Disclosure Avoidance for the 2020 Census: An Introduction

Issued November 2021

United States® Census Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Acknowledgments

Beth Jarosz, Program Director, U.S. Programs, Population Reference Bureau (PRB), **Mark Mather**, Associate Vice President, U.S. Programs, PRB, and **Linda A. Jacobsen**, Vice President, U.S. Programs, PRB, drafted portions of this handbook in partnership with the U.S. Census Bureau's 2020 Census Data Products and Dissemination Team.

Jason Devine, **Michael Hawes**, **Michele Hedrick**, **Cynthia Hollingsworth**, **Meghan Maury**, **Thomas Morton**, **Kimberly Quick**, **Letha Rubin**, **Matthew Spence**, and **James Whitehorne**, Census Bureau, contributed to the planning and review of the handbook.

The Data Products and Dissemination Operation is under the direction of **Albert E. Fontenot Jr.**, Associate Director for Decennial Census Programs, **Deborah M. Stempowski**, Assistant Director for Decennial Census Programs, and **Jennifer Reichert**, Division Chief, Decennial Census Management Division.

Other individuals from the Census Bureau who contributed to the review and release of this handbook include **John M. Abowd**, **Victoria A. Velkoff**, **Karen Battle**, **Philip Leclerc**, **Dan Kifer**, **Ryan Cumings**, and **Pavel Zhuravlev**.

Stacey Barber, **Corey Beasley**, **Faye Brock**, **Christine Geter**, and **Paula Lancaster** provided publication management, graphic design and composition, editorial review, and 508 compliancy for electronic media and print under the direction of **Linda Chen**, Acting Chief of the Graphic and Editorial Services Branch, Public Information Office.

Disclosure Avoidance for the 2020 Census: An Introduction

Issued November 2021

U.S. CENSUS BUREAU
Ron S. Jarmin,
Acting Director

Suggested Citation

U.S. Census Bureau,
Disclosure Avoidance for the 2020 Census: An Introduction,
U.S. Government Publishing Office,
Washington, DC,
November 2021.

U.S. CENSUS BUREAU

Ron S. Jarmin,
Acting Director

Ron S. Jarmin,
Deputy Director and
Chief Operating Officer

Michael T. Thieme,
Senior Advisor to the Deputy Director for
Information Technology and Operations

Albert E. Fontenot, Jr.,
Associate Director for
Decennial Census Programs

Deborah M. Stempowski,
Assistant Director for
Decennial Census Programs

Victoria A. Velkoff,
Associate Director for
Demographic Programs

John M. Abowd,
Associate Director for
Research and Methodology

Contents

This page is intentionally blank.

1. DISCLOSURE AVOIDANCE FOR 2020 CENSUS REDISTRICTING DATA: AN INTRODUCTION

Background

What is disclosure avoidance and why does it matter? At the U.S. Census Bureau, disclosure avoidance is defined as a process used to protect the confidentiality of respondents' personal information. Since the 1990 Census, the Census Bureau has protected confidentiality by adding "noise"—or variations from the actual count—to the collected data.

In 2020, millions of Americans responded to the decennial census. The decennial census determines congressional apportionment, is often used by states for redistricting purposes, and informs the allocation of hundreds of billions of dollars in federal funding. The 2020 Census counted more than 331 million people in more than 140 million housing units.

The challenge for the Census Bureau is balancing the need to collect and report these data with the statutory obligation to protect their confidentiality.[1] The Census Bureau's work toward that balance is guided by our privacy principles including necessity, openness, respectful treatment of respondents, and confidentiality.[2]

For data users, the main challenge is understanding how disclosure avoidance works, how it may affect the 2020 Census results (Box 1-1), and how it differs from the disclosure avoidance performed on the 2000 Census and 2010 Census. This report provides an overview of how and why the Census Bureau is applying new disclosure avoidance techniques to the 2020 Census and some of the key implications for those who rely on the data.

Responses Are Protected by Law

The Census Bureau is bound by federal law to protect data provided by or on behalf of respondents and to keep them strictly confidential. Not only is this protection a legal and ethical responsibility, but it also underpins the public trust in the Census Bureau. That trust is critical to the public's willingness to respond to censuses and surveys, which in turn is critical to the quality of data that is central to our mission.

Title 13 of the U.S. Code prohibits the Census Bureau from disclosing any "information reported by, or on behalf of, any particular respondent" and from "[making] any publication whereby the data furnished by any particular establishment or individual under this title can be identified."[3] Office of Management and Budget (OMB) guidance on interpreting confidentiality standards further clarifies that federal agencies are required to consider the broader context of disclosure risk (known as the "mosaic effect") when performing their disclosure reviews: "Before disclosing potential PII or other potentially sensitive information, agencies must consider other publicly available data—in any medium and from any source to determine whether some combination of existing data and the data intended to be publicly released could allow for the identification of an individual or pose another security concern."[4]

In fact, every employee at the Census Bureau takes a lifelong oath to protect all respondent information gathered by the Census Bureau. This oath forms the cornerstone of the Census Bureau's broader culture of data stewardship.

Data stewardship is a comprehensive framework designed to protect information over the course of the information life cycle—from collection to dissemination—and it starts with a commitment to confidentiality that is required by law and designed to maintain public trust. Research conducted by both the Census Bureau and nongovernmental researchers has shown that concerns about privacy and confidentiality are among the reasons most often given by potential respondents for unwillingness to participate in surveys and censuses.[5, 6]

Many commercial vendors collect, sell, and publish data about people living in the United States. While these vendors have access to their own data on name, address, and date of birth, fewer vendors have access to the type of rich demographic data the census collects on characteristics like race, ethnicity, and household relationships.

The information on demographic characteristics that these vendors lack is precisely the sort of information collected by the decennial census. The disclosure of these types of characteristics could not only make it

[1] U.S. Constitution, Article I, Section 2; Title 13 U.S. Code, Sections 8–9; Title 13 U.S. Code, Section 141.

[2] "Our Privacy Principles," <www.census.gov/about/policies/privacy/data_stewardship/our_privacy_principles.html>.

[3] Title 13 U.S. Code, Sections 8–9.

[4] OMB Memorandum M-13-13, <https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>, pp. 4–5.

[5] More information on research conducted by the Census Bureau is available at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/final-analysis/2020-report-cbams-study-survey.html>.

[6] More information on nongovernmental researchers is available at <www.srl.uic.edu/newsletter/issues/2000s/04v35n2-3.pdf>.

Box 1-1. Disclosure Avoidance: Key Considerations for Data Users Working With 2020 Census Redistricting Data

In this handbook, the U.S. Census Bureau's Disclosure Avoidance System is described in the context of the 2020 Census Redistricting Data (P.L. 94-171) Summary File because those are the first 2020 Census data protected using the new privacy procedures. The apportionment counts released earlier, in April 2021, were not subject to these new privacy procedures and were the actual enumerated population counts for each state.

Here is a summary of key considerations and recommendations for data users working with the 2020 Census redistricting data:

- Data for very small geographic areas, such as census blocks, may be noisy and should be aggregated into larger geographic areas before use. (Note: this was also the case for 2000 Census and 2010 Census data.)
- Small population groups may experience larger relative uncertainty. While the absolute error is the same for all groups within the same table, the noise added to small groups will result in higher relative error because the underlying population (the denominator) is smaller. (Note: this was also the case for characteristics data in the 2000 Census and 2010 Census.)
- Counts are consistent within tables, across person tables (P1–P5), across the housing unit table (H1), and across geographies. For example, rows within a table sum up to the parent row and counts for geographic levels add up to totals for parent geographies.
- The disclosure avoidance methods for the redistricting data were designed to allow users to transform the published person-level tables by addition and subtraction across tables. For example, you can subtract Table P3 (voting-age population by race) from Table P1 (total population by race) to obtain the population under the age of 18 by race.
- For a given geography, particularly at the block level, the uncertainty introduced by disclosure avoidance may result in apparent inconsistencies between the population and housing tables, such as more occupied housing units than people.
- Data should not be divided across population and housing tables for small geographic areas such as block groups. For example, values from Table P2 should not be divided by values from Table H1 to obtain the average number of people per household. Users who need less noisy statistics on people per household should wait for the release of the Detailed Demographic and Housing Characteristics File (Detailed DHC).
- As with any census, noise infusion is not the only source of uncertainty in 2020 Census data. In most cases, these other sources of uncertainty in census data are more significant than the uncertainty due to confidentiality protection.[1]

The redistricting data files include certain "invariants"—data that are kept exactly as enumerated with no noise added. Invariant statistics for the 2020 Census redistricting data are:

- Total number of people in each state, the District of Columbia, and Puerto Rico.
- Total number of housing units in each census block.
- Number of occupied group quarters facilities by major group quarters type in each census block (e.g., correctional facilities, nursing facilities, college dorms, and military quarters).

All other population and housing characteristic data, including population counts for every geography below the state level, had noise introduced.

In addition to the invariants noted above, the Census Bureau applies the following additional constraints to the redistricting dataset:

- Population and housing counts must be integers and may not be negative.
- The voting-age population count must not exceed the total population count.
- Counts must be consistent within tables, across tables, and across geographies. For example, the population by race must sum to the total population, and the number of occupied and vacant housing units must sum to the total number of housing units.
- If there are zero housing units and zero group quarters facilities in a geography, then no people may be assigned to that geography.
- Blocks with group quarters facilities must include at least one person for each type of group quarters facility present.

[1] "2020 Census Data Quality," <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/data-quality.html#evaluating>; Declaration of John Abowd, U.S. Census Bureau, State of Alabama v. U.S. Department of Commerce, Appendix B, United States District Court for the Middle District of Alabama Eastern Division, filed April 13, 2021.

easier to target individuals—particularly in vulnerable populations—such as communities of color, same-sex couples, older adults, or parents of very young children—for fraud, enforcement actions, disinformation, or physical or virtual abuse, but it could also undermine the public's trust in the confidentiality of its census response, which could cause people to be less likely to respond to future censuses.

To protect information against disclosure in published tabulations, the Census Bureau uses disclosure avoidance procedures—techniques to disguise data to protect the confidentiality of those data.

Disclosure Avoidance Is Not New

Disclosure avoidance at the Census Bureau is not new. Figure 1.1 provides a summary overview of how census privacy protections have evolved from the 1930 Census to the 2020 Census.

For the 1930 Census, the Census Bureau stopped publishing certain tables for small geographic areas to avoid indirect disclosure. In 1954, privacy protection rules were consolidated into Title 13, U.S. Code. For the 1970 Census, the Census Bureau suppressed certain tables based on the number of people or households in a given area.[7]

In 1990, the Census Bureau began using more sophisticated techniques, such as data swapping, to protect against disclosure. With data swapping, the Census Bureau injects "noise" into the data by swapping records for certain households with those from households with similar characteristics in a nearby area. The Census Bureau does not release information about its specific methods for swapping. While this confidentiality around swapping techniques is important to protect against disclosure, it means that the practice is not transparent to data users, which prevents data users from assessing the impact of those protections on the published data.

The Census Bureau continued to use data swapping to avoid disclosure in the 2000 and 2010 decennial censuses. It also used techniques such as top- and bottom-coding, blank-and-impute algorithms, table and cell suppression, and other methods to protect responses against disclosure.[8]

Big Data, Big Potential Threats

Advances in computing technology and rapid growth in the number of commercially available databases on people and households have increased concerns about data privacy. Published tables from the Census Bureau are increasingly vulnerable to database reconstruction and re-identification attacks—that is, an outside party could, by combining information in published tables, reconstruct the original census responses without names or addresses; link these to external databases (or using personal knowledge about a person) on variables shared in common with the census responses; and, from this linking, infer confidential information about individual census respondents. When a person's census record (including block-level location and name) is correctly inferred by linking with an external dataset, we refer to this as a confirmed or correct re-identification.

Some inferences about confidential information can be achieved with purely statistical information (especially for blocks with many identical records). These inferences rely on aggregate statistical information about groups and do not rely on any individuals' confidential census responses. For example, suppose Alice is trying to learn how Bob responded to the race question, and she already knows Bob lived in Montana at the time of the 2010 Census enumeration. Alice could then review the 2010 Census tables, and because she can find that 89.4 percent of respondents reported "White Alone" in Montana, Alice can guess with high confidence that Bob's census

[7] "Disclosure Avoidance Techniques Used for the 1970 Through 2010 Decennial Censuses of Population and Housing," <www.census.gov/library/working-papers/2018/adrm/cdar2018-01.html>.

[8] Ibid.

Figure 1.1. **A History of Census Privacy Protections**

Source: U.S. Census Bureau.

response was "White Alone." This is an example of an inference based on aggregate statistical information about groups, rather than knowledge of Bob's confidential census response. The Disclosure Avoidance System (DAS) permits accurate inferences based on aggregate statistical information about groups. Bob's census response was one of 989,415 in Montana in 2010, and so, even if Bob had never participated in the census, it would still be easy for Alice to guess that Bob's race is probably "White Alone," just by reviewing the responses of the other participants and guessing that Bob's response would match the most common response.

