, together are making it easier for outside parties to potentially identify individual respondents in USCB data (Ruggles 2018; Garfinkel et al. 2018; Jarmin 2018). The potential disclosure of individual-level information threatens confidentiality protections guaranteed under federal law. Today, it is more difficult for the Bureau to provide quality statistical information while simultaneously safeguarding individual confidentiality (Abowd 2018).

The USCB has protected individual privacy in the past using traditional data disclosure techniques, including data aggregation, variable suppression, data swapping, and various other approaches (Ruggles 2018). While these approaches were relatively effective at protecting respondent confidentiality, they are inadequate safeguards in today's data environment. In recent years, the Bureau has explored various approaches for modernizing disclosure avoidance to ensure that census data remain private.

Beginning with the 2020 Census, the USCB is implementing differential privacy[1] as its primary tool to ensure census data remain confidential. By injecting "noise", or error, into data, differential privacy makes it virtually impossible to identify individuals from aggregate census data with certainty (Garfinkel et al. 2018). However, data privacy—the chief advantage of differential privacy—comes at a cost: data distortion. Proponents of data privacy generally argue for more noise (i.e., less accurate data) while data user advocates support less noise (i.e., more accurate data). Striking a balance between data privacy and accuracy is the primary challenge of differential privacy.

Balancing data privacy and accuracy is more difficult for population subgroups, sub-county geographies, and less-populated areas. More data noise is required to guarantee the same level of privacy compared to larger populations (NCSL 2020; Nagle and Kuhn 2019). Differentially-private census data will thus likely "produce wildly inaccurate numbers" for sub-county geographies, for example (Wezerek and Van Riper 2020). In this white paper, we analyze and report the extent to which differential privacy error disproportionately affects individuals of color at sub-county geographies. Our concern is that less-accurate census data will almost certainly make it more difficult to achieve data-driven, equity-focused policy in Southern Arizona.

In 2019, the USCB released demonstration data to illustrate the potential impact of differential privacy in Census 2020 data. These data allow us to assess the reliability and accuracy of differentially-private data among communities of color across Southern Arizona. To better understand the impact of differential privacy across Southern Arizona, we examine two questions: 1) how reliable are differentially-private data for Arizonans of color? 2) to what extent does differential privacy introduce unequal data distortion among Arizonans of color at sub-county geographies? The first research question allows us to establish a descriptive baseline of noise-induced error by race/ethnicity; the

[1] The USCB relies on a "top-down algorithm" to modernize disclosure avoidance in 2020 Census data products. The top-down algorithm uses differential privacy as a tool for data privacy and disclosure control. In this report, we refer to disclosure avoidance using the term "differential privacy." For more information, see Van Riper et al. (2020).

second allows us to determine if any inferential relationship exists between data distortion and communities of color across sub-county geographies.

The recent adoption of differential privacy means that the questions we address here are largely underexplored. Our work advances the understanding of differential privacy in two key ways. First, we ground our analysis by highlighting the implications for public policy at the local level. A recent study showed that Arizona cities would experience some of the greatest levels of data distortion in the nation under differential privacy (Nagle and Kuhn 2019). Differentially-private Census 2020 data will almost certainly affect the Making Action Possible (MAP) data indicators, making it more difficult to measure progress and guide community action around equitable housing policy; sustainable transportation options; economic justice; erasing racial/ethnic discrimination, and many other equity-driven goals.

Second, by protecting individual confidentiality, differential privacy yields equitable privacy benefits. The question, however, is at what cost? Certain populations, notably rural residents and communities of color, are more likely to feel the disproportionate effects of differential privacy—more noise that renders less accurate data (Brasch 2020). This underscores issues of data justice. In other words, the burden of protecting respondent confidentiality is not shared equitably across populations. Moreover, the well-intentioned adjustments of differential privacy could distort and potentially silence the stories of already marginalized communities. This will almost certainly challenge our ability to realize data-driven and equity-focused governance, especially for socially disadvantaged groups and communities of color.

U.S. Census Bureau Data

Data from the USCB capture who we are, where we live, and how we're growing as a nation. For decades, the decennial U.S. census has served as the primary source of individual and household sociodemographic data. Conducted every 10 years since 1790, the decennial census provides data that support: 1) reapportionment and redistricting, 2) allocating federal funds to states, including $20.5 billion dollars to the State of Arizona in FY 2016 (e.g., Medicare/Medicaid, Supplemental Nutrition Assistance Program (SNAP), Section 8 Housing Vouchers, and others) (Reamer 2020), and 3) more than 130 surveys and programs (U.S. Census Bureau 2020a).

The USCB administers the nation's largest continuous household survey, the American Community Survey (ACS). The ACS and other USCB programs and surveys provide social, demographic, and economic data that are foundational to local, regional, and state public policy. Most survey and program data, including the ACS, are statistical sample estimates that are derived, in part, using decennial census data (among other sources, including administrative data, for example). The advantage of decennial census data is that the data represent a complete enumeration of individuals and households at a given point in time, allowing statisticians to control ACS estimates using recent censuses as anchor points. Accurate decennial census population and housing data yield poverty estimates, mortality rates, and countless other indicators that are foundational to research and governance.

The technical utility of decennial census data, combined with its foundational role in public policy, together underscore the importance of decennial census data. We highlight the on-the-ground implications of our findings by illustrating how differentially-private data negatively affect the utility of decennial census as denominator data for demographic and economic estimates, civic indicators, public

health metrics, and public policy more broadly (Jarosz 2019). A recent study by Santos-Lozada et al. (2020), for example, shows that data distortion introduced by differential privacy biases mortality rates for non-Hispanic Black/African American and Hispanic/Latino individuals. Grounding the implications for social, economic, civic, and health metrics—critical guideposts for data-driven policy—is essential for underscoring the day-to-day policy impacts of differential privacy (Hotz and Salvo 2020; Weldon Cooper Center for Public Research 2020).

Data Stewardship and Individual Confidentiality

The USCB collects sensitive data from individuals and households, publishing demographic data that is essential for telling the story of Americans and their communities. Over the decades, the Bureau has safeguarded data by implementing various privacy protections (U.S. Census Bureau 2019a). One of the most significant breaches of data privacy occurred during the early-to-mid 1940s. Federal officials from the War Department requested individual-level census data of Japanese Americans—including names and addresses—to identify and support the internment of Japanese Americans during World War II (Gates 2000). In 1954, the U.S. Congress addressed this data violation by passing Section 9 of the Census Act, U.S. Code Title 13. The law stipulates that the Bureau must uphold principles of data stewardship by protecting individual confidentiality in all publicly available data (U.S. Census Bureau 2020b).

Since the passage of Title 13, the Bureau has relied on traditional data disclosure techniques, including data aggregation (e.g., reporting data at aggregate census geographies), variable suppression (e.g., making data unavailable that fail to meet a minimum case threshold), data swapping (e.g., exchanging potentially identifiable information of one respondent with another respondent in a nearby geography) and other approaches to ensure that individual census data remain private (Ruggles 2018; Moore 1996). In today's digitally-mediated world, however, traditional data disclosure techniques are largely ineffective for protecting individual confidentiality under Title 13. Innovations in computational science, combined with widely available sources of public data, together make it easier for outside parties to potentially identify individual respondents in USCB data products (Ruggles 2018; Garfinkel et al. 2018). For example, USCB researchers were able to re-identify the race, age, sex, and location of 52 million individuals, down to the census block, when combining publicly available Census 2010 data with commercially available data sources (Van Riper et al. 2020; Hawes 2020; Jarmin 2018).

What is Differential Privacy?

The USCB Disclosure Review Board released a statement in August 2016 citing the ineffectiveness of traditional disclosure avoidance techniques in today's digital environment. They noted that "...Census Bureau and other researchers will need to develop, test, and apply new methodologies and techniques to Census Bureau data'' (Wisniewski 2016). Following extensive research, the USCB decided to implement differential privacy as the primary tool to ensure census data remain confidential, beginning with 2020 Census data products.

By introducing "noise", or error, differential privacy mathematically guarantees a level of protection from individual data reconstruction from aggregate census data (Garfinkel et al. 2018). Noise infusion is not a new disclosure avoidance technique. The Bureau has injected noise into decennial census data

since 1980 to meet data privacy requirements (Abowd 2018). The difference under differentially private algorithms is an increased transparency of how noise impacts data privacy and accuracy.[2]

Injecting error into published data makes it virtually impossible to reconstruct a database of individual characteristics. The database reconstruction process operates much like the popular puzzle game, Sudoku. In Sudoku, players follow a set of mutually exclusive rules to solve unknown number puzzles based on a unique combination of available numbers. Published census data sans differential privacy operate much like predefined numbers in Sudoku; more publicly available cross-tabulated data combinations make it easier to potentially re-identify individuals. The risk of reconstruction is reduced when summarized data are less representative of the true population values. This can be achieved by either changing the median age of a census tract from 39.2 to 40.1, for example, or by adding people to one geographic area and subtracting them from another (Wezerek and Van Riper 2020; Santos-Lozada et al. 2020).

Differential privacy creates a number of challenges arising from the tension between accuracy and privacy. Establishing the "acceptable" limit for the amount of disclosure avoidance is the most important and controversial decision. The "correct" decision is situated along a continuum with complete privacy (i.e., no data) anchored at one end of the continuum and total data availability (i.e., no privacy) at the opposite end (Van Riper and Spielman 2019; Abowd 2018). Proponents of data privacy generally argue for more noise (i.e., less accurate data) while data user community advocates support less noise (i.e., more accurate data).

