Making Differential Privacy Work for Census Data Users

Cory McCartan[1] Tyler Simko[2] Kosuke Imai[2,3]

[1]Center for Data Science, New York University, New York, New York, United States of America,
[2]Department of Government, Harvard University, Cambridge, Massachusetts, United States of America,
[3]Department of Statistics, Harvard University, Cambridge, Massachusetts, United States of America

ABSTRACT

The U.S. Census Bureau collects and publishes detailed demographic data about Americans, which are heavily used by researchers and policymakers. The bureau has recently adopted the framework of differential privacy in an effort to improve confidentiality of individual census responses. A key output of this privacy protection system is the Noisy