Re-identification of an individual's confidential census responses, however, can occur when an outside party is able to leverage information from statistics in the published data to reconstruct the individual-level records that were used to generate the published tables. When combined with outside information, this approach allows an outside party to infer with high confidence what an individual's confidential census responses were. Suppose, for example, that on his 2010 Census form Bob reported being "Some Other Race Alone," that Bob was the only resident of his census block, and that Alice knows Bob's address (and subsequently his block). Alice could then easily review the published tables for Bob's block, find that a single person reported "Some Other Race Alone," and, if not for the disclosure avoidance techniques used in 2010 (swapping, especially), guess with complete confidence that Bob reported "Some Other Race Alone." This is an example of a privacy-violating inference—if Bob had not participated in the census, Alice would not be able to infer Bob's race in this way as his block would have a reported count of zero.[9] Because Alice could only learn this information about Bob as a direct result of Bob's data being present in the confidential census responses, this kind of learning is about information unique to Bob's confidential response. Both the household swapping procedures used by the Census Bureau in 2010 and the differentially private algorithms used in the 2020 DAS are intended to control how much can be learned about confidential information, while still allowing users of census data to learn about statistical information.

In the examples with Bob and Alice, Alice already had enough auxiliary or "side" information about Bob to learn about Bob directly from the published census tables, but advances in mathematics and computing now allow Alice to go a step further. She can take the published tables and infer highly accurate, complete record-level responses from them for a large proportion of the U.S. population. This process of inferring complete census records from the published tables is like filling in the missing cells in a giant Sudoku puzzle. In Sudoku, players use logic to infer missing numbers in a grid based on the numbers that are available.

Database reconstruction works in a similar way; every piece of published data makes it easier to infer the underlying records. For example, a person's age may not be published, but it may be possible to reconstruct that person's age based on other data available in the Census Bureau's statistical tables.[10]

For Alice, reconstructing complete records could be useful. Suppose that Alice knows Bob's address and that he is over the age of 18, but a second person also lives on Bob's block, so that Bob's block table had one "Some Other Race Alone" and one "White Alone" person reported in it. Alice could not be sure just from these two counts if Bob reported "Some Other Race Alone" or "White Alone." However, if Alice can reconstruct complete record-level responses and finds that the "Some Other Race Alone" person is of voting age, while the "White Alone" person is under the age of 18, then Alice can infer that Bob must correspond to the "Some Other Race Alone" response.

More generally, Alice might be an outside party armed with not just a small amount of knowledge about a single person, but a large external database. Using this database of information on many different people, Alice could then frequently re-identify individuals simply by finding another set of data that is consistent with the reconstructed records.[11] Once individual-level records have been reconstructed, re-identification of specific individuals in those data is often quite easy. In fact, re-identifications have already occurred with datasets outside of the Census Bureau. In 2006, Netflix released an anonymized list of movie ratings from nearly 500,000 users. Researchers described how they could use this database—in combination with a separate Internet Movie Database that included raters' identities—to identify a Netflix user 96 percent of the time based on just eight movie ratings and the approximate timeframe when a rating occurred.[12] The Census Bureau has recently

[9] We emphasize that the key issue here is that Alice's inference could not have been made without Bob's data being present in the census and could only be made with his data present; this is what makes the inference unique to Bob's census response. That Bob is the only resident of his block and the inference is 100 percent certain, rather than just highly confident, both help to make the example simple. Privacy-violating inference can still take place in blocks with large populations (even if it is more common in small populations) and when an attacker can be confident but not certain.

[10] John M. Abowd et al., "The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau," 2020, available at <www.census.gov/library/working-papers/2020/adrm/modernization-statistical-disclosure-limitation.html>, accessed August 11, 2021.

[11] Simson Garfinkel, John M. Abowd, Christian Martindale, "Understanding Database Reconstruction Attacks on Public Data," *Communications of the ACM*, Volume 62, Number 3, March 2019, pp. 46–53, <https://cacm.acm.org/magazines/2019/3/234925-understanding-database-reconstruction-attacks-on-public-data/fulltext>.

[12] Arvind Narayanan and Vitaly Shmatikov, "How to Break Anonymity of the Netflix Prize Dataset," 2006, <https://arxiv.org/abs/cs/0610105>.

documented re-identification attacks made on its data products by outside researchers who provided the documentation.[13]

Census data present an enticing target for re-identification attacks. As the federal government's largest statistical agency, the Census Bureau publishes a very large number of statistics. The 2010 Census data products included over 150 billion statistics based on 309 million people and 1.9 billion confidential data points. This wealth of published statistics suggests that highly accurate reconstruction of census records may be possible, and, if it is possible, that many re-identifications not attributable purely to statistical information may also be possible, especially in small blocks and subpopulations.

In 2018, the Census Bureau conducted an experiment to simulate database reconstruction based on tables published from the 2010 Census. Analysts began by reconstructing the geographic location (i.e., census block), sex, age, race, and ethnicity of all 309 million individuals in the census. On these records, location (census block) and whether the person was voting age or not were always correct. In addition, for 144 million people or 46 percent of the U.S. population, all five variables were identical to the census responses; an additional 76 million were also identical except for variation of 1 year of age.[14] Next, they linked the reconstructed records with information available through commercial databases and were able to find likely matches for 138 million individuals. From those 138 million likely matches, they were able to confirm 38 percent. Overall, they were able to correctly re-identify about 52 million people or 17 percent of the total U.S. population in 2010.

Reconstructing 100 percent of the 2010 Census records with full accuracy for 46 percent of the U.S. population is alarming. It implies that the combined effect of the released tables no longer meets the existing 2010 Census standards for microdata releases. The 2010 standards for microdata releases allowed a sample of microdata to be published only for geographic areas with at least 100,000 people and with demographic categories of at least 10,000 people nationally. However, the 144 million exactly reconstructed records respect none of these constraints. This set of 144 million records includes census blocks—all with population less than 100,000—and many demographic subpopulations with national counts much smaller than 10,000. This "implicit release" of microdata led the Census Operating Committee in January 2018 to elevate reconstruction to an enterprise-level 2020 Census risk.

A correct re-identification rate of 38 percent in the 138 million linked records is still more alarming. This involves linking names and addresses from an external database to the reconstructed records and checking that a record with that name, address, and the reconstructed demographic characteristics is present in the unprotected census data in the predicted census block. From this kind of linking, an attacker could infer confidential, sometimes sensitive information about individuals that was not already present in the external database, like race and ethnicity. However, some of these re-identifications are purely statistical, in the sense described above. The re-identification, and the inferences it enables, would have been possible by reasoning just from statistical aggregates, even if the person re-identified had never participated in the census.

When focusing on small-population blocks, a single person's participation has a much larger influence on whether they could be re-identified. In the extreme case where a person lives in a one-person block, if their data were not included in the census, the specific re-identification procedure used in the Census Bureau's simulation would never re-identify this person. In small blocks, the 38 percent rate of confirmed re-identifications jumps to 72 percent. This increase in confirmed re-identification rates in small blocks suggests millions of records exist for which the re-identifications in the simulated attack could not have been reasonably achieved purely from statistical information about their communities.

More concerning still is that the simulated attack discussed above was just a "lower bound"—a single, relatively simple, reconstruction-abetted, re-identification attack, with just a single set of external information in use. External attackers may have more resources, better external databases, and more clever algorithms. While the simulated reconstruction-abetted re-identification attack focused on inferences about race and ethnicity, future attacks could focus on other characteristics. It is difficult to predict what kinds of inferences might be harmful to the confidentiality of respondents in future censuses. The questions for the 2030 Census have not been determined, but the 2020 Census included data on children, same-sex relationships, household composition, older adults, and parents who are a different race or ethnicity than their children. Controlling the rate of inference an attacker may try to make about individuals is exactly the problem that the 2020 Census DAS was designed to address.

[13] Laura McKenna, "U.S. Census Bureau Reidentification Studies," U.S. Census Bureau, Washington, DC, 2019, <https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-Reidentification%20studies-20210331FinRed.pdf>, accessed August 11, 2021.

[14] Declaration of John Abowd, U.S. Census Bureau, State of Alabama v. U.S. Department of Commerce, Appendix B, United States District Court for the Middle District of Alabama Eastern Division, filed April 13, 2021.

Methods like data swapping that were used in the 2010 Census were designed to protect data for individuals who were considered most likely to be re-identified. But new computing technologies, by enabling large-scale, complete-record-level reconstructions, have drastically expanded the number of people who are vulnerable to re-identification. Older disclosure avoidance methods were not designed to defend against potential database reconstruction and re-identification attacks. If traditional disclosure avoidance techniques were applied to the 2020 Census data, the amount of noise required to protect against new attacks would make census data unfit for most uses. This vulnerability prompted the Census Bureau's Data Stewardship Executive Policy Committee (DSEP) to modernize disclosure avoidance for the 2020 Census.

Differential Privacy Enters the Scene

For the 2020 Census data, the Census Bureau applied a relatively newer disclosure avoidance framework based on "differential privacy." What is differential privacy and how does it differ from previous disclosure avoidance frameworks? The "goal of differential privacy is to obscure the presence or absence of any individual (in a database), or small groups of individuals, while at the same time preserving statistical utility."[15] The basic idea behind differential privacy is that the level of disclosure risk can be quantified, even when we cannot know what kinds of algorithms or external databases an attacker might deploy, which is important for transparency in setting disclosure review standards.[16]

Differential privacy works by adding "noise" to the collected data. Imagine the image on a television screen: what appears to be a clear, crisp picture is actually composed of millions of pixels, tiny dots of color. If you were to zoom in, you could identify individual pixels. Adding noise to the census data is like introducing small changes to the pixels. The noise reduces the risk that you can correctly identify any one individual but retains the overall picture when you zoom back out (Figure 1.2).

Adding noise into the data is a tradeoff. Adding more noise increases confidentiality protection, but it also makes the data less accurate. With differential privacy, we can now quantify that tradeoff (Figure 1.3).

Differential privacy is a framework in which the outcome of any data analysis—from a simple tabulation to a complex regression—is nearly equally likely, whether any individual is, or is not, included in the dataset. Because of this statistical property of the framework, differential privacy allows the Census Bureau to limit the disclosure risk for published data. If the output of an analysis is essentially the same, regardless of whether a given individual is in the dataset, then that person's confidential information is protected. There are numerous ways to implement differential privacy. This means that differential privacy is a characteristic of an algorithm or process, not a specific algorithm.

Differential privacy has some clear advantages over prior Census Bureau approaches to disclosure avoidance:

- Differential privacy allows the Census Bureau to track and address potential privacy loss as the list of published tables is expanded.
- Unlike prior methods of table suppression or record swapping, differentially private data can be published, analyzed, and linked to other data without any increased risk of disclosure; once the data have been processed, there is no more privacy loss regardless of how the data are used.
- Differential privacy provides mathematically provable guarantees against a wide range of potential privacy attacks.
- Differential privacy is transparent, unlike prior data protection methods such as data swapping. The programming code and decisions for differential privacy are available to the public; the only information not published is the exact value of the noise that is added to a given data point.[17]

[15] C. Dwork, "Differential Privacy: A Cryptographic Approach to Private Data Analysis," in Privacy, Big Data, and the Public Good, Cambridge University Press, New York, NY, 2014, pp. 296–322.

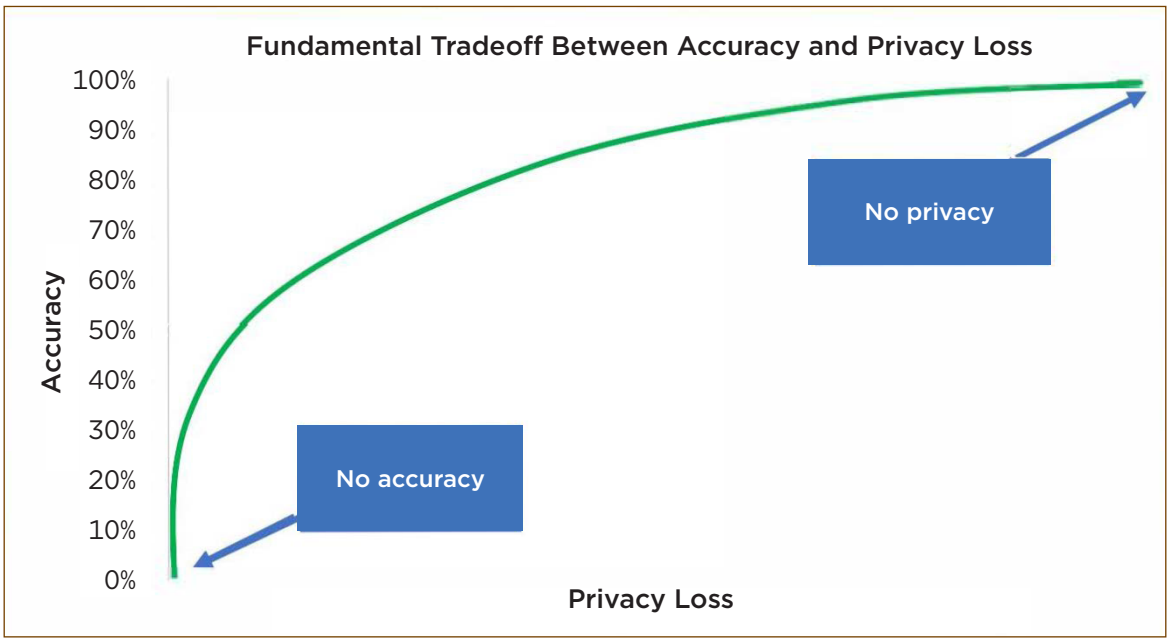
[16] C. Dwork, "Differential Privacy: A Survey of Results," Theory and Applications of Models of Computation, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2008, Vol. 4978, <https://doi.org/10.1007/978-3-540-79228-4_1>.