Providing quality statistical information while simultaneously safeguarding individual confidentiality is a difficult and onerous responsibility. Bureau officials acknowledge that differential privacy "lacks a well-developed theory for measuring the relative impact of added noise on the utility of different data products, tuning equity trade-offs, and presenting the impact of such decisions'' (Garfinkel et al. 2018). Toward this end, in 2019 the USCB asked the Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and Medicine to convene a group of census data users, including planners, policy makers, scholars, and community advocates. The goal of the meeting was to solicit feedback on 1) how decennial data products are used in day-to-day tasks and, 2) how differentially private data might impede data analysis, if at all.

Effects of Differential Privacy Among Communities of Color

Presentations at the 2019 CNSTAT meeting generated helpful insight around differential privacy. Several experts confirmed that differentially-private data are less accurate for sub-county geographies, population subgroups, and less-populated areas. For these groups, the parameters of data privacy require more distortion to guarantee the same level of privacy compared to larger populations (Beveridge 2019; Nagle and Kuhn 2019). This is true in Southern Arizona. Under the 2019 demonstration data scenario, the 2010 housing occupancy rate for Santa Rosa, AZ, a census-designated place in the Tohono O'Odham Nation, for example, increases from 73% to 100% with differential privacy (ESRI 2019). Our concern is that unequal adjustments across groups and space disproportionately affect the true representation of communities of color, in particular. Consider also the implications for racial

[2] Under traditional data disclosure methods, the USCB applied noise-induced error to populations vulnerable to potential reidentification. Now, under differential privacy, all data are subject to noise infusion.

and ethnic representation in Southern Arizona where the non-white, non-Hispanic population makes up a small proportion of the population in many areas. Less accurate data will not only affect the representation of these communities, but also make it difficult to address the structural problems these communities face.

Differential privacy will affect denominator data that are the foundation of social, economic, and health metrics used in data-driven policy (Hotz and Salvo 2020, Weldon Cooper Center for Public Research 2020; Ruggles et al. 2019). Epidemiologists in the Arizona Department of Health Services exploring the disparate impact of COVID-19 across the Navajo Nation, for example, need accurate data of the American Indian and Alaska Native (AIAN) population. A recent study suggests that unreliable denominator data at the census tract-level will make it difficult to accurately measure COVID-19 mortality rates for certain populations of color (Santos-Lozada et al. 2020).

A second and equally important takeaway from the 2019 CNSTAT meeting is that differential privacy may introduce issues of data justice. As Daniel Barth-Jones (2019, p. 3) points out, data injustice occurs "when benefits of research for individuals are unequally denied and when risks are inequitably distributed." By protecting individual confidentiality, differential privacy yields equitable privacy benefits. The question, however, is at what cost? Given that certain populations, notably individuals of color, are more likely to feel the disproportionate effects of differential privacy—principally more noise rendering less-accurate census data—society must now wrestle with how, and to what extent, differential privacy will impede data-driven and equity-focused governance (Brasch 2020; Ruggles et al. 2019). These are important questions that require more careful consideration.

USCB Data Error

Currently, policymakers and public officials need to consider two primary types of error in decennial census data: 1) content error and, 2) coverage error. Content error refers to demographic errors that emerge from a respondent incorrectly answering a question, either unintentionally or purposefully. In general, content error is not a major source of error in census data. Coverage error, alternatively referred to as a census undercount and overcount, refers to the incorrect omission or inclusion of individuals in a census enumeration, respectively (National Research Council 2007). Census undercounts are a primary concern in any census enumeration.

Certain individuals, including young children, individuals of color, non-English speakers, rural residents, immigrants, non-citizens, renters, the homeless, and many others are unfortunately at risk of not being counted in the census (Jurjevich 2020). These groups are referred to as "hard-to-count" (HTC) populations. In the 2010 census, for example, 1.5% of Hispanic/Latino individuals and 2.1% of African Americans were undercounted, representing roughly 750,000 and 850,000 individuals, respectively (U.S. Census Bureau 2012).

The goal of a census enumeration, as outlined by the USCB, is to count everyone once, only once and in the right place. Census undercounts/overcounts are the benchmarks for assessing the degree to which this goal is realized. Based on these metrics, Census 2010 was one of the most accurate censuses in American history. However, data users need to consider the extent to which coverage error reduces data accuracy, especially among HTC populations, and communicate coverage error to stakeholders. Now, policymakers, academics, and data users must navigate another source of error in Census 2020—

an intentional error introduced by differential privacy. Together, these errors will make it more difficult to accurately represent the sociodemographic stories of marginalized communities.

Differentially-Private Census 2020 Data

Differential privacy will materially affect the availability and reliability of Census 2020 data products, particularly for cross-tabulated data (e.g., age by race/ethnicity), microdata (e.g., Public Use Microdata Sample or PUMS), and for the smallest census geographies (e.g., census blocks). Data users need to recognize that the new Disclosure Avoidance System, under differential privacy, will affect data availability and reliability in three important ways. First, the USCB will implement multiple disclosure avoidance protocols across Census 2020 data products (U.S. Census Bureau 2020c). For Group 1 products, including the redistricting file (i.e., Public Law 94-171), the Bureau will rely on the "Top-Down Algorithm" as its primary safeguard. Group 2 products, which include detailed race/ethnicity files, contain data requiring more extensive privacy safeguards. Therefore, the Bureau will rely on a secondary system of disclosure avoidance protocol for these products. Second, a series of metrics measuring data distortion, bias, and overall accuracy will summarize the error introduced by differential privacy. These metrics, called "fitness-for-use" statistics, will accompany census data products (Devine, Borman, and Spence 2020; U.S. Census Bureau 2020d). Third, certain data products available in previous censuses may not be available in Census 2020 data products, depending on data detail (Devine and Hollingsworth 2019). This will likely impede longitudinal analyses.

Data and Methods

This section summarizes 1) data sources, 2) definitions of race/ethnicity, and 3) methodological approaches for the research questions.

Demonstration Data

In October 2019, the USCB released demonstration data so data users could assess the potential effects of differential privacy. We accessed the October 2019 file from the IPUMS National Historical Geographical Information System (NHGIS) at the University of Minnesota (U.S. Census Bureau 2019b; Manson et al. 2020). The file includes demographic and housing data from Census 2010, and Census 2010 data with the proposed 2020 disclosure avoidance protocol.[3] These data allow us to examine the intentional noise-induced error introduced by differential privacy, expressed as the difference between the "actual" Census 2010 data published under traditional data privacy protocols[4] and differentially-private Census 2010 data.

Race and Ethnicity

The USCB currently relies on individual self-reporting of race and ethnicity according to categories outlined in a directive from the Office of Management and Budget (OMB). Last updated in 1997, the federal government records and reports race and ethnicity separately. Ethnicity refers to Hispanic, Latino, or Spanish origin, while race refers to the following categories: White, Black or African American, American Indian or Alaska Native (AIAN), Asian,[5] Native Hawaiian and Pacific Islander, and Two or More Races.

Differential privacy is based on noise generated at random, regardless of the composition of the population under consideration. In practical terms, this means that in census tracts where Asian individuals make up a small percentage of the population, for example, Asians will wind up with more noise (as a percentage of their population) compared to racial groups with larger populations. To quantify the disproportionate effects among specific racial/ethnic groups, we compare the largest racial/ethnic subgroup, non-Hispanic Whites, as a reference group to the following populations of color: Black/African American, American Indian and Alaskan Natives (AIAN), Asian, Native Hawaiian and Pacific Islanders, Other Race, Two or More Races, as well as Hispanic/Latino. Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

[3] The October 2019 demonstration data products represent an "interim version of the 2020 Disclosure Avoidance System (DAS)". After reviewing the feedback from the December 2019 Committee on National Statistics (CNSTAT) meeting (see Abowd and Velkoff 2020), the USCB made adjustments to the algorithm and released a revised demonstration file in May 2020. Based on our analysis, the 2020 vintage produced greater data distortion for populations of color at the census tract level compared to the 2019 vintage.

[4] Traditional privacy protocols in the 2010 census, for example, include data swapping by race/ethnicity. This means that 2010 census race/ethnic data contain unreported amount of data distortion.

[5] The Asian race category is an aggregation of racial origins, including: Chinese, Filipino, Asian Indian, Other Asian, Vietnamese, Korean, and Japanese.

Research Approach

The geographic scope of our analysis covers Southern Arizona, which includes Cochise, Pima, Santa Cruz, and Yuma Counties. The geographic unit of analysis is the census tract, which is a small, relatively permanent county sub-division containing roughly 2,500 to 8,000 individuals (U.S. Census Bureau 2019c). There are three main advantages for conducting our analysis at the census-tract level: 1) census tracts roughly approximate neighborhoods, 2) census tracts are consistent, allowing for longitudinal analysis and, 3) census tracts are often preferred census geographies for public policy.

In this white paper, we assess the impact of differential privacy across Southern Arizona by examining two questions: 1) how reliable are differentially-private data for Arizonans of color? 2) to what extent does differential privacy introduce unequal data distortion among Arizonans of color at sub-county geographies (i.e., census tracts)?

To assess the reliability of differentially-private data for Arizonans of color, we calculate the median absolute error and median relative (i.e., percent) noise-induced error from the Census 2010 demonstration data. We report the mean and median error values by race/ethnicity for Southern Arizona, and also by county (i.e., Cochise, Pima, Santa Cruz, and Yuma). These data points establish a descriptive baseline for quantifying the unequal adjustments among communities of color across Southern Arizona.

We first construct an inferential statistical model to help establish whether differential privacy introduces unequal data distortion among Arizonans of color across census tracts.[6] We classify census tracts as urban if at least 50% of the population is within a census designated urban area; all other census tracts are considered rural. We expect that after controlling for population size and urban-rural status, census tracts where communities of color make up a small percentage of the population will experience greater data distortion from differential privacy.

Second, we create an "Error Index" as an aggregate score of racial/ethnic groups whose data are disproportionately affected by differential privacy error. The Error Index ranges from 0 to 5,[7] and is calculated by comparing the absolute and relative errors for each racial/ethnic group to the respective median error across all Southern Arizona census tracts. A value of 1 is assigned to a specific racial/ethnic group if both the absolute and relative error are at least one standard deviation greater than the median errors across all census tracts. In other words, a racial/ethnic group with a value of 1

[6] The statistical model is an Ordinary Least Squares (OLS) regression model. The model's dependent variable, differential privacy error (relative error), is predicted using the following independent variables: urban-rural census tract classification; race/ethnicity population (i.e., White, Black/African American, American Indian or Alaska Native (AIAN), Asian, Native Hawaiian and Pacific Islanders, Two or More Races, as well as Hispanic/Latino), and; small (i.e., 0% - 25% quartile), medium (i.e., 25-50% quartiles), and large (i.e., 75-100% quartile) populated census tracts.

[7] Our analysis examines differential privacy error for eight racial/ethnic groups, so the Error Index theoretically ranges from 0 to 8. However, our analysis shows that the maximum Error Index across all Southern Arizona census tracts is 5.

indicates that the noise-induced error under differential privacy is greater than 85% of tract-level errors across Southern Arizona.[8]

The Error Index for Census Tract 33.04, located in the Midtown-East area of Tucson, is shown in Figure 1. This census tract has an Error Index of 3, which indicates there are three racial/ethnic groups—Hispanic/Latino, Two or More Races, and White—with disproportionately high noise-induced error. Put another way, the absolute and relative errors among Hispanic/Latino, Two or More Races, and White populations in Census Tract 33.04 rank in the top 15 percent of cases across Southern Arizona. The errors for the other racial/ethnic categories—designated with "0"—are below the 85th percentile.

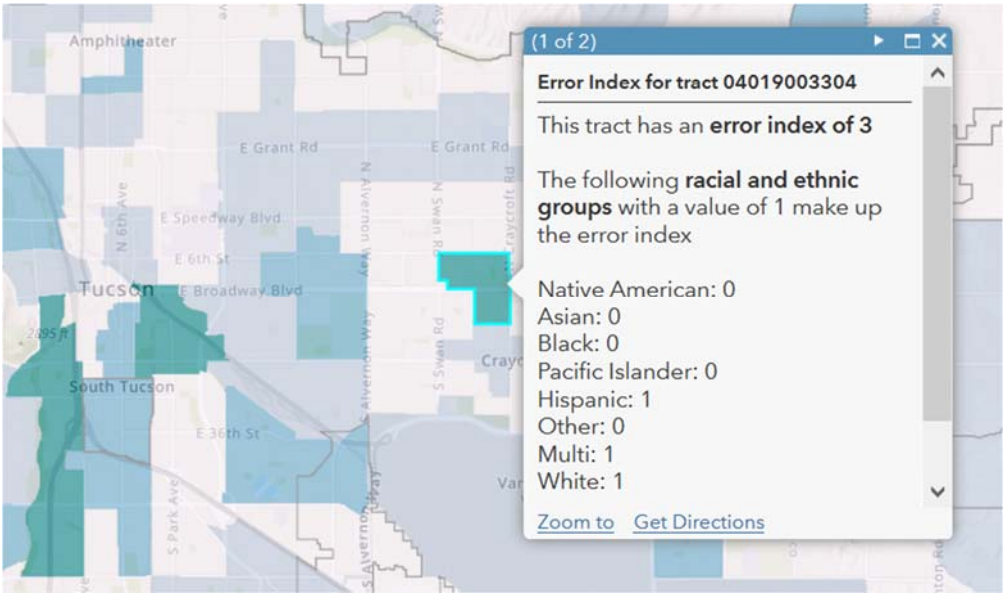

Figure 1. Illustration of Error Index by Racial/Ethnic Group, Census Tract 33.04

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

The components of the Error Index for the Hispanic/Latino and Black/African American populations in Census Tract 33.04 are shown in Table 1. For the Hispanic/Latino population, we calculate the index by first comparing the Census 2010 reported population of 809 with the differentially-private value of 888. The difference yields an absolute error of 79 individuals and a relative error of 9.8%. Second, we compare these error values to the corresponding error values across all Southern Arizona census tracts (i.e., median absolute error of 27 individuals, and a relative error of 2.0%). Third, we calculate the absolute and relative deviation using the following formula:

$$Deviation = \frac{Error - median(error)}{median\ absolute\ deviation(error)}$$

[8] In a normal distribution, one standard deviation contains 34.1% of cases. The 85 percent statistic is based on one standard deviation above the mean/median, so 50% + 34.1% is roughly 85%.

In this example, data distortion introduced by differential privacy for the Hispanic/Latino population in Census Tract 33.04 yields an absolute deviation of 2.2 (i.e., [79-27]/24) and a relative deviation of 3.6 (i.e., [9.8%-2.0%]/2.2%). Given that both of these error values exceed 1.0, the Error Index is set to 1 for the Hispanic/Latino population (see Figure 1). This means that the Hispanic/Latino population in Census Tract 33.04 experiences above-average data distortion under differential privacy, and the noise-induced error is within the top 15 percent of all census tracts in Southern Arizona.

	Population		Error		Deviation		Group Index (i.e. Abs. & Rel. Dev > 1?)
Hispanic/Latino	Actual	DP	Absolute	Relative	Absolute	Relative	
Census Tract 33.04	809	888	79	9.8%	**2.2**	**3.6**	**Yes**
Southern AZ Median			*27*	*2.0%*	*24*	*2.2%*	
Black/African American							
Census Tract 33.04	171	122	49	28.7%	**4.3**	**0.8**	**No**
Southern AZ Median			*11*	*15.0%*	*9*	*17.4%*	

Table 1. Deviation and Error Index Calculations for Hispanic/Latino and Black/African American Population, Census Tract 33.04

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

We repeat this calculation for the Black/African American population in Census Tract 33.04 (see Table 1). Here, the actual Census 2010 population of 171 and differentially-private value of 122 yields an absolute error of 49 individuals and a relative error of 28.7%. The median absolute error among the Black/African American population across all Southern Arizona census tracts is 11 individuals, and a relative error of 15.0%. Using the deviation formula above produces an absolute deviation of 4.3 (i.e., [49-11]/9) and a relative deviation is 0.8 (i.e., [28.7%-15.0%]/17.4%). Given that both error values do *not* exceed 1.0, the Error Index is set to 0 for the Black/African American population in this census tract (see Figure 1).

The takeaway for Census Tract 33.04 is that noise-induced error for the Hispanic/Latino population is in the top 15 percent of tracts across Southern Arizona, but not for the Black/African American population. This does not mean, however, that the error introduced by differential privacy for the Black/African American population in Census Tract 33.04 is not significant or relevant; it simply means that the error is not one of the most egregious examples of data distortion for the Black/African American population in Southern Arizona.

Analysis and Findings

How reliable are differentially-private data for Arizonans of color?

The differential privacy error by race/ethnicity across Southern Arizona census tracts is presented in Table 2. As expected, noise-induced error is higher among non-Hispanic/Latino populations of color. The degree to which differential privacy error renders data unreliable is particularly significant. For example, the median absolute error among the Native Hawaiian and Pacific Islander population (i.e., 3) equals the median actual population (i.e., 3), yielding a median relative error of 100%. Sizable relative errors also exist among the Other (75%), American Indian and Alaska Native (AIAN) (26.7%), Two or More Races (24.1%), Asian (16.1%), and Black/African American (14.5%) populations (Table 2). Differential privacy error among the non-Hispanic White and Hispanic/Latino populations, however, is negligible with a median relative error of 0.6% and 2.7%, respectively. In the end, smaller populations (e.g., AIAN with a median actual population of 30) wind up with relatively more data distortion than larger populations (e.g., non-Hispanic White and Hispanic/Latino populations[9] of more than 1,000) (Table 2). The result is unequal data distortion that disproportionately disadvantages non-Hispanic/Latino populations of color in Southern Arizona.