[17] The code base can be found at <https://github.com/uscensusbureau>.

Figure 1.2. **Adding Noise to Population Data Is Like Blurring Faces in a Photo**

Source: Population Reference Bureau.

Figure 1.3. **The Accuracy/Privacy Loss Tradeoff**



Source: U.S. Census Bureau.

Publishing the code base is an important step toward transparency because it allows data users to assess the impact of disclosure avoidance on the data, which was not possible with traditional disclosure avoidance methods like swapping. Documenting the impact of this noise infusion allows data users to assess whether the published data are suitable for their specific applications. We call this assuring the data's "fitness for use."

Differential privacy has been in use for Census Bureau products for more than a decade. In 2008, the Census Bureau published the world's first differentially private dataset through the Longitudinal Employer-Household Dynamics OnTheMap application—a revolutionary data system that links federal, state, and Census Bureau data on employers and employees.[18] Differential privacy is also used for other datasets, such as the Post-Secondary Employment Outcomes tabulations[19] and the Opportunity Atlas.[20] These data—which serve as an important resource for local planning, decision-making, and research—would not be available without modern disclosure avoidance methods such as differential privacy.

Differential privacy forms the foundation of the DAS used to protect the confidentiality of the 2020 Census data.

[18] More information on "OnTheMap" is available at <https://onthemap.ces.census.gov>.

[19] More information on "Post-Secondary Employment Outcomes (PSEO)" is available at <https://lehd.ces.census.gov/data/pseo_experimental.html>.

[20] More information on "The Opportunity Atlas" is available at <www.opportunityatlas.org/>.

2. HOW DOES THE DISCLOSURE AVOIDANCE SYSTEM WORK FOR REDISTRICTING DATA?

This handbook describes the Disclosure Avoidance System (DAS) in the context of the 2020 Census redistricting data because those are the first 2020 Census data that are protected using differential privacy. (The apportionment counts released earlier, in April 2021, were not subject to these new disclosure avoidance procedures and were instead the actual enumerated population counts for each state.)

As of the publication of this handbook (November 2021), the U.S. Census Bureau is still determining how to optimize the DAS for the next scheduled 2020 Census data products—the Demographic Profile and the Demographic and Housing Characteristics File. Information about confidentiality protection methods for these later data products will be published when more information is available.

Public Law 94-171, enacted by Congress in December 1975, requires the Census Bureau to provide states with census data they may use for legislative redistricting. The redistricting data files contain housing unit counts by occupancy status, total population, and population counts by race/ethnicity and voting age (aged 18 and over). For the first time, the 2020 redistricting data files also include data on the population living in seven major group quarters types, such as correctional facilities, college/university student housing, or military quarters.

The Census Bureau's DAS for redistricting data has two parts: differential privacy algorithms and post-processing. Both take place within a framework known as the TopDown Algorithm (TDA). The differentially private algorithms add noise to the data, while post-processing imposes certain consistencies (for example, ensuring that the population totals for counties within a state sum to the state's total population). Steps in the TDA process are described in more detail below.

How Noise Is Added to the Data

How does the Census Bureau apply differential privacy algorithms to the 2020 Census data? Working with input from stakeholders, the Census Bureau first compiled a list of tables for the 2020 Census redistricting data files.[21]

[21] A detailed list of tables is available in the Census Bureau's "2020 Census State Redistricting Data (Public Law 94-171) Summary File Technical Documentation," <https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/2020Census_PL94_171Redistricting_StatesTechDoc_English.pdf>.

Next, the Census Bureau consolidated all the redistricting data tables into one detailed cross-tabulation that reflects all the variables for each geographic level (from the nation, to states, down to census blocks), all categories for each variable in the dataset, and combinations of those categories (Table 2.1). For example, there are two categories for ethnicity—Hispanic or Latino and Not Hispanic or Latino.

In the published redistricting data files, there are 252 possible combinations of race, ethnicity, and age (63 x 2 x 2 = 252), plus eight residency types for people (housing unit plus seven group quarters types) and two occupancy status categories for housing units, which constitute 262 (252 + 8 + 2) distinct published data elements for each geographic unit.

To generate these published data, the TDA uses an even more detailed cross-tabulation that crosses the 252 race, ethnicity, and age categories with eight residential categories (lives in a housing unit and seven group quarters types) to get 2,016 (252 x 8) distinct data elements per geographic unit.

There are approximately 8 million census blocks in the 2020 Census—the smallest geography at which redistricting data are available. With 2,016 data elements per block, this means that there are more than 16 billion data cells for people in TDA. There are more than 12 million cells for housing units in that part of TDA.

Providing highly accurate information for every data cell would pose a disclosure risk; so, noise is added to protect the confidentiality of individual respondents. Adding noise to the data means that for any given data point, the TDA may add or subtract a small amount from the count to obscure the original value.

Table 2.1. **Number of Categories in the 2020 Census Redistricting File**

Variable	Number of categories
Race (6 race alone groups; 57 multiple race combinations)	63
Ethnicity (Hispanic or Latino; Not Hispanic or Latino)	2
Age (voting age, total population)	2
Occupancy status (occupied, vacant)	2
Population in group quarters (7 types)	7

Note: This table shows the number of categories for each variable, not the publication data layouts.
Source: Population Reference Bureau.

The level of noise introduced is guided by a "privacy-loss budget"—the budget defines the absolute upper bound of privacy loss that can occur. The privacy-loss budget can be set higher or lower, acting like a dial that tunes the amount of noise that is added to the data. As the privacy-loss budget rises, noise decreases (a greater share of the random noise numbers drawn are at or close to zero), meaning the data will be more accurate, but the likelihood that the reconstructed data can be used for re-identification also rises.

This privacy-loss budget can be set anywhere on a spectrum from "no accuracy but high protection" to "high accuracy but no protection." Choosing the privacy-loss budget is a policy decision based on a desired balance between accuracy and confidentiality, and the decision must be simultaneously informed by the Census Bureau's legal obligations and feedback on data utility from stakeholders. The lower the budget, the higher the protection and the less precise is each data point.

The total privacy-loss budget must be allocated across population characteristics, housing characteristics, and geographic levels. This process happens for selected topics referred to as "queries," rather than for the whole tabulation at once. More on this process is available in the "Multipass Optimization" section.

Technical Appendix A provides more information about the overall privacy-loss budget for redistricting data and how the budget is allocated across characteristics and geographic levels.

Privacy-Loss Budget Allocation

The overall privacy-loss budget must be distributed across all published census products (tables and microdata). Spending some of the budget to improve accuracy for one dimension of the data (such as more accurate total population counts for blocks) may mean that there is less budget for accuracy in another dimension (such as race detail). A detailed description of the privacy-loss budget for the 2020 Census and a listing of the budget for each table type and geographic hierarchy level is available in Technical Appendix A.

An illustrative example of noise infusion is shown in Table 2.2. In this example, noise is added to a tabulation of data by voting age and nonvoting age for the five census blocks in a hypothetical census block group. In the first step, noise is added independently to each of the individual tabulations.[22] In a second step, the noisy data are then controlled to the block group's tabulations and most inconsistencies are fixed. More information on the types of adjustments made in this second step is available in the "Additional Constraints" and "Example of Post-Processing" sections.

Within the TDA, the noise added to any given cell in a table is randomly drawn from a statistical distribution (described in more detail in Technical Appendix A).

The amount of noise added to any cell is independent of the size of the population in the cell. For example, it is equally likely that five people could be added to an area with a population of 100,000 or 100. This means that while the *absolute* error is the same for both areas, the noise added to small population cells will result in higher *relative* error because the underlying population (the denominator) is smaller. This higher relative error for small populations is an advantageous feature inherent to most disclosure avoidance methods including swapping, as re-identification risk is typically highest for data about small populations.

Notice in Table 2.2 that the amount of noise added to each cell is independent of the size of the cell—meaning a small cell may include a larger amount of noise or vice versa. Some cells may have zero noise added, meaning their values remain unchanged.

Noise is also added independently for each characteristic in each cell such as total population and population by voting age. The independence of the noise across cells in the same table may, however, lead to logically inconsistent data, such as the population aged 18 and over being larger than the total population in the hypothetical example for block 5.

[22] Noise is not added to state total populations nor to the national total but is introduced at lower geographic levels.

Table 2.2. **Hypothetical Example of Noise Infusion for a Census Block Group**

Step 1: Adding Noise to Tabulations

Block	Enumerated counts			Noise			Preliminary noisy table		
	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population
Block 1	25	75	100	0	−4	2	25	71	102
Block 2	20	70	90	−3	2	3	17	72	93
Block 3	10	40	50	2	−3	−2	12	37	48
Block 4	1	9	10	−2	1	1	−1	10	11
Block 5	1	2	3	0	2	0	1	4	3

Source: U.S. Census Bureau.

Noise may be positive or negative. For small cells, negative numbers make it possible that the noise-infused counts will be negative. Adding –2 to a population of 1 would result in a noise-infused value of –1 (as shown in the hypothetical example for the nonvoting-age population of block 4). Negative results are evidence of the uncertainty caused by the disclosure avoidance but are often confusing to data users, so a post-processing step is needed to adjust the noisy results and eliminate negative numbers.

Post-Processing the Noisy Statistics to Produce Tables

Invariants

The DAS departs from "textbook" differential privacy in one important way. The redistricting data include certain invariants—data that are kept exactly as enumerated with no noise added. Unlike traditional approaches to disclosure avoidance, differentially private noise infusion offers quantifiable and provable confidentiality guarantees. These guarantees, reflected in the global privacy-loss budget and its allocation to each statistic, serve as a promise to data subjects that there is an inviolable upper bound to the risk that an attacker can learn or infer something about those data subjects through publicly released data products. While that upper bound is ultimately a policy decision, and may be low or high depending on the balancing of the countervailing obligations to produce accurate data and to protect respondent confidentiality, the level of the global privacy-loss budget is central to the ability of the approach to protect the data. Invariants are, by their very nature, the equivalent of assigning infinite privacy-loss budget to particular statistics, which compromises the central promise of differentially private solutions to controlling disclosure risk. By excluding the accuracy of invariant data elements from the control of the privacy-loss budget, invariants exclude the disclosure risk and potential inferences that can be drawn from those data elements from the formal privacy guarantees. Thus, instead of being able to promise data subjects that the publication of data products will limit an attacker to being able to infer, at most, a certain amount about them (with that amount being determined by the size of the privacy-loss budget and its allocation to each characteristic), the inclusion of one or more invariants fundamentally excludes attacker inferences about the invariant characteristic(s) from the very nature of that promise. The qualifications and exclusions to the privacy guarantee weaken the strength of the approach and make communicating the resulting level of protection substantially more difficult. For these reasons, the Census Bureau chose to limit the number of invariants for the 2020 Census.

State population counts from the census are used to reapportion seats in the U.S. House of Representatives across the 50 states. The Census Bureau held the total population for each state invariant. Other statistics are held invariant for operational purposes, such as the total number of housing units in each census block and the number and type of group quarters facilities in each census block.

Invariant statistics for the 2020 Census redistricting data are:

- Total number of people in each state, the District of Columbia, and Puerto Rico.
- Total number of housing units (but not population counts) in each census block.
- Number of occupied group quarters facilities (but not population counts) in each census block by the following types:
  - Correctional facilities for adults.
  - Juvenile facilities.
  - Nursing facilities/skilled-nursing facilities.
  - Other institutional facilities.
  - College/university student housing.
  - Military quarters.
  - Other noninstitutional facilities.

All other population and housing characteristics, including population counts for every geography below the state level, have had noise introduced.

Additional Constraints

In addition to the invariants noted above, there are some constraints within TDA that are applied at all geographic levels. These constraints include the following:

- Population and housing counts must be integers and may not be negative.
- The cells of a table must sum to its row and column margins, which must in turn sum to the total population for the table.
- Counts must be consistent within tables, across tables, and across geographies for a given universe (i.e., population tables are consistent with population tables, and housing tables are consistent with housing tables). For example, the population by race must sum to the total population, the number of occupied and vacant housing units must sum to the total number of housing units, and the population in each county within a state must sum to the state's total population.
- If there are zero housing units and zero group quarters (GQ) facilities in a geography, then no people may be assigned to that geography.
- The number of people per GQ facility is greater than or equal to 1.

- The number of people per housing unit is less than or equal to 99,999, and the number of people per GQ facility is less than or equal to 99,999.
- There are zero people aged less than 18 in GQ type 301, "Nursing facilities/skilled nursing facilities."

While these constraints have been applied in TDA, some inconsistencies may remain in the redistricting data files. These inconsistencies are described in detail in the section "Improbable and Impossible Results."