	Median Actual Population	Median Absolute Error	Relative Error
White	1,928	12	0.6%
Hispanic/Latino	1,002	27	2.7%
American Indian and Alaskan Native	30	8	26.7%
Asian	62	10	16.1%
Black/African American	76	11	14.5%
Native Hawaiian and Pacific Islander	3	3	100.0%
Other	4	3	75.0%
Two or More Races	58	14	24.1%

Table 2. Median Population and Error by Race/Ethnicity for Southern Arizona Census Tracts

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

In Table 3, we report the median actual population and differential privacy error by race/ethnicity for Pima, Cochise, Santa Cruz, and Yuma Counties. Here, the relative differential privacy error for the non-Hispanic White and Hispanic/Latino populations is fairly consistent across counties and is generally similar to data presented in Table 2. Data presented in Table 3 illustrate again that differential privacy error varies according to population size. For example, because the Black/African American population is larger in Pima County compared to Santa Cruz County (i.e., median actual population of 78 and 15 individuals, respectively), the relative differential privacy error is lower in Pima County (i.e., 14.1%

[9] The non-Hispanic, White and Hispanic/Latino populations comprise more than 75% of the population in Southern Arizona.

compared to 44.8%). Differential privacy, under this scenario, introduces error rendering census data unreliable for most communities of color across Southern Arizona census tracts.

	Pima		Cochise		Santa Cruz		Yuma	
	Median Actual Population	Relative error	Median Actual Population	Relative error	Median Actual Population	Relative error	Median Actual Population	Relative error
White	2,169	**0.6%**	2,218	**0.6%**	217	**6.7%**	1,223	**0.7%**
Hispanic/Latino	929	**2.9%**	989	**2.9%**	3,784	**0.9%**	1,448	**1.9%**
American Indian and Alaskan Native	32	**28.1%**	23	**30.4%**	12	**41.7%**	26	**23.1%**
Asian	91	**13.3%**	44	**27.6%**	7	**35.7%**	31	**25.8%**
Black/African Americai	78	**14.1%**	30	**28.8%**	15	**44.8%**	19	**42.1%**
Native Hawaiian and Pacific Islander	3	**133.3%**	7	**76.9%**	0	--	2	**100.0%**
Other	5	**80.0%**	6	**50.0%**	2	**100.0%**	2	**100.0%**
Two or More Races	68	**25.0%**	71	**21.8%**	11	**81.0%**	33	**39.4%**

The percentage error for Pacific Islanders in Santa Cruz cannot be calculated because the denominator, average population, is 0

Table 3. Median Population and Error by Race/Ethnicity by Southern Arizona County

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

Next, we break down differential privacy error by race/ethnicity according to the population size of each census tract (Table 4). A census tract with a racial/ethnic population in the bottom 25% of all Southern Arizona census tracts is considered small; a population in the middle 50% of tracts (i.e., between the 25th and 75th quartile) is medium; and a population in the top 25% of tracts is large. Results presented in Table 4 underscore two important points. First, absolute differential privacy error by race/ethnicity is consistent across different population sizes, but relative error increases as population size decreases. For the AIAN population, for example, census tracts with large populations have a relative error of 10% (i.e., 10/100) compared to 75% error (i.e., 6/8) in census tracts with small populations. And among census tracts with small populations, all non-Hispanic populations of color have a relative error that is nearly 50% or more of the actual population. Second, the relative error for the largest non-Hispanic populations of color is greater than the relative error for the smallest non-Hispanic White and Hispanic/Latino populations, on average. In other words, data distortion for non-Hispanic White populations in small census tracts (i.e., 1.7%)—with the greatest risk of data disclosure—is still smaller than the relative error in census tracts with large populations of color (e.g., 7.9% among the Asian population). Clearly, the burden of protecting respondent confidentiality is not shared equitably across populations and disproportionately affects non-Hispanic populations of color in Southern Arizona.

	Small: 0% - 25%		Medium: 25% - 75%		Large: 75% - 100%	
	Median Pop	Rel. error	Median Pop	Rel. error	Median Pop	Rel. error
White (NH)	461	**1.7%**	1,928	**0.6%**	3,546	**0.4%**
Hispanic	351	**6.0%**	1,002	**2.5%**	3,362	**0.9%**
Nat. Am. (NH)	8	**75.0%**	30	**30.0%**	100	**10.0%**
Asian (NH)	11	**54.5%**	60	**18.3%**	178	**7.9%**
Black (NH)	12	**58.3%**	76	**14.5%**	253	**5.5%**
Pacific Is. (NH)	0	--	3	**100.0%**	13	**69.2%**
Other (NH)	0	--	4	**75.0%**	11	**63.6%**
Multi (NH)	17	**47.1%**	58	**25.9%**	121	**17.4%**

The percentage error for Pacific Islanders and Other cannot be calculated because the denominator, average population, is 0

Table 4. Median Population and Percentage Error by Race/Ethnicity for Quartile Ranges of Southern Arizona Census Tracts

Note: Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

In conclusion, how reliable are differentially-private data for Arizonans of color?

Our analysis reveals four key takeaways:

1. Differential privacy error is severe for non-Hispanic populations of color and negligible for non-Hispanic White and Hispanic/Latino populations in Southern Arizona. The effects are consistent across Pima, Cochise, Santa Cruz, and Yuma Counties.
2. Differential privacy error increases as population size decreases, although the data distortion is unequal. For example, the largest non-Hispanic populations of color experience relatively greater data distortion than the smallest non-Hispanic White and Hispanic/Latino populations.
3. Differential privacy error renders data for non-Hispanic populations of color unreliable across many census tracts. Across census tracts with small populations, non-Hispanic populations of color have a relative error that is roughly 50% of the actual population.
4. Differential privacy introduces data injustice because non-Hispanic populations of color are more likely to shoulder the disproportionate effects of differential privacy—more noise rendering less accurate data. This makes it impractical to achieve equity-focused policy for small-area geographies across Southern Arizona.

To what extent does differential privacy introduce unequal data distortion among Arizonans of color across census tracts?

To tackle this question, we constructed an inferential statistical model to establish the extent to which differential privacy error disproportionately affects Arizonans of color. The statistical results—detailed in the Table A1 of the Appendix —reveal that census tract population size and race/ethnicity, not urban-

rural status,[10] are statistically significant explanatory factors of differential privacy error. In other words, Southern Arizona policymakers must consider both the population size (i.e., small, medium, and large) *and* the race/ethnicity population of a given census tract to fully account for the effects of differential privacy error. Finally, data distortion among most non-Hispanic populations of color[11] in Southern Arizona remains statistically significant even after controlling for the effect of census tract population size. This finding empirically suggests that the burden of protecting respondent confidentiality is not shared equitably across populations and disproportionately affects non-Hispanic populations of color in Southern Arizona. Taken as a whole, these results further underscore issues of data justice.

Next, we present Error Index maps (Figures A1-A4) to illustrate the geography of unequal data distortion across Southern Arizona. The Error Index is an aggregate measure of absolute and relative error introduced by differential privacy. An index value is calculated for each racial/ethnic group and for each census tract in Southern Arizona. Census tracts with lower levels of data distortion (i.e., fewer racial/ethnic groups with disproportionately high error) are in light blue, while census tracts with higher data distortion (i.e., more racial/ethnic groups with disproportionately high error) are designated in dark green. The Error Index maps reveal the following county-level takeaways:

- Santa Cruz. Differential privacy error severely distorts census data for two racial/ethnic groups (Other, and Two or More Races) in a census tract near Nogales, AZ (Figure A1). Data distortion is also significant in six other census tracts, which have severe distortion for one race/ethnic group.
- Pima. More than half of all census tracts in Pima County—127 out of 234, or 54%—have at least one racial/ethnic group with severe differential privacy error (Figure A2). Of these 127 census tracts, 36 (15%) have two racial/ethnic groups with severe differential privacy error. Differential privacy error severely distorts census data for three or more racial/ethnic groups in 13 (6%) census tracts. Most census tracts are located in and around the greater Tucson metropolitan area.
- Cochise. Census tracts with severe differential privacy error in Cochise County are near Sierra Vista and Douglas (Figure A3). Overall, 20 out of 32, or 63% of census tracts in Cochise County have at least one racial/ethnic group with extreme levels of data distortion.
- Yuma. Differential privacy error is severe in at least one racial/ethnic group in 28 out of 55, or 51% of census tracts in Yuma County (Figure A4). Several census tracts east of the City of Yuma contain more than one racial/ethnic group with severe data distortion.

To allow for hands-on exploration and greater understanding of how differential privacy error affects census data, we created a web-based map application and dashboard (Figure 2). The dashboard reports the data we used in our analysis and is accessible here. The primary upside of the dashboard is that it allows users to compare the actual 2010 population to the 2010 demonstration data by race/ethnicity. The tool also reports the absolute and relative error for each racial/ethnic group. Users can customize their selection to one or multiple census tracts to see how differential privacy impacts community

[10] Urban-rural status emerged as a non-statistically significant factor for explaining differential privacy error when controlling for race/ethnicity and census tract population size.

[11] Specifically, the following racial/ethnic groups are statistically significant in the model: Black/African American, American Indian or Alaska Native, Asian, Native Hawaiian and Pacific Islander, and Two or More Races.

representation for a given area, and begin to strategize for the looming consequences related to policy and planning.