How Does the TopDown Algorithm (TDA) Work?

1. After the confidential Census Edited File (CEF)[23] is input into the DAS, the system's TDA takes an extensive series of differentially private "noisy" measurements.
2. The algorithm uses these measurements to generate privacy-protected microdata records for the entire nation.
3. These individual records contain every level of geography on the Census Bureau's geographic hierarchy based on the noisy measurements taken at each of those geographic levels and subject to the population invariants and other constraints.
4. These microdata records are exported into the tabulation system to generate the redistricting data products.
5. The resulting data reflect the privacy guarantees established by the privacy-loss budget for the 2020 Census, incorporating the greatest level of uncertainty at the census block level (where privacy risk is usually greatest), while providing increasingly accurate measures of the nation's population at each higher level of geography.

[23] The 2020 CEF—the individual census responses that have been processed through quality control routines such as filling in missing information.

Moving From the Top to the Bottom of the Geographic Hierarchy

The Census Bureau also considers geographic nesting, such as counties within states, as it applies noise at different geographic levels.

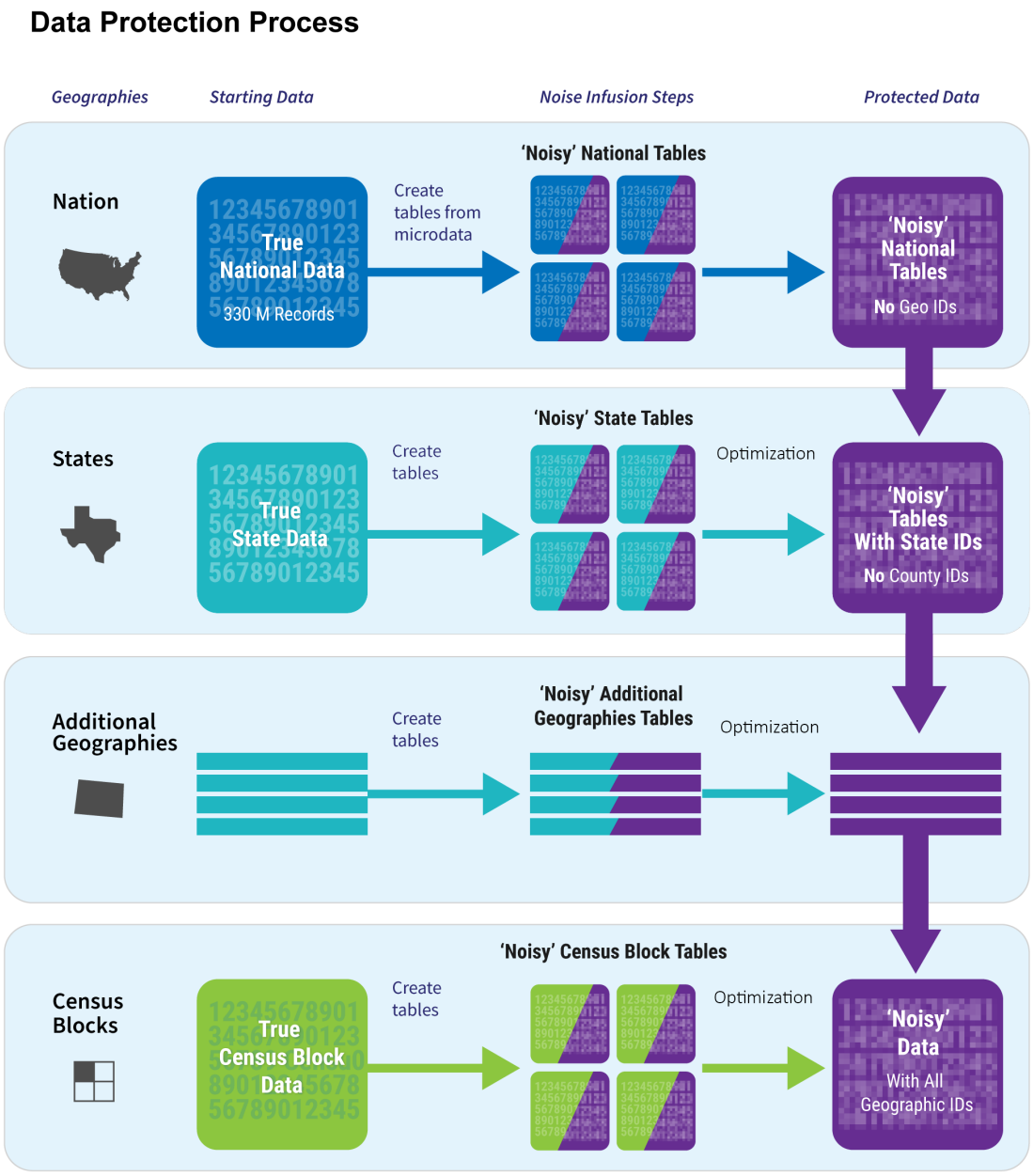
Starting with the list of redistricting tabulations described above, the Census Bureau queries the 2020 CEF to produce certain tabulations, such as counts of the voting-age population, for every geographic area in the country. The TDA adds noise to cells in those tabulations using a differential privacy mechanism. Then starting at the national level, the noise-infused tabulations are used to adjust a detailed cross-tabulation—representing all of the combinations of characteristics across all of the data—to create a new nationwide, noise-infused set of data. These data include a "noisy" record representing every person in the United States but do not yet include geographic information. (Figure 2.1)

Once the national data are set, the process is repeated for states. In the state step, mathematical optimization routines ensure that the state totals for different population or housing characteristics are as close as possible to the noisy measurements and that these state totals, when added together, are consistent with the national data from the prior step. The result is an updated set of data that now includes state identifiers.

This optimization process is repeated for a series of ever-smaller geographic units, ending with census blocks. The geographic hierarchy is described in more detail in the section on "Geographies and the Geographic 'Spine.'"

In the very last step, the tabular census block data are converted back into microdata.

Figure 2.1. **Creating Differentially Private Data for the 2020 Census Redistricting Files**



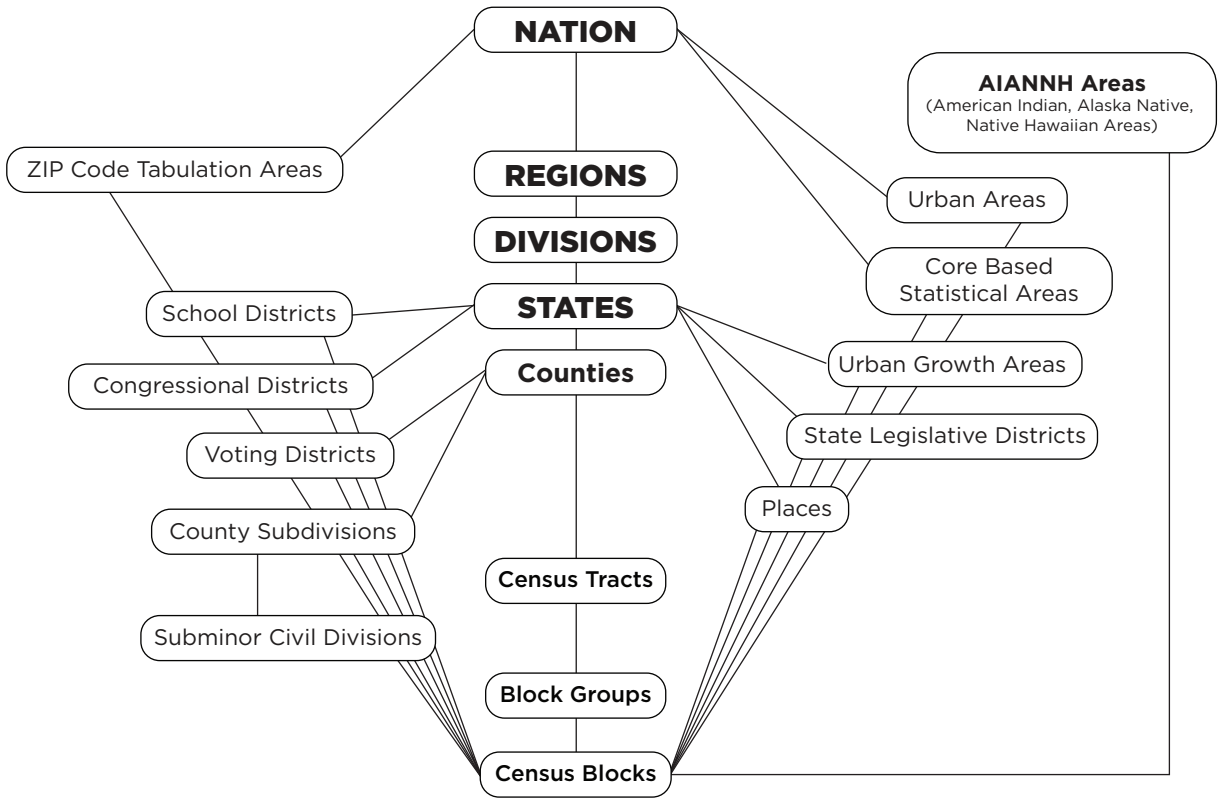
Source: Population Reference Bureau.

Geographies and the Geographic "Spine"

The hierarchy, or nesting scheme, of geographies for census data products is sometimes called the geographic "spine" (Figure 2.2). Along the spine, each "child" geography perfectly nests within its "parent" geography. For example, all counties nest within one (and only one) state. Starting with the smallest unit along the geographic spine and working upward: blocks nest within block groups, block groups within census tracts, census tracts within counties, counties within states, states within divisions, divisions within regions, and regions within the nation.

Some geographies, however, do not fit within the nesting scheme. School districts, for example, can be summed up from blocks and fit within states but do not necessarily follow block group or tract boundaries. Because these nonnested, or "off-spine," geographies are not part of the TDA processing routine, the noise-infused data for these areas may be noisier than those for the on-spine geographies. To address feedback from data users about the importance of accurate data for off-spine geographies, the Census Bureau made changes to the geographic hierarchy used for TDA.

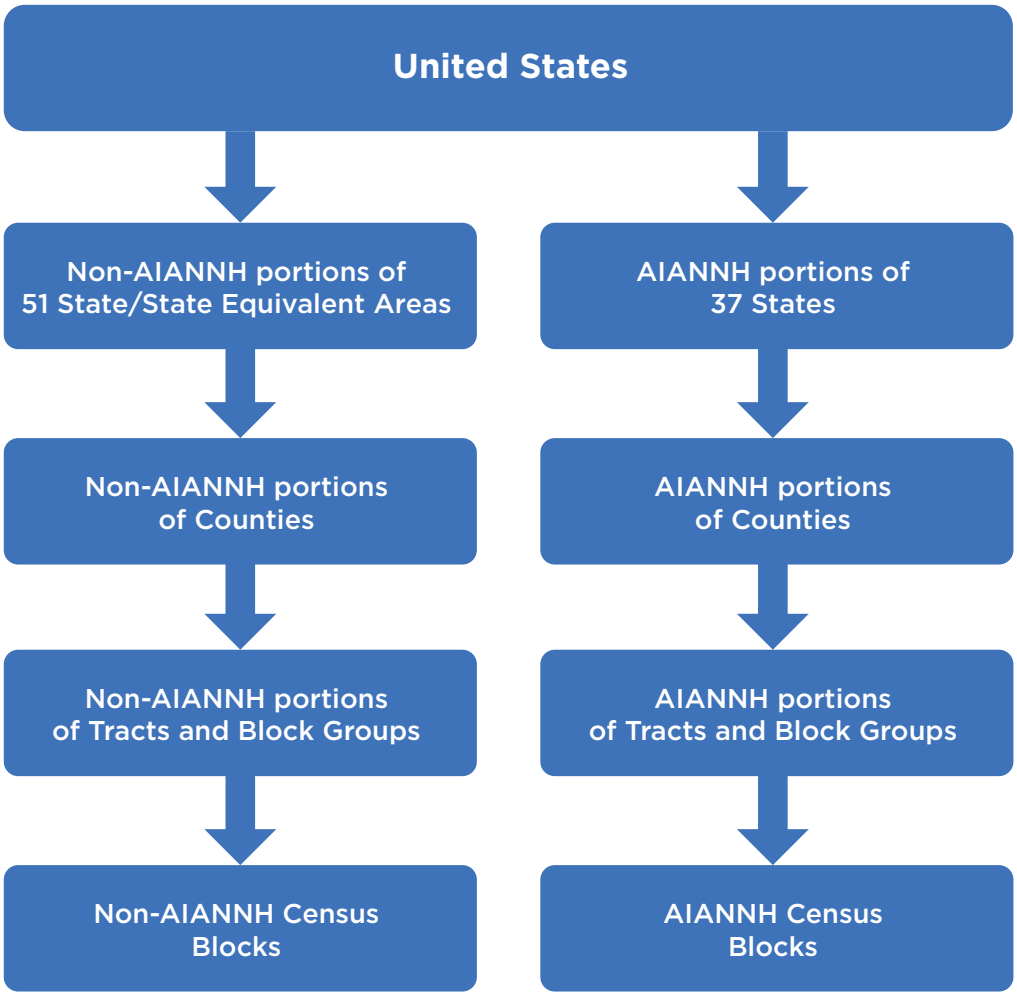
Figure 2.2. **Standard Hierarchy of Select Geographic Areas**



Source: U.S. Census Bureau.

The hierarchy used for TDA differs from the standard hierarchy of census geography in important ways. First, for states with American Indian/Alaska Native/ Native Hawaiian (AIANNH) areas, the AIANNH and non-AIANNH portions of the state are split to improve data accuracy for AIANNH areas (Figure 2.3).

Figure 2.3. **Hierarchy for Disclosure Avoidance System Processing**



Source: U.S. Census Bureau.

Within TDA, all AIANNH areas in a state are grouped together for data processing. This minimizes the likelihood that post-processing could result in systematic undercounts. For example, at the state level, three American Indian areas in Kansas—the (IA-KS-NE) Reservation and Off-Reservation Trust, Kickapoo (KS) Reservation, and Prairie Band of Potawatomi Nation Reservation—are processed together, separate from the rest of Kansas (Figure 2.4). At lower geographic levels, these individual tribal areas are then processed separately from each other.

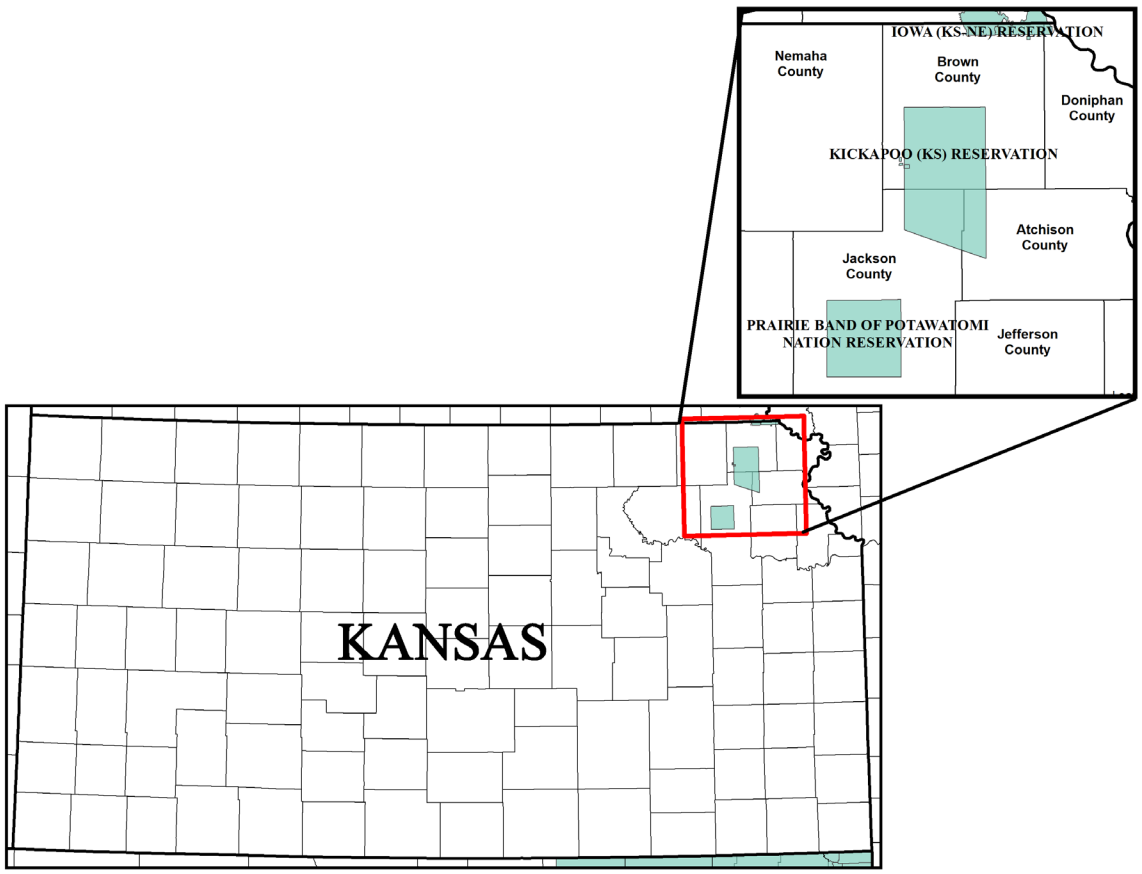
Another important departure from the standard geographic hierarchy is in how blocks are grouped before being aggregated to tracts. Rather than using the Census Bureau's standard block groups, blocks are aggregated—sometimes in groups of nonbordering blocks—to improve the TDA's processing efficiency and reduce post-processing error especially for GQ residents.

In most states, the District of Columbia, and Puerto Rico, these block aggregations (called "optimized block groups" in the technical documentation) were redefined to more closely approximate places (such as cities). In 12 states, blocks were aggregated to more closely approximate minor civil divisions (cities, boroughs, and towns/townships).

While some of the TDA geographic groupings differ from those in the standard geographic hierarchy, data products will still be released for the standard tabulation geographic entities. TDA geographies are intended for data processing, not for reporting.

In TDA, the Census Bureau processes all of the geographic units within a larger geographic area at the same time to ensure that they add up to the parent geography. For example, the Census Bureau examines the noisy population counts for all tracts within a county, and then finds the set of counts for each tract

Figure 2.4. **Example of Grouping American Indian, Alaska Native, and Native Hawaiian Areas for TopDown Algorithm in Kansas**



Source: U.S. Census Bureau.

that is closest to its noisy count but that also adds up to the total population for the county.

Example of Post-Processing

After the DAS produces noise-infused counts, the data undergo further post-processing. More information about constraints integrated into the post-processing step is available in the "Additional Constraints" section.

Table 2.3 builds on Table 2.2, adding the post-processing step to the example of noise-infused data. The noise introduced into each table cell results in population totals that are different from the original data. The processing steps handles several issues from the noisy data step. First, negative population counts, such as the –1 value for the Block 4 population aged 18 and over are adjusted to be nonnegative. Some inconsistencies, such as population aged 18 and over being larger than total population (as occurs for the Block 5 population), are also resolved. Then, the noisy characteristics are adjusted to match the total noisy population across all relevant geographies. In this example, the preliminary noisy block population totals summed to 257, but must be adjusted to sum to 254, the privacy-protected block group total.

Multipass Optimization

While the noisy measurements themselves do not introduce bias into the results because noise is drawn from a symmetrical distribution centered on zero (with an equal distribution of positive and negative noise values), the post-processing step may introduce bias, by e.g., removing negative values or to impose other constraints on the resulting data. More information about this can be viewed in "Additional Constraints."

The Census Bureau implemented a new post-processing routine, called "multipass optimization," to reduce bias. Multipass optimization is described in more detail in the next section, but the routine is intended to reduce bias for small geographic areas and population subgroups.

In prior iterations of the TDA, the Census Bureau observed that small populations tended to have a positive bias, where the published count was higher than in the original, confidential data; larger populations tended to have a corresponding negative bias. For example, there was a slight bias for total population toward rural areas. The Census Bureau reconfigured the TDA parameters to largely eliminate this impact.[24]

Detailed Summary Metrics published with each of the model runs provide specific information about bias at varying levels of geography.[25]

A key feature of the final version of the TDA used to produce the redistricting data is that the accuracy and reliability of statistics should increase as the underlying population being measured increases. To address this objective, the Census Bureau implemented a multipass framework that processes certain elements of the data first and then uses those results as input to subsequent steps.

At the national level, the state level, and then for lower levels of geography, multipass first determines the population count for each unit within that geographic level (for example, the population for each county

[24] The Detailed Summary Metrics (2021-06-08) can be found at <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20210608/2021-06-08-data-metrics-tables_production-settings.xlsx>.

[25] More information on "Developing the DAS: Demonstration Data and Progress Metrics" is available at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

Table 2.3. **Hypothetical Example of Post-Processing**

Step 2: Post-processing

Block	Enumerated counts			Noise			Preliminary noisy counts			Post-processed counts		
	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population
Block 1	25	75	100	0	–4	2	25	71	102	27 (+2)	71 (–4)	98 (–2)
Block 2	20	70	90	–3	2	3	17	72	93	19 (–1)	72 (+2)	91 (+1)
Block 3	10	40	50	2	–3	–2	12	37	48	12 (+2)	37 (–3)	49 (–1)
Block 4	1	9	10	–2	1	1	–1	10	11	0 (–1)	11 (+2)	11 (+1)
Block 5	1	2	3	0	2	0	1	4	3	1 (+0)	4 (+2)	5 (+2)
									Block group	**59**	**195**	**254**

Source: U.S. Census Bureau.

within a state or each census tract within a county). Next, the algorithm generates the remaining statistics, constraining those statistics to the population counts determined in the first pass.[26]

Improbable and Impossible Results

It is possible that noise infusion could result in some improbable results in the redistricting data. For example:

- A block might have only one occupied housing unit but dozens of people (implying that those dozens of people live in the same household).
- A block may have resident children under the age of 18, but no adults present.

The data could also include mathematically impossible statistics. For example:

- A block may have people living in households in an area with only vacant housing units.
- A block may have more occupied housing units than people to occupy those units.

These inconsistent and improbable results are often associated with geographic units having very small populations. For example, as shown in Table 2.4, 4.8 percent of blocks with people living in households have zero occupied housing units. But only about 0.1 percent of block groups and tracts have this kind of inconsistency.

Data users will find that the frequency of improbable and impossible results diminishes, and the accuracy of the estimates increases, as data are aggregated to larger geographic areas. For many use cases, such as detailed housing or household population analysis, block-level data may be too noisy. Block groups, census tracts, or other larger geographies may be better choices as units of analysis. Data users are encouraged to combine block-level data into geographic areas with larger populations. Doing so reduces the noise due to disclosure avoidance.

The next section provides more guidance on how users can deal with impossible and improbable results.

[26] John M. Abowd and Victoria A. Velkoff, "Modernizing Disclosure Avoidance: A Multipass Solution to Post-Processing Error," U.S. Census Bureau, Washington, DC, 2020, <www.census.gov/newsroom/blogs/research-matters/2020/06/modernizing_disclosu.html>.

Table 2.4. **Inconsistent or Implausible Results by Geographic Summary Level**

Inconsistency	Blocks affected		Block groups affected		Tracts affected		Counties affected	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Zero occupied housing units but more than zero household population	392,921	4.80	223	0.09	90	0.11	0	0.00
Zero household population but more than zero occupied housing units	91,415	1.10	30	0.01	17	0.02	0	0.00
Everyone in area under age 18 (excludes areas with group quarters population)[1]	101,127	1.80	27	0.02	17	0.05	0	0.00

[1] Share of areas that have no group quarters population.
Source: U.S. Census Bureau.

3. RECOMMENDATIONS AND CONSIDERATIONS WHEN USING THE REDISTRICTING DATA

What do data users need to know before they start using statistics from the 2020 Census redistricting data files? This section provides some considerations and recommendations for working with the data.

Block-level data should be aggregated before use. The amount of noise added to statistics does not depend on population or geographic size, so block-level data are most affected by disclosure avoidance procedures. For example, it is equally likely that five people could be added to an area with a population of 10,000 or a population of 100. As data are aggregated across blocks or across demographic groups, the accuracy of the resulting data will increase.

U.S. Census Bureau researchers found that for block groups, a minimum total population between 450 and 499 is sufficient to provide reliable characteristics of various demographic groups, whereas a minimum total population between 200 and 249 provides reliable characteristics for places and minor civil divisions.[27]

Counts are consistent within tables, across tables, and across geographies. For example, rows within a table sum up to the parent row and universe. The total population count in Table P1 is consistent with the total population count in Table P2. In addition, block-level tables sum to their corresponding block-group-level tables, block-group-level tables sum up to their tract-level tables, and so forth.

Data should not be divided across tables in low population areas. For example, values from Table P2 should not be divided by values from Table H1 at low levels of geography or for low population areas to obtain the average number of people per household. The separation of the people universe from the housing universe introduces some inconsistencies, particularly at low levels of geography (tract and smaller) such as more households than people. More on this topic is available in the "Improbable and Impossible Results" section. Users who want more accurate statistics on people per household should wait for the release of the Detailed Demographic and Housing Characteristics (Detailed DHC) File.

Data may be subtracted across tables to obtain new counts. For example, you can subtract Table P3 from Table P1 or Table P4 from Table P2 to obtain the population under 18 years old. However, subtracting data across tables at the block level may yield improbable results such as a large number of children under 18 years old relative to the number of adults. Aggregating to larger geographies reduces the likelihood of these improbable results.

The Disclosure Avoidance System is not the only source of uncertainty in 2020 Census data. Noise introduced by disclosure avoidance may compound underlying errors or may offset those errors. (Examples of these types of errors are available in the 2010 Census Post-Enumeration Survey.)[28]

[27] Tommy Wright and Kyle Irimata, "Empirical Study of Two Aspects of The TopDown Algorithm Output for Redistricting: Reliability & Variability (August 5, 2021 Update)," Working paper #2021-02, U.S. Census Bureau, Washington, DC, 2021, <www.census.gov/library/working-papers/2021/adrm/SSS2021-02.html>.