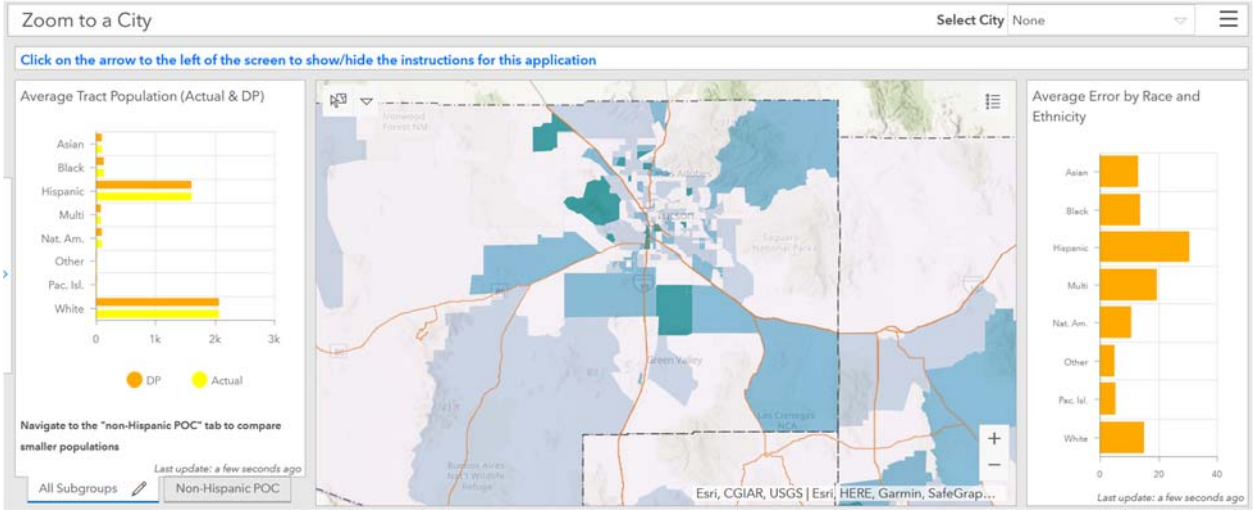

Figure 2. Southern Arizona Differential Privacy Dashboard

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

In conclusion, to what extent does differential privacy introduce unequal data distortion among Arizonans of color across census tracts? Our analysis reveals three important takeaways:

1. Population size and race/ethnicity, not urban-rural status, are statistically significant explanatory factors of differential privacy error in Southern Arizona census tracts.
2. More than half of all census tracts—55%—in Santa Cruz, Pima, Cochise, and Yuma Counties have at least one racial/ethnic group with severe data distortion.
3. Our web-based map application and dashboard allows users to explore the geography of differential privacy error by race/ethnicity across all Southern Arizona census tracts.

Conclusion

Providing quality statistical information and protecting respondent confidentiality in today's digitally-mediated world is a difficult and onerous responsibility for the U.S. Census Bureau. Both objectives are achievable vis-a-vis differential privacy, provided there is balance between data privacy and accuracy. This challenge is more difficult for sub-county geographies, population subgroups, and less-populated areas because these groups wind up with proportionately more noise, to guarantee the same level of privacy, compared with larger populations (NCSL 2020; Nagle and Kuhn 2019).

For non-Hispanic populations of color and smaller communities in Southern Arizona, the price for data privacy may be too high. Our analysis of the 2010 demonstration data reveals three important findings for Southern Arizona. First, population size and race/ethnicity are statistically significant factors of differential privacy error. Southern Arizona policymakers must consider both the population size (i.e., small, medium, and large) and the race/ethnicity of a given census tract to fully account for the effects of differential privacy error. Second, results also reveal that differential privacy introduces unequal data distortion among Arizonans of color across census tracts. The largest non-Hispanic populations of color, on average, experience more relative error than the smallest non-Hispanic White and Hispanic populations. In other words, populations of color are more likely to shoulder the disproportionate costs of differential privacy—more noise that renders less accurate data. Third, we find that differential privacy error renders data for non-Hispanic populations of color unreliable across many Southern Arizona census tracts. Generally, non-Hispanic populations of color pay a higher price for data privacy due to their smaller average population size. Among non-Hispanic populations of color in census tracts with small populations, for example, relative error is roughly 50% of the actual population (e.g., a noise-induced population of 100 compared to an actual population of 200). In Santa Cruz, Pima, Cochise, and Yuma Counties, more than half of all census tracts—55%—have at least one racial/ethnic group with severe data distortion.

These findings provide the following salient takeaways:

1. By protecting respondent confidentiality, differential privacy yields equitable benefits, but the burden of protecting respondent confidentiality is not shared equitably across populations. Communities of color, for example, are more likely to feel the disproportionate effects of differential privacy—more noise that renders less accurate data. This inequity introduces issues of data justice. According to Daniel Barth-Jones (2019, p. 3), data injustice occurs "when benefits of research for individuals are unequally denied and when risks are inequitably distributed." Under this definition, differential privacy may not meet the textbook definition of data injustice. However, the unequal burden of costs introduces important moral-ethical questions. The data illustrate how the well-intentioned adjustments of differential privacy can produce considerable harm by distorting and potentially erasing the stories of marginalized communities.

2. Differential privacy introduces data distortion that could make it difficult, and in some cases improbable, to realize data-driven and equity-focused governance (i.e., compared to years past). This is especially true for small-area geographies and disadvantaged groups, including communities of color. For the Making Action Possible (MAP) coalition, measuring progress and fostering collective civic action around equitable housing policy; sustainable transportation

options; economic justice; erasing racial/ethnic discrimination; and many other equity-driven goals could be increasingly difficult to achieve in the years to come.

Measuring disparate health outcomes will be difficult with differentially-private census data. Consider efforts to measure disparities in COVID-19 death rates among communities of color, for example. Recall the data distortion for the Black/African American population in Census Tract 25.01 (Table ES2); the 2010 census reported a population of 243 individuals compared to a differentially-private value of 381. Hypothetically, if 10 Black/African American Tucsonans died due to COVID-19 in this census tract, the noise-induced death rate (per 1,000 people) would be 26.2 (i.e., 1,000* [10/381]). This death rate is considerably lower than the actual death rate under this scenario of 41.2 (i.e., 1,000* [10/243]). Moving forward, researchers and communities will have to contend with similarly latent error in social, economic, and health indicators.

3. Differential privacy error will make it more difficult to use and interpret census data compared to previous censuses. Data users will need to consult a series of metrics—fitness-for-use statistics—to assess data distortion, bias, and overall accuracy of 2020 census data. Our concern is that fitness-for-use statistics will be difficult for many novice data users to interpret, apply, and communicate. Previous research examining how users engage with statistical uncertainty in ACS data suggests that many data users might ignore fitness-for-use statistics, not reporting the accompanying error, or potentially reject census data altogether (Jurjevich et al. 2018).[12]

4. Our analysis contains two limitations. First, our findings are based on the first round of demonstration data. We are unsure how the Bureau's final adjustments will affect 2020 census data. Second, the findings only highlight data distortion for total population by race/ethnicity. The impacts of differential privacy error across other decennial census topics (e.g., age, sex, and housing characteristics) in Southern Arizona remain unexplored. Our analysis also raises an equally important question: how will differentially-private census data affect the accuracy of ACS data products and social science research, more broadly? The ACS, which relies on decennial census data to estimate median household income, commute mode, income/poverty, and other socioeconomic indicators, will now have to contend with differential privacy error, in addition to sampling error. Understanding the potential cascading effects of differential privacy is critically important to fully quantifying the costs and benefits of differential privacy.

In February 2021, the USCB made two important announcements. First, the Bureau announced that the privacy-loss budget for 2020 Census data products will be finalized in Summer 2021 (U.S. Census Bureau 2021). Second, the final privacy-loss budget will contain less noise, yielding greater accuracy for 2020 Census data compared to earlier privacy-loss budgets (including the budget used in the October 2019

[12] In addition to navigating differential privacy error, data users need to be mindful of the challenges and barriers of Census 2020, leading to a potential census undercount. These include, but are not limited to, the proposed citizenship question; increasing distrust in government; inconsistent and insufficient federal funding; the COVID-19 pandemic; unprecedented and devastating fires in the Western United States; growing fears among immigrants in the current social and political environment.

demonstration data).[13] We believe this decision is a step in the right direction and is good news for data users. This also means that the data distortion in 2020 Census data might be less severe than what we report here. In the end, our high-level findings remain relevant for public policy, research, and governance because the well-intentioned adjustments of differential privacy disproportionately distort and potentially silence the stories of marginalized communities. This data distortion could make it difficult, and in some cases improbable, to realize data-driven and equity-focused governance, especially for socially disadvantaged groups and communities of color.

Bringing together proponents of data privacy and the data user community remains a necessary step to ensure that census data continue to empower data-driven and equity-focused governance. In a presentation at the ACS Data Users Group (DUG) meeting in May 2019, Dr. Connie Citro, former director of the Committee on National Statistics of the National Research Council, called on the Bureau to "institutionalize systematic, two-way, transparent interaction—structured input, dialog, preliminary decision, [repeat], and document the final decision" (Citro 2019). Partnering with the Committee on National Statistics of the National Academy of Sciences and others, the U.S. Census Bureau (e.g., National Advisory Committee, NAC) has championed many of these recommendations in recent years.

Moving forward, we encourage even greater outreach to data users and policymakers. Community leaders of color, urban and regional planners working with small-area geographies, and program leaders championing locally focused social justice and equity initiatives, for example, are well-positioned to illustrate how noise-induced distortion affects their work. These testimonials of data distortion are important for grounding the potential erasure of underserved and marginalized communities, and for more fully quantifying the costs of balancing privacy and accuracy. Moreover, this effort could advance discussion surrounding the morality and ethics of unequal data distortion, including how principles of data stewardship might be augmented to ensure data justice for underrepresented and marginalized communities. Efforts that continue democratizing differential privacy among community leaders, academics, policymakers, legislators, and government officials will be important in the months and years ahead.

References

Abowd, John and Victoria Velkoff. 2020. "Modernizing Disclosure Avoidance: What We've Learned and Where We Are now." U.S. Census Bureau. March 13. https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing_disclosu.html

Abowd, John. 2018. "Protecting the Confidentiality of America's Statistics: Ensuring Confidentiality and Fitness-for-Use." U.S. Census Bureau. September 4. https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi0.html

Barth-Jones, Daniel C. 2019. "Differential 'Privacy Guarantees' and Ethical Equipoise." Presentation at *Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations*. Convened by the Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and

[13] The USCB will release one more demonstration dataset available for analysis using a less conservative privacy-loss budget by April 30, 2021.

Medicine. December 11-12, Washington, DC.
https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_197518.pdf

Beveridge, Andrew A. 2019. "Impacts of Redistricting: The Case of New Rochelle, NY." Presentation at *Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations*. Convened by the Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and Medicine. December 11-12, Washington, DC.
https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_197494.pdf

Brasch, Alex. 2020. "A Comparison of Census 2010 SF1 & Differentially Private Data in Oregon."
https://github.com/a-brasch/PSU-USP522-Practicum

Citro, Constance F. 2019. "Dissemination of ACS Data: Looking Ahead—Discussion." Presentation at 2019 ACS Data Users Conference. May 14-15, Washington, DC. Link

Devine, Jason, Borman, Christine, and Matthew Spence. 2020. "2020 Census Disclosure Avoidance Improvement Metrics." U.S. Census Bureau. Presentation to the Committee on National Statistics, Disclosure Avoidance Working Group. March 18. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf

Devine, Jason and Cynthia Hollingsworth. 2019. "Status Update on 2020 Census Data Products Plan." U.S. Census Bureau. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products.html

ESRI. 2019. Census 2010 Summary File 1 vs. Differential Privacy: Compare Places. [Interactive Dashboard]. https://arcgis-content.maps.arcgis.com/apps/opsdashboard/index.html#/04451f90e7b049f39aa6647a41b986ac

Garfinkel, Simon L., Abowd, John M., and Sarah Powazek. 2018. "Issues Encountered Deploying Differential Privacy." *WPES 18: Proceedings of the 2018 Workshop on Privacy in the Electronic Society*. 133–137. https://doi.org/10.1145/3267323.3268949

Gates, Gerald W. 2000. "Confidentiality: Keeping the Promise" in *Encyclopedia of the U.S. Census.* Edited by Margo J. Anderson. Washington, DC: CQ Press.

Hawes, Michael B. 2020. "Implementing Differential Privacy: Seven Lessons From the 2020 United States Census." *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.353c6f99

Hotz, V. Joseph and Joseph Salvo. 2020. "Assessing the Use of Differential Privacy for the 2020 Census: Summary of What We Learned from the CNSTAT Workshop."
https://www.amstat.org/asa/files/pdfs/POL-CNSTAT_CensusDP_WorkshopLessonsLearnedSummary.pdf

Jarmin, Ron. 2018. "The Balancing Act of Producing Accurate and Confidential Statistics." U.S. Census Bureau. December 14.
https://www.census.gov/newsroom/blogs/director/2018/12/the_balancing_actof.html

Jarosz, Beth. 2019. "Importance of Decennial Census for Regional Planning in California." Presentation at *Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations*. Convened by the Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and Medicine. December 11-12, Washington, DC.
https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_197498.pdf

Jurjevich, Jason R. 2020. "Toward an Accurate Census: Projections of Arizona's Hard-to-Count (HTC) Population in Census 2020." *Census 20/20*.
https://www.census2020now.org/s/AZ_Census2020_HTCReport.pdf

Jurjevich, Jason R., Griffin, Amy L, Spielman, Seth E., Folch, David C., Merrick, Meg and Nicholas N. Nagle. 2018. "Navigating Statistical Uncertainty: How Urban and Regional Planners Understand and Work with American Community Survey (ACS) Data for Guiding Policy." *Journal of the American Planning Association*. 84(2): 112-126. DOI: 10.1080/01944363.2018.1440182

Manson, Steven, Schroeder, Jonathan, Van Riper, David, Kugler, Tracy and Steven Ruggles. 2020. IPUMS National Historical Geographic Information System: Version 15.0 [dataset]. Minneapolis, MN: IPUMS. http://doi.org/10.18128/D050.V15.0

Moore, Richard A. 1996. *Controlled Data Swapping Techniques for Masking Public-Use Microdata Files*.
https://www.census.gov/srd/CDAR/rr96-04_Controlled_DataSwapping.pdf

Nagle, Nicholas and Tim Kuhn. 2019. "Implications for School Enrollment Statistics." Presentation at *Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations*. Convened by the Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and Medicine. December 11-12, Washington, DC.
https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_197492.pdf

National Research Council. 2007. *Research and Plans for Coverage Measurement in the 2010 Census: Interim Assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11941

National Council of State Legislatures (NCSL). 2020. "Differential Privacy for Census Data Explained."
https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx

Reamer, Andrew. 2020. "Counting for Dollars: The Role of the Decennial Census in the Geographic Distribution of Federal Funds." George Washington University Institute of Public Policy.
https://gwipp.gwu.edu/sites/g/files/zaxdzs2181/f/downloads/IPP-1819-3%20CountingforDollars_AZ.pdf

Ruggles, Steven, Fitch, Catherine, Magnuson, Diana, and Jonathan Schroeder. 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

Ruggles, Steven. 2018. "Implications of Differential Privacy for Census Bureau Data Dissemination." Presentation to Federal Economics Statistics Advisory Committee (FESAC).
https://www.census.gov/content/dam/Census/about/about-the-bureau/adrm/FESAC/meetings/Ruggles%20Presentation%20Revised.pdf

Santos-Lozada, Alexis R., Howard, Jeffrey T., and Ashton M. Verdery. 2020. "How Differential Privacy Will Affect Our Understanding of Health Disparities in the United States." *PNAS*. 117(24): 13405-13412. https://doi.org/10.1073/pnas.2003714117

U.S. Census Bureau. 2012. "2010 Census Coverage Measurement Estimation Report #2010G-01." Authored by Thomas Mule. https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g01.pdf

U.S. Census Bureau. 2019a. "A History of Census Privacy Protections." October 10. https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html

U.S. Census Bureau. 2019b. 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

U.S. Census Bureau. 2019c. Glossary for "Census Tract." https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13

U.S. Census Bureau. 2020a. List of Surveys and Programs. https://www.census.gov/programs-surveys/surveys-programs.html

U.S. Census Bureau. 2020b. Data Stewardship. https://www.census.gov/about/policies/privacy/data_stewardship.html

U.S. Census Bureau. 2020c. "Why aren't all 2020 Census data products protected using the same Disclosure Avoidance System?" Frequently Asked Questions (FAQ).

U.S. Census Bureau. 2020d. "Revised Data Metrics for 2020 Disclosure Avoidance." November 16. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

U.S. Census Bureau. 2021. "2/23/2021: The Road Ahead: Upcoming Disclosure Avoidance System Milestones." *2020 Disclosure Avoidance System Updates.* https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html

Van Riper, David, Kugler, Tracy, and Steven Ruggles. 2020. "Disclosure Avoidance in the Census Bureau's 2010 Demonstration Data Product." In *Privacy in Statistical Databases*. Edited by Domingo-Ferrer, J and K. Muralidhar. PSD 2020, LNCS 12276, 353-368. https://doi.org/10.1007/978-3-030-57521-2_25

Van Riper, David and Seth Spielman. 2019. "Geographic Review of Differentially Private Demonstration Data." Presentation at *Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations*. Convened by the Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and Medicine. December 11-12, Washington, DC. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_197491.pdf

Weldon Cooper Center for Public Research. 2020. “2020 Census Data Distortion.” https://sdcclearinghouse.files.wordpress.com/2020/01/censusdistortionprogram_vagovernor_2020-01-23.pdf

Wezerek, Gus and David Van Riper. 2020. “Changes to the Census Could Make Small Towns Disappear.” *New York Times*. February 6. https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html

Wisniewski, William. 2016. “Challenges Facing the Disclosure Review Board.” U.S. Census Bureau. August 9. https://www.census.gov/newsroom/blogs/research-matters/2016/08/challenges-facing-the-disclosure-review-board.html

Appendix

Urban/Rural, Race & Ethnicity, and Quartile Range Predicting Pct. Error				
	B	std.error	statistic	sig
(Intercept)	0.57	0.11	5.27	*
Urban	-0.10	0.08	-1.18	
Black	0.44	0.10	4.29	*
Nat. Am	0.39	0.10	3.83	*
Asian	0.36	0.10	3.53	*
Pac. Isl.	1.69	0.10	16.14	*
Other	0.16	0.10	1.57	
Multi	0.31	0.10	3.05	*
Hispanic	0.12	0.10	1.21	
25% - 75% (Medium)	-0.58	0.06	-8.89	*
75% - 100% (Large)	-0.79	0.08	-10.27	*
R-squared= 15%				

Table A1. Statistical Model Results for Predicting Relative Differential Privacy Error

Note: A (*) indicates statistical significance at the 95% confidence level. Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