[28] More information is available at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/data-quality.html#metrics>.

This page is intentionally blank.

4. EVALUATING THE 2020 CENSUS DATA

The formal privacy methods of the 2020 Disclosure Avoidance System (DAS) will allow data users, for the first time, to understand the extent to which a statistic or data cell may have been altered and whether it is suitable for their inferences. While the actual noise in an individual data cell will not be published, the amount of expected noise can be inferred from published model parameters, the privacy-loss budget, and summary "fitness-for-use" metrics.

There have been numerous assessments of the impact of the DAS on 2010 Census data, including with using the 2020 production parameters on that 2010 data. The production parameters and privacy-loss budget allocations used for the 2020 Census redistricting data are included in Technical Appendix A.

In October 2019, the U.S. Census Bureau released an initial set of demonstration data products using 2010 Census data that had been run through an interim version of the DAS. The purpose was to demonstrate that the noise-infused data were fit for use.[29] Although the DAS did very well at ensuring the data's fitness-for-use for some important use cases, it fell short in others.

The Census Bureau released three additional demonstration data products using the same privacy-loss budget as the initial set of demonstration data products. The privacy-loss budget was held roughly the same across those four releases to allow analysts and data users to compare the effects of incremental algorithmic improvements in the system. The fifth demonstration dataset included two versions: a version using an increased privacy-loss budget and a version using the earlier, development-focused privacy-loss budget. The version using higher allocation of privacy-loss budget allowed data users to evaluate demonstration data that more readily approximated the anticipated confidentiality/accuracy tradeoff of the 2020 Census data products.

Through this process, the Census Bureau received invaluable feedback from external stakeholders through the 2020 DAS e-mail, advisory meetings, tribal consultations, and comments provided during presentations at conferences and the Differential Privacy Webinar Series that informed our efforts and decision-making. The Census Bureau and external data users identified several issues with the DAS that needed to be resolved before it could be applied to the 2020 Census data, including:

- Situations where small populations tended to gain population, whereas larger populations tended to lose population.
- Limitations of the noise-infused data for emergency planning operations.
- Issues for populations living on American Indian reservations.
- Problems with the accuracy of census data for "off-spine" geographies.[30]
- Identification of extreme outliers.
- Distortions in the data that effectively moved individuals from high- to low-density populations (e.g., from cities to rural areas or from larger race groups to smaller race groups).

The Census Bureau used these assessments to make improvements to the DAS and to make targeted increases and reallocations of the privacy-loss budget in order to improve overall accuracy for geographic areas and other characteristics, but never to favor a particular subpopulation over another. As a result of this work, the Census Bureau was able to greatly reduce or eliminate all of these limitations. Details of all demonstration datasets, including "fitness-for-use" metrics for each model run, can be found on the Census Bureau's Web site.[31] Internally, the Census Bureau also conducted over 600 experimental data runs to optimize and tune the parameters of the DAS algorithm. These internal assessments of the DAS were informed by various applications such as enforcement of the Voting Rights Act (Box 4-1), the creation of population estimates and projections, and demographic reasonableness analysis.

[29] John M. Abowd and Victoria A. Velkoff, "Modernizing Disclosure Avoidance: A Multipass Solution to Post-Processing Error," U.S. Census Bureau, Washington, DC, 2020, <www.census.gov/newsroom/blogs/research-matters/2020/06/modernizing_disclosu.html>.

[30] Committee on National Statistics, workshop on "2020 Census Data Products: Data Needs and Privacy Considerations," <https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518>.

[31] More information on "Developing the DAS: Demonstration Data and Progress Metrics" is available at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

Box 4-1. Data for the Voting Rights Act

The published data from the 2020 Census are available for jurisdictions to use in devising redistricting plans for offices from the U.S. House of Representatives to local school boards and for the analysis of such plans by the U.S. Department of Justice (DOJ) for compliance with federal voting rights laws, including the Voting Rights Act of 1965, Title 52 U.S. Code, Section 10301. To assess the effect of the Disclosure Avoidance System on redistricting data, U.S. Census Bureau researchers measured the effects of applying the production version of the TopDown Algorithm (TDA) to the 2010 Census data by analyzing the results using previous redistricting plans provided by the DOJ.

Their starting point for this analysis was the published 2010 Census Redistricting Data (Public Law 94-171) Summary File that resulted from applying data swapping to the 2010 Census Edited File (CEF). The comparison to published data, rather than the CEF, allows for external data users to replicate or extend the analysis. In addition, block-level counts of total population and population aged 18 and over were the same in the 2010 CEF and the published data. (Note that while this analysis relied on comparisons to published, swapped data, our internal team did conduct additional analyses that compared the differentially private data to the unswapped CEF with similar results.)

Next, the researchers used data where the TDA (production version) had been applied to the 2010 CEF 25 different times. The TDA adds noise randomly and there was interest in how results would vary among the 25 runs. The privacy-loss budget for each run of the TDA was ε=17.41 (rho=2.56, delta=10-10) for the person file. An explanation of the privacy-loss budget is available in Technical Appendix A.

Thus, the researchers had 26 different national datasets—one where the 2010 CEF had been treated with data swapping and 25 where the 2010 CEF had been treated 25 different times with the production version of the TDA. Their approach had two parts: (1) to report observations on variability of results among the 25 runs of the TDA relative to the average of the 25 runs, and (2) to report observations on variability between the results among the 25 runs of the TDA relative to the data swapping (i.e., the published 2010 Census Redistricting Data [P.L. 94-171] Summary File data).

In the first part of their analysis, the researchers sought to determine the minimum population size necessary for geographic areas to have reliable demographic characteristics for the purposes of redistricting. Examining census block groups as well as places and minor civil divisions (MCDs), they demonstrated that for any block group with a total population between 450 and 499 people or larger, and for MCDs and places between 200 and 249 or larger, the difference in the largest demographic group as a proportion of the total population between the published 2010 Census tabulations and the 2010 Demonstration Privacy-Protected Microdata File (2021-06-08) is less than or equal to 5 percentage points at least 95 percent of the time. No congressional or state legislative district fails this test; that is, for these districts, the 5-percentage-point criterion holds 100 percent of the time.

The second part of their analysis examined districts in Rhode Island and in three specific jurisdictions provided by the DOJ. The three cases are Panola County, MS (2,180 blocks); Tate County (School District), MS (784 blocks); and Tylertown (Walthall County), MS (136 blocks). Additional jurisdictions of various sizes were also included in internal reviews but were not the subject of this particular analysis. Overall, the researchers observed empirically "that variability in data results from the TDA increases as we consider smaller pieces of geography and population" but the relative accuracy of the data increases substantially as the noisy block-level data are aggregated together into their jurisdictions. Specifically, for the Rhode Island districts analyses, they observed "that counts and percentages put in place from swapping being applied to the 2010 CEF have very similar counts and percentages after the TDA is applied to the same 2010 CEF." Moreover, variability with the 2021-04-28 version of the TDA (privacy-loss budget ε=10.3) is less than what they reported with the 2019-10-31 version (privacy-loss budget ε=4.0).

Overall, the comparisons showed that differences from the 2010 Census Public Law 94-171 data decreased as geographic and population size increased.

Census Bureau researchers also examined the impact of the TDA production settings on the ability to identify majority-minority districts (districts in which a demographic group constitutes a majority of the total population or of the voting-age population). This research examined the proportion of 26 race and Hispanic origin demographic categories in each of the nation's 436 congressional districts (including the District of Columbia's nonvoting delegate district), 1,946 state upper legislative districts, and 4,785 state lower

legislative districts, comparing the published 2010 Census tabulations to the 2010 Demonstration Data Privacy-Protected Microdata File (2021-06-08) with the production settings. Comparing these data, researchers identified 25 districts out of 7,167 (0.3 percent of all districts) where a demographic group could be considered to flip from having a majority in the published 2010 Census tabulations to being a minority in the demonstration data or vice versa. In every case, slight changes to the district boundaries could restore the original determination and the boundaries represent what was drawn with the original data, not what would have been drawn had the differentially private data been the basis. Flips occurred in both directions (11 groups went from majority to minority, 14 went from minority to majority). No flips involved both a racial or ethnic group's total population and their voting-age population; that is, districts drawn such that a demographic group constitutes a majority relative to both the total population and to the voting-age population are more stable. All observed flips involved very small numbers of individuals in districts that were tightly drawn (usually within a few hundredths of a percent of the 50 percent mark) using the published 2010 Census tabulations (a level of precision that would be greatly impacted by the noise injected into racial and Hispanic origin characteristics by the 2010 Census swapping algorithms). Detailed results from these analyses are available in two working papers available on the Census Bureau's Web site and on a recorded Webinar.[1]

[1] Tommy Wright and Kyle Irimata, "Empirical Study of Two Aspects of the TopDown Algorithm Output for Redistricting: Reliability & Variability," Working paper #2021-01, U.S. Census Bureau, Washington, DC, 2021, <www.census.gov/library/working-papers/2021/adrm/SSS2021-01.html>; Tommy Wright and Kyle Irimata, "Empirical Study of Two Aspects of the TopDown Algorithm Output for Redistricting: Reliability & Variability (August 5, 2021 Update)," Working paper #2021-02, U.S. Census Bureau, Washington, DC, 2021, <www.census.gov/library/working-papers/2021/adrm/SSS2021-02.html>; U.S. Census Bureau Webinar, "Understanding the 2020 Census Disclosure Avoidance System: Analysis of Production Settings for Redistricting and Voting Rights Act Use Cases," Recorded August 10, 2021, available at <www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/analysis-of-demonstration-data-for-redistricting-and-voting-rights-act-use-cases-production-settings.html>.

This page is intentionally blank.

5. FREQUENTLY ASKED QUESTIONS

This section provides answers to some frequently asked questions about disclosure avoidance. The U.S. Census Bureau also provides a wealth of information about disclosure avoidance on its "Frequently Asked Questions" and "2020 Census Data Products: Disclosure Avoidance Modernization" Web pages.[32]

How is a differentially private system different from the Census Bureau's prior disclosure avoidance techniques?

The disclosure avoidance techniques that were used in the 2010 Census and in the American Community Survey rely on "swapping" characteristics in the underlying data between a subset (millions) of households in different geographic areas. In this era of Big Data, these methods are insufficient. Were we to use our prior disclosure avoidance techniques, the amount of noise we would have to inject into the data to comply with our statutory confidentiality obligations would make census data unfit for most uses.

With the current method, the noise is added to the statistics in the tables themselves. This allows the U.S. Census Bureau to precisely control the amount of noise that we add. By documenting the properties of this noise, we can help data users determine if published estimates are suitable for their specific applications. We call this assuring "fitness for use." Documenting the impact of this noise is similar to the way we provide margins of error for our current statistical products. For the same level of protection, a differentially private 2020 dataset will be significantly more accurate than datasets produced using our prior disclosure avoidance methods.

Do 2020 Census state population totals reflect actual reported totals, exempted from disclosure avoidance methods?

Yes. As always, state population totals from the 2020 Census will reflect the actual population numbers as enumerated in the census. The totals determine congressional apportionment and are protected only by aggregation. We call such statistics "invariants," meaning that their value will not be modified by the Disclosure Avoidance System. We use invariants sparingly in our disclosure avoidance algorithms as they impact the calibration of noise that must be applied to other statistics and weaken the overall confidentiality guarantee.

How did the Census Bureau involve data users in the design of the Disclosure Avoidance System (DAS)?

The U.S. Census Bureau's Data Stewardship Executive Policy Committee (DSEP) relies on input from a variety of sources when making decisions about the adoption, implementation, and parameters of the DAS. These include internal subject matter experts, the Census Bureau's advisory panels (the National Advisory Committee on Racial, Ethnic, and Other Populations and the Census Scientific Advisory Committee), the Committee on National Statistics of the National Academy of Sciences, academic experts and researchers, privacy advocates, professional associations, federal and state partners (including the DOJ with regards to Voting Rights Act matters), and many others. We also solicited public comments in a July 2018 Federal Register notice and have conducted formal consultations with American Indian and Alaska Native tribal leaders.

Engagement with these and other stakeholders is ongoing. The Census Bureau will continue to solicit and consider feedback to improve our disclosure avoidance methods. This process of enhanced data user engagement in the design and implementation of disclosure avoidance methods marks a significant shift from prior censuses, where data users were largely unaware of the impact of the methods being applied.

How will the Disclosure Avoidance System work for other 2020 Census products?

TopDown Algorithm as designed can provide consistency between redistricting data and other 2020 Census data products, such as the Demographic Profile and Demographic and Housing Characteristic File (DHC). However, methods for disclosure avoidance in the DHC files were not finalized at the time of publication.

Future data products will include additional data on household and relationship-to-householder characteristics, age detail, and other demographic and housing information.

The U.S. Census Bureau will continue to seek input from stakeholders as they make decisions about disclosure avoidance procedures for these products.

[32] More information on "Frequently Asked Questions" is available at <https://ask.census.gov>. More information on "2020 Census Data Products: Disclosure Avoidance Modernization" is available at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html>.

What harms could arise if the basic demographic data collected in the decennial census is exposed?

Data stewardship is a comprehensive framework designed to protect information over the course of the information life cycle, from collection to dissemination, and it starts with a commitment to confidentiality that is required by law and designed to maintain public trust. Research conducted by both the U.S. Census Bureau and nongovernmental researchers has shown that concerns about privacy and confidentiality are among the reasons most often given by potential respondents for unwillingness to participate in surveys and censuses.[33, 34]

In addition to the impact of confidentiality protections on response rates, our disclosure avoidance system protects against direct threats to the disclosure of our respondents' data. Many vendors collect, sell, and publish data about people living in the United States. While many commercial vendors have access to their own data on name, address, and date of birth, fewer vendors have access to the type of rich demographic data the census collects on characteristics like race, ethnicity, and household relationships.

The information on demographic characteristics these vendors lack is precisely the sort of information collected by the decennial census. The disclosure of these types of characteristics could not only make it easier to target individuals—particularly in vulnerable populations such as communities of color, same-sex couples, older adults, or parents of very young children—for fraud, enforcement actions, disinformation, or physical or virtual abuse, but it could also undermine the public's trust in the confidentiality of its census response, which could cause people to be less likely to respond to future censuses.

Could external attackers know whether they've correctly re-identified individuals in census data even if the attackers don't have access to confidential census records?

Yes, if they have access to additional outside data sources or perform some minimal fieldwork to verify their results.[35] Vulnerability of the published data to reconstruction of the confidential microdata could be an unintentional violation of existing disclosure avoidance rules for published microdata that were in place for the 2010 Census. Re-identification of those records is not required to trigger strengthening the necessary disclosure avoidance standards for tabular data releases.

This is one reason why the U.S. Census Bureau must seriously address the threat of disclosure and apply a comprehensive and coordinated program of disclosure avoidance.

The Census Bureau has the only copy of the confidential microdata, but an adversary could have access to many different outside data sources. Unless we protect the data, an adversary could independently confirm their re-identifications with reasonable certainty.

As the volume and quality of outside data sources—such as names, addresses, and birth dates—grow and improve, so do adversaries presumed and actual matches. Our analysis of 2010 Census re-identification vulnerability used a large database of commercial information available at the time of that census. The risks associated with using 2010 disclosure avoidance methods today and into the future will only increase.

Is there any evidence of successful re-identifications by attackers?

To date, we are not aware of successful re-identifications by bad actors, though we would not necessarily expect bad actors to publicize their results. We have, however, documented re-identifications that users have brought to our attention through Reidentification Studies.[36] There has been a dramatic increase in the availability of both large-scale computing resources and commercial-strength optimizers that can solve systems of billions of simultaneous equations.

Together, these resources and tools have changed the threat of database reconstruction from a theoretical risk to an issue that the U.S. Census Bureau must address. The adoption of differential privacy for 2020 Census data releases is intended to guard against successful reconstructions and re-identifications by those who seek to reverse-engineer the census data. This includes those who would be especially difficult to identify like state actors (e.g., foreign governments), corporations, and cybercriminals, all of whom would be unlikely to publicly announce a successful reconstruction or re-identification attack.

[33] More information on research conducted by the Census Bureau is available at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/final-analysis/2020-report-cbams-study-survey.html>.

[34] More information on nongovernmental researchers is available at <www.srl.uic.edu/newsletter/issues/2000s/04v35n2-3.pdf>.

[35] Simson L. Garfinkel, "De-Identification of Personal Information," NISTIR 8053, National Institute of Standards and Technology, Washington, DC, 2015, <https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>.

[36] More information on Reidentification Studies is available at <www.census.gov/library/working-papers/2019/adrm/2019-04-ReidentificationStudies.html>.

Why is the Census Bureau adopting modernized disclosure avoidance for the 2020 Census instead of waiting until the 2030 Census?

Our research verified that traditional disclosure avoidance methods leave personal data exposed with today's faster computers, high-powered machine learning software, and large public databases. This left us with two choices: we could publish significantly less information, or we could adopt a modernized approach to confidentiality protection. We chose the latter, and there is no other statistical technique that can be reliably employed to assure the confidentiality of the underlying data while simultaneously assuring the highest quality statistical product for our data users.

The U.S. Census Bureau has a dual mandate to produce quality statistical information and protect the confidentiality of respondent data. We know that the nation needs timely and accurate information to make informed decisions. People have to know that we will safeguard their privacy and the confidentiality of their data zealously if we want them to entrust us with their personal information.

Can I compare 2020 Census data with previous census data?

Yes, data users can compare 2020 Census data with data from prior censuses. Data users should be cautious about drawing strong inferences based on changes observed for very small geographies, such as blocks, as they will tend to have a higher amount of noise relative to larger areas. As with every census, data users should review guidance regarding methodology changes, geographic boundary changes, etc., when making comparisons.

Can I compare 2020 Census data and American Community Survey data?

Yes, data users can compare 2020 Census data with estimates from the American Community Survey. Data users should keep in mind the differences between the two sources. For example, the American Community Survey includes sampling error, whereas the decennial census does not.

How do I calculate the accuracy of user-defined geographies based on the published data?

As in prior censuses, data users may combine tabulated quantities from several geographies to create information about new user-defined geographies. Users should be advised, however, that the accuracy of these combined tabulations will depend on both the overall population size of the created geography and the created geography's distance from the geographic spine.[37] Generally, areas that include more people and areas with boundaries closer to tract or county geographies have more relative accuracy.

Technical users may download demonstration data, called privacy-protected microdata files (PPMFs), that have run 2010 Census data through the 2020 Disclosure Avoidance System software. The latest PPMF vintage 2021-06-08 is the Production Settings run, which uses the same software and settings for the 2020 production run of the redistricting data. Users can compare tabulated values from the PPMFs to published 2010 data to identify the amount of uncertainty that can be expected for a given geography or characteristic. Users can also calculate new measures of the spread of the uncertainty. For example, comparing tabulations from the PPMFs with the published 2010 data will show that 90 percent of counties have a privacy-protected total population that is within ± four people of their published total population.

The latest summary metrics are available at Developing the DAS: Demonstration Data and Progress Metrics <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

[37] John M. Abowd et al., "Geographic Spines in the 2020 Census Disclosure Avoidance System Topdown Algorithm," Working Paper CED-21-01, U.S. Census Bureau, Washington, DC, 2021, <www.census.gov/library/working-papers/2021/adrm/geographic-spines.html>.

This page is intentionally blank.

6. ADDITIONAL RESOURCES

2020 Census Data Products: Disclosure Avoidance Modernization

<www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html>

2020 Census Results

<www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-results.html>

2020 Decennial Census Visualizations and Infographics

<www.census.gov/programs-surveys/decennial-census/decade/2020/2020-visualizations.html>

2020 Census State Redistricting Data (Public Law 94-171) Summary File

<https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/2020Census_PL94_171Redistricting_StatesTechDoc_English.pdf>

Developing the DAS: Demonstration Data and Progress Metrics

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

Disclosure Avoidance Techniques Used for the 1960 Through 2010 Census

<www.census.gov/library/working-papers/2019/adrm/six-decennial-censuses-da.html>

A History of Census Privacy Protections

<www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html>

Census Protections Evolve Continuously to Address Emerging Threats

<www.census.gov/library/stories/2020/02/through-the-decades-how-the-census-bureau-protects-your-privacy.html>

2020 Disclosure Avoidance System Updates

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

GitHub Repository

<https://github.com/uscensusbureau/census2020-das-2010ddp>

Redistricting Data Program

<www.census.gov/rdo>

Decennial Census P.L. 94-171 Redistricting Data

<www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html#P1>

This page is intentionally blank.

7. GLOSSARY

Accuracy. One of four key dimensions of survey quality. Accuracy refers to the difference between the published estimate and the true value. Attributes are measured in terms of sources of error (for example, coverage, sampling, nonresponse, measurement, processing, and disclosure avoidance). Throughout this handbook, we use accuracy in the context of the Disclosure Avoidance System to refer to difference between the published data and the as-enumerated data.

Block group. A statistical subdivision of a census tract, generally defined to contain between 600 and 3,000 people and between 240 and 1,200 housing units, and the smallest geographic unit for which the U.S. Census Bureau tabulates sample data. A subdivision of a census tract (or, before 2000, a block numbering area), a block group is a cluster of blocks having the same first digit of their four-digit identifying number within a census tract.

Census Edited File (CEF). A file created by implementing edits and characteristic imputation on the CEF. Edits are used to ensure certain consistencies among characteristics. Characteristics imputation is used to ensure that each person and housing unit on the final census file has valid values in the person and housing items—sex, age, date of birth, Hispanic origin, race, relationships to householder, group quarters type, tenure, and detailed vacancy status.

Census geography. A collective term referring to the types of geographic areas used by the U.S. Census Bureau in its data collection and tabulation operations. With connecting lines, the diagram in the "Geographies and the Geographic Spine" section shows the hierarchical relationships between geographic types. For example, a line extends from states to counties because a state is comprised of many counties, and a county can never cross a state boundary.

If no line joins two geographic types, then an absolute and predictable relationship does not exist between them. For example, many places do not cross a county boundary (i.e., only one county). However, some places extend over more than one county like New York City. Therefore, an absolute hierarchical relationship does not exist between counties and places, and any tabulation involving both of these geographic types may represent only a part of one county or one place.

Census tract. A small, relatively permanent statistical subdivision of a county delineated by a local committee of census data users for presenting data. Census tracts nest within counties and their boundaries normally follow visible features but may follow legal geography boundaries and other nonvisible features in some instances. Census tracts ideally contain about 4,000 people and 1,600 housing units.

Confidentiality. The confidentiality of census data is protected under Title 13 of the U.S. Code, which prohibits the U.S. Census Bureau from disclosing any "information reported by, or on behalf of, any particular respondent" and from "(making) any publication whereby the data furnished by any particular establishment or individual under this title can be identified."[38]

Data swapping. A disclosure avoidance method used for prior censuses that "swaps" data between households in different locations that have similar characteristics on a set of variables. Which households were swapped is not public information. The selection process is highly targeted, so it is most often applied to the data with the highest disclosure risk. Often, swapping occurs within a specific geographic area so there is no effect on the population or characteristics totals for that geographic area. Because of data swapping, users should expect that tables with cells having a value of one or two do not reveal information about specific individuals. As a consequence, these cells typically do not have a high degree of accuracy.

Decennial census. The census of population and housing, taken by the U.S. Census Bureau in years ending in 0 (zero). Article I of the Constitution requires that a census be taken every 10 years for the purpose of reapportioning the U.S. House of Representatives among the states.

Differential privacy. The scientific term for a mathematical framework that quantifies the disclosure risk associated with each published statistic. By quantifying the disclosure risk of the statistics we publish, we can then use statistical noise to slightly alter the data so the link between the data and a specific person or business can't be certain. Differentially private disclosure avoidance methods precisely control the amount of statistical noise added using sophisticated mathematical formulas to assure that enough noise is added to protect confidentiality but not so much as to damage the statistical validity of our publications. The idea of using statistical noise to protect confidentiality is not new. The U.S. Census Bureau has used similar techniques for decades.

[38] Title 13 U.S. Code, Sections 8–9.

Disclosure avoidance. Statistical methods used to treat data prior to release to ensure the confidentiality of responses.

Editing and imputation. Editing is the process of ensuring consistencies among characteristics for a person or people in a household. Characteristic imputation is the process used to fill in missing or misreported data via assignment, allocation, or substitution.

Epsilon. A measure of privacy loss. Higher values of epsilon result in more privacy loss, whereas lower values result in less privacy loss. Epsilon may also be referred to as the privacy-loss budget, although in the TopDown Algorithm, the privacy-loss budget is allocated using the parameter rho defined in Zero-Concentrated Differential Privacy.

Group quarters (GQ) facilities. A GQ facility is a place where people live or stay that is normally owned or managed by an entity or organization providing housing and/or services for the residents. These services may include custodial or medical care, as well as other types of assistance. Residency is commonly restricted to those receiving these services. People living in GQ facilities are usually not related to one another. There are two general categories of group quarters facilities: institutional group quarters (such as correctional facilities) and noninstitutional group quarters (such as college/university student housing).

Group quarters population. Includes all people living in group quarters instead of housing units. Group quarters are places where people live or stay, in a group living arrangement that is owned or managed by an entity or organization providing housing and/or services for the residents.

Housing unit. A housing unit is a house, an apartment, a mobile home or trailer, a group of rooms, or a single room occupied as separate living quarters, or if vacant, intended for occupancy as separate living quarters. Separate living quarters are those in which the occupants live separately from any other individuals in the building and which have direct access from outside the building or through a common hall. For vacant units, the criteria of separateness and direct access are applied to the intended occupants whenever possible.

Indirect identification. Indirect identification refers to using information in conjunction with other data elements to reasonably infer the identity of a respondent. For example, data elements, such as a combination of gender, race, date of birth, geographic indicators, or other descriptors, may be used to identify an individual respondent.

Invariant. A number reported exactly as enumerated.

Post-processing. In the context of disclosure avoidance, a process used by the U.S. Census Bureau to impose certain consistencies on the published data (for example, ensuring that the population for counties within a state sums up to the state's total population, converting protected tables to microdata).

Privacy-loss budget. A measure of global disclosure risk. Higher values for the privacy-loss budget result in more privacy loss, whereas lower values result in less privacy loss. Privacy-loss budget may also be referred to as epsilon or, in the case of the TopDown Algorithm as rho, a related parameter.

Rho. A measure of disclosure risk used in the Zero-Concentrated Differential Privacy framework that is used by the TopDown Algorithm. Higher values of rho result in more disclosure risk, whereas lower values result in less privacy loss. Rho may also be referred to as the privacy-loss budget.

Table suppression and cell suppression. When published statistics could result in potential disclosure of individual information, it may be necessary to suppress data from publication—either by suppressing cells within a table or suppressing entire tables of data. Refer to "Disclosure avoidance" above.

Top- and bottom-coding. Top- and bottom-coding refer to the practice of not reporting the largest (or smallest) characteristics, but grouping those with others near the top, such as reporting household sizes 1 through 3, but then reporting 4 or more to include sizes 4, 5, 6, etc.

TopDown Algorithm. An algorithm used by the U.S. Census Bureau based on the privacy-loss accounting framework of Differential Privacy that injects noise into 2020 Census data to protect the confidentiality of respondents.[39]

[39] DAS 2020 Redistricting Production Code Release, <https://github.com/uscensusbureau/DAS_2020_Redistricting_Production_Code>.

8. TECHNICAL APPENDIX A: THE PRIVACY-LOSS BUDGET FOR 2020 REDISTRICTING DATA

To achieve a given level of confidentiality protection (i.e., set the maximum possible amount of disclosure risk for a given dataset), the privacy-loss budget (PLB) acts like a dial that impacts the range of random noise that is drawn from a statistically defined probability distribution (Figure 8.1). Higher values of PLB imply more accuracy and less confidentiality. As the PLB (reflected in the terms epsilon or rho) rises, the increasingly peaked shape of the distribution means that the noise added to any given cell is increasingly likely to be zero. Lower values of PLB imply less accuracy/more protection, as the noise distribution spreads out away from zero, and larger amounts of noise added to a cell become increasingly likely. In the most extreme cases, a PLB of zero would reflect complete noise with no accuracy. A PLB value of infinity would reflect complete accuracy with no noise.

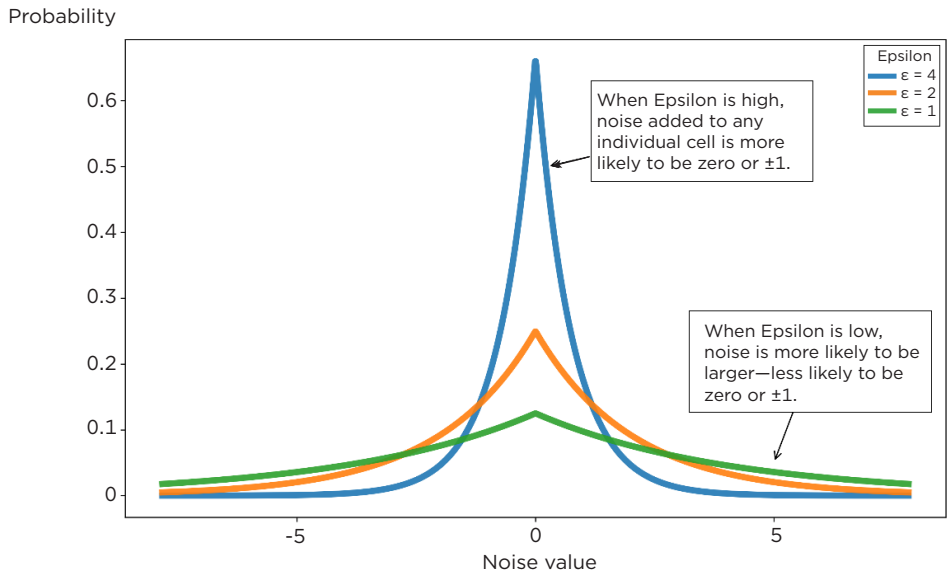
The privacy-loss budget is not the only factor that influences the shape of the distribution. The type of distribution (such as Laplace, geometric, or Gaussian) also plays a role. In "pure" differential privacy, the statistical distributions that are most commonly used, such as Laplace, allow for sizeable "outliers"—places where the amount of noise added is unusually large (very far from 0 or ±1).

For the purposes of decennial census data, confidentiality concerns need to be balanced with the accuracy of the data and adding large amounts of noise to some cells may harm the data's fitness for use.

To address this issue, the U.S. Census Bureau chose to implement a framework of Zero-Concentrated Differential Privacy (zCDP), based on a different statistical distribution (discrete Gaussian). This shift means that for the same level of privacy-loss budget, zCDP has lower probability of injecting unusually large amounts of noise than pure differential privacy would.[40]

[40] Statisticians would refer to this as the zCDP distribution having thinner "tails" (lower probability that an observation will be very far from zero) than the distributions most commonly used in pure differential privacy.

Figure 8.1. **The Privacy-Loss Budget (Epsilon) Acts as a Dial on the Level of Noise**



Source: U.S. Census Bureau.

The switch to zCDP significantly reduces the likelihood of outliers, yielding substantially greater accuracy for comparable privacy risk. It does so in part by modifying the mechanics of the mathematical guarantee.

The privacy-loss budget is often referred to by a single value of epsilon or rho for the entire dataset, but the budget itself is allocated across various dimensions of the dataset. Within the mechanics of zCDP, the privacy-loss budget is allocated to queries using the parameter, rho. Through the inclusion of a third parameter delta, which interprets the strength of the privacy guarantee represented by rho, rho can be used to calculate the global epsilon for any given value of delta.

For the 2020 Census Public Law 94-171 redistricting files, the privacy-loss budget was allocated as shown below, where the fractions in each cell represent the share of privacy-loss budget allocated.

Privacy-Loss Budget Allocations for the P.L. 94-171 Redistricting Data: United States

Global Privacy-Loss Budget: People

Global rho	2.56
Global epsilon	17.14
Delta	10^{-10}

Source: U.S. Census Bureau.

Privacy-Loss Budget: People

Geographic level	Rho allocation
United States	104/4,099
State	1,440/4,099
County	447/4,099
Tract	687/4,099
Optimized block group[1]	1,256/4,099
Block	165/4,099

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

Source: U.S. Census Bureau.

Per Query Privacy-Loss Budget: People

Query	Geographic level and rho allocation					
	United States	State	County	Tract	Optimized block group[1]	Block
TOTAL (1 cell)[2]	N	3,773/4,097	3,126/4,097	1,567/4,102	1,705/4,099	5/4,097
CENRACE (63 cells)	52/4,097	6/4,097	10/4,097	4/2,051	3/4,099	9/4,097
HISPANIC (2 cells)	26/4,097	6/4,097	10/4,097	5/4,102	3/4,099	5/4,097
VOTINGAGE (2 cells)	26/4,097	6/4,097	10/4,097	5/4,102	3/4,099	5/4,097
HHINSTLEVELS (3 cells)	26/4,097	6/4,097	10/4,097	5/4,102	3/4,099	5/4,097
HHGQ (8 cells)	26/4,097	6/4,097	10/4,097	5/4,102	3/4,099	5/4,097
HISPANIC*CENRACE (126 cells)	130/4,097	12/4,097	28/4,097	1,933/4,102	1,055/4,099	21/4,097
VOTINGAGE*CENRACE (126 cells)	130/4,097	12/4,097	28/4,097	10/2,051	9/4,099	21/4,097
VOTINGAGE*HISPANIC (4 cells)	26/4,097	6/4,097	10/4,097	5/4,102	3/4,099	5/4,097
VOTINGAGE*HISPANIC*CENRACE (252 cells)	26/241	2/241	101/4,097	67/4,102	24/4,099	71/4,097
HHGQ*VOTINGAGE*HISPANIC*CENRACE (2,016 cells)	189/241	230/4,097	754/4,097	241/2,051	1,288/4,099	3,945/4,097

N Not applicable.

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

[2] The TOTAL query (total population) is held invariant at the state level. This note pertains to the interpretation of the entry in the State column of this row only. This rho allocation assigned to TOTAL at the state level is the amount assigned to the state-level queries for the total population of all American Indian and Alaska Native (AIAN) tribal areas within the state and for the total population of the remainder of the state, for the 36 states that include AIAN tribal areas.

Source: U.S. Census Bureau.

Global Privacy-Loss Budget: Units

Global rho	0.07
Global epsilon	2.47
Delta	10^{-10}

Source: U.S. Census Bureau.

Privacy-Loss Budget: Units

Geographic level	Rho allocation
United States	1/205
State	1/205
County	7/82
Tract	364/1,025
Optimized block group[1]	1,759/4,100
Block	99/820

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

Source: U.S. Census Bureau.

Per Query Privacy-Loss Budget: Units

Query	Geographic level and rho allocation					
	United States	State	County	Tract	Optimized block group[1]	Block
Detail (2 cells)	1/1	1/1	1/1	1/1	1/1	1/1

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

Source: U.S. Census Bureau.

Per Attribute Epsilons

Attribute	Epsilon allocation
HHGQ	7.24
VOTINGAGE	7.57
HISPANIC	10.04
CENRACE	10.08
H1	2.47

Source: U.S. Census Bureau.

Cross-Universe Epsilons: People + Units

Geographic level	Epsilon allocation
Block within block group	3.06
Block within tract	9.62
Block within county	12.04
Block within state	13.40
Block within United States	17.18
All levels	**17.44**

Source: U.S. Census Bureau.

Cross-Universe Rho: People + Units

Geographic level	Rho allocation
Block within block group	0.11
Block within tract	0.93
Block within county	1.38
Block within state	1.67
Block within United States	2.56
All levels	**2.63**

Source: U.S. Census Bureau.

Privacy-Loss Budget Allocations for the P.L. 94-171 Redistricting Data: Puerto Rico

Global Privacy-Loss Budget: People

Global rho	2.56
Global epsilon	17.14
Delta	10^{-10}

Source: U.S. Census Bureau.

Privacy-Loss Budget: People

Geographic level	Rho allocation
Puerto Rico	689/4,099
Municipio	695/4,099
Tract	772/4,099
Optimized block group[1]	1,778/4,099
Block	165/4,099

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

Source: U.S. Census Bureau.

Per Query Privacy-Loss Budget: People

Query	Geographic level and rho allocation				
	Puerto Rico	Municipio	Tract	Optimized block group[1]	Block
TOTAL (1 cell)	N	3,126/4,097	1,467/4,102	1,876/4,103	5/4,097
CENRACE (63 cells)	11/108	10/4,097	13/4,102	4/4,103	9/4,097
HISPANIC (2 cells)	11/108	10/4,097	1/586	4/4,103	5/4,097
VOTINGAGE (2 cells)	11/108	10/4,097	1/586	4/4,103	5/4,097
HHINSTLEVELS (3 cells)	11/108	10/4,097	1/586	4/4,103	5/4,097
HHGQ (8 cells)	11/108	10/4,097	1/586	4/4,103	5/4,097
HISPANIC*CENRACE (126 cells)	53/513	28/4,097	866/2,051	749/4,103	21/4,097
VOTINGAGE*CENRACE (126 cells)	53/513	28/4,097	15/2,051	10/4,103	21/4,097
VOTINGAGE*HISPANIC (4 cells)	11/108	10/4,097	1/586	4/4,103	5/4,097
VOTINGAGE*HISPANIC*CENRACE (252 cells)	56/513	101/4,097	50/2,051	27/4,103	71/4,097
HHGQ*VOTINGAGE*HISPANIC*CENRACE (2,016 cells)	25/342	754/4,097	725/4,102	1,417/4,103	3,945/4,097

N Not applicable.

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

Source: U.S. Census Bureau.

Global Privacy-Loss Budget: Units

Global rho	0.07
Global epsilon	2.47
Delta	10^{-10}

Source: U.S. Census Bureau.

Privacy-Loss Budget: Units

Geographic level	Rho allocation
Puerto Rico	5,047/876,580
Municipio	18,746/219,145
Tract	94,451/262,974
Optimized block group[1]	281,911/657,435
Block	99/820

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products wil be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

Source: U.S. Census Bureau.

Per Query Privacy-Loss Budget: Units

Query	Geographic level and rho allocation				
	Puerto Rico	Municipio	Tract	Optimized block group[1]	Block
Detail (2 cells)	1/1	1/1	1/1	1/1	1/1

[1] The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

Source: U.S. Census Bureau.

Per Attribute Epsilons

Attribute	Epsilon allocation
HHGQ	8.69
VOTINGAGE	9.49
HISPANIC	11.65
CENRACE	11.69
H1	2.47

Source: U.S. Census Bureau.

Cross-Universe Rho: People + Units

Geographic level	Rho allocation
Block within block group	0.11
Block within tract	1.25
Block within municipio	1.76
Block within Puerto Rico	2.20
All levels	**2.63**

Source: U.S. Census Bureau.

Cross-Universe Epsilons: People + Units

Geographic level	Epsilon allocation
Block within block group	3.06
Block within tract	11.39
Block within municipio	13.82
Block within Puerto Rico	15.72
All levels	**17.44**

Source: U.S. Census Bureau.