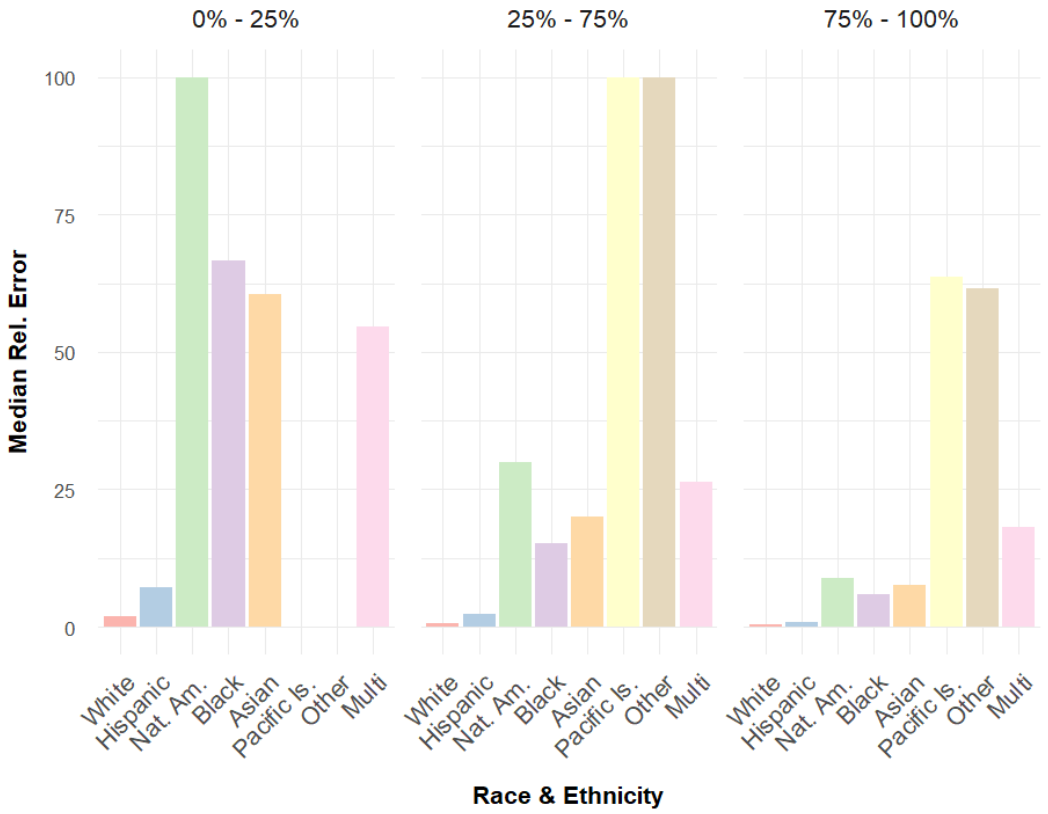

Figure A1. Median Relative Error by Race/Ethnicity for Quartile Ranges of Southern Arizona Census Tracts

Note: This figure visualizes data presented in Table 3. Racial groups are reported as individuals *not* identifying as Hispanic/Latino.

Source: Calculated by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

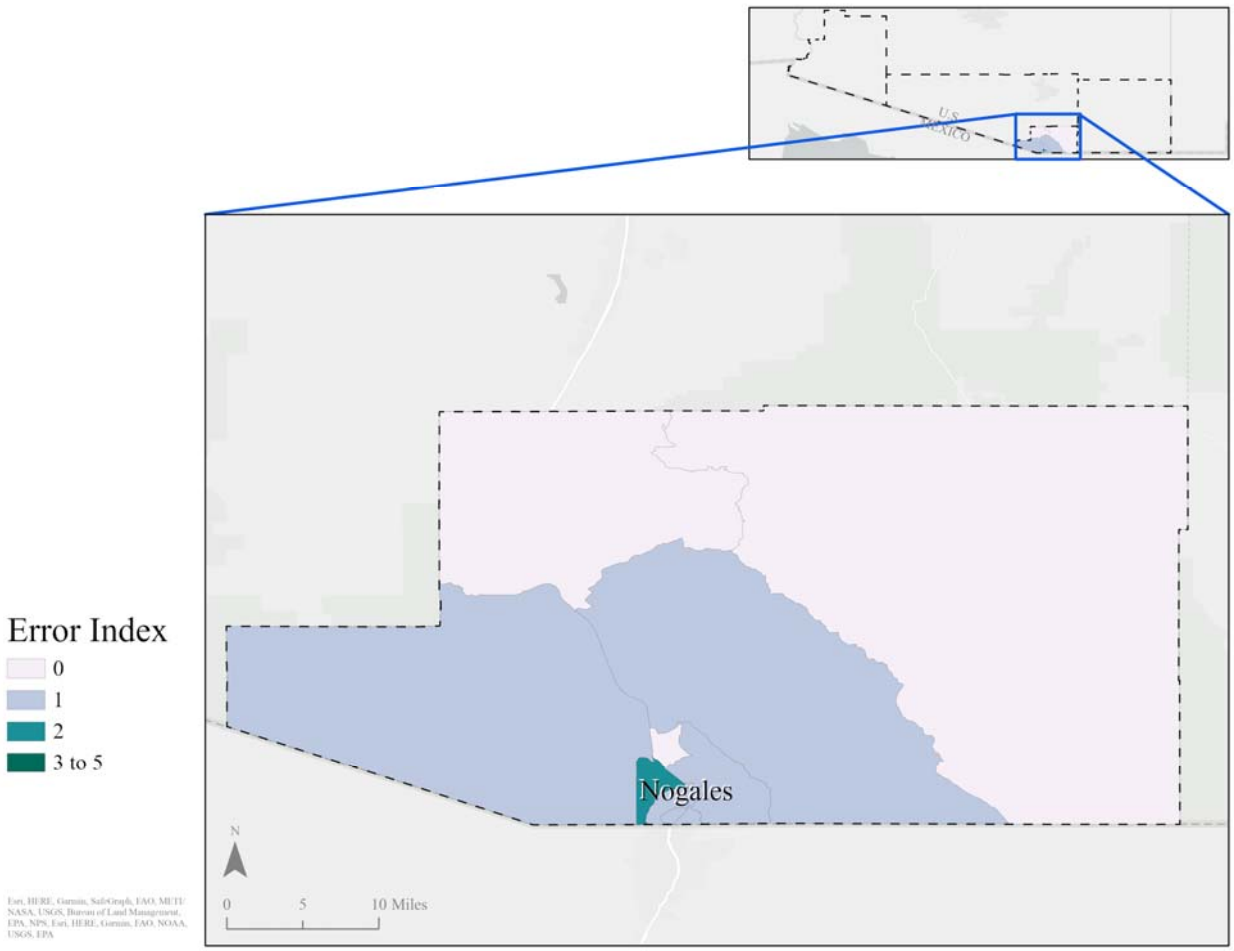

Figure A2. Error Index for Santa Cruz County Census Tracts

Source: Map created by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

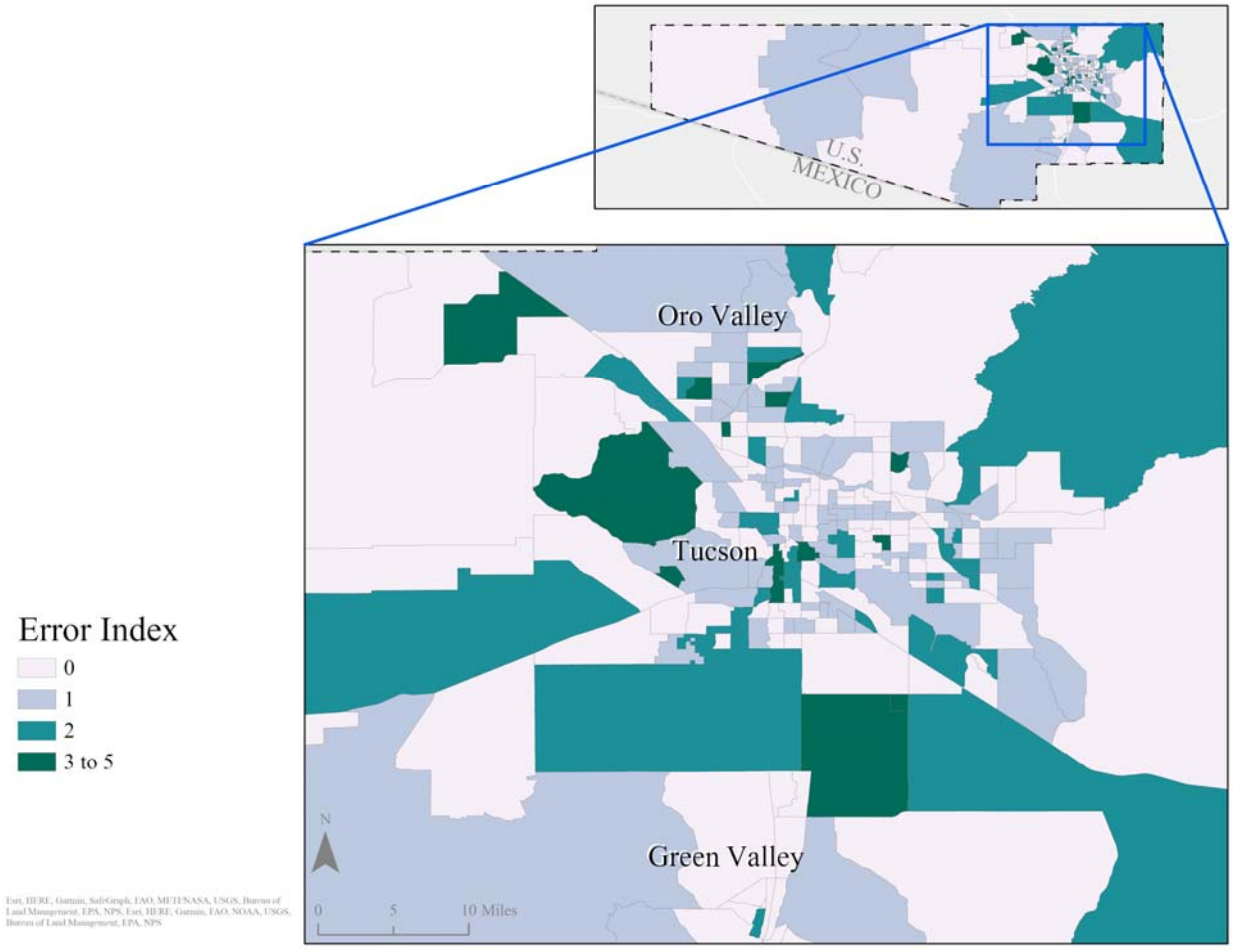

Figure A3. Error Index for Pima County Census Tracts

Source: Map created by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

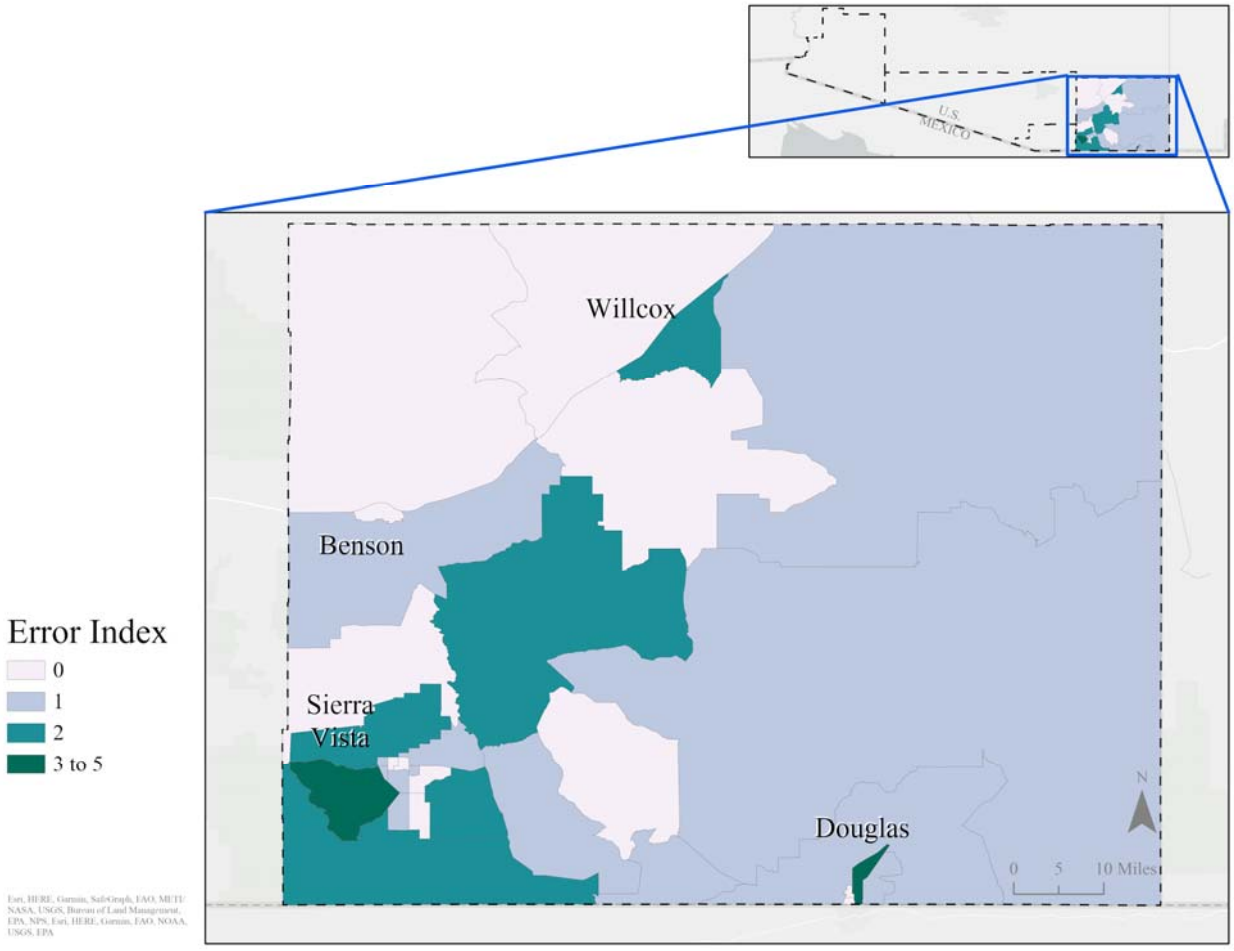

Figure A4. Error Index for Cochise County Census Tracts

Source: Map created by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).

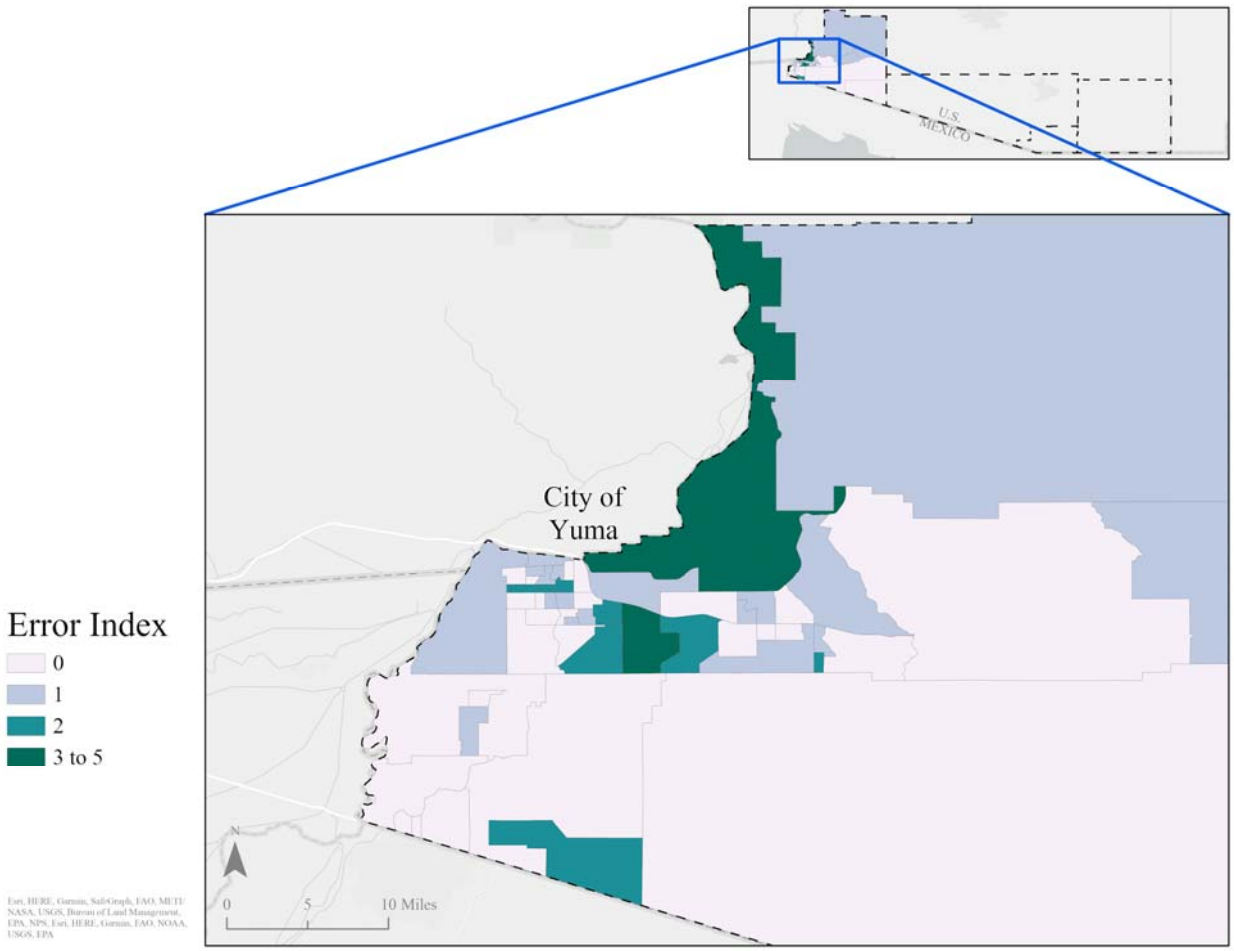

Figure A5. Error Index for Yuma County Census Tracts

Source: Map created by authors using data from the U.S. Census Bureau (2019b), accessed through Manson et al. (2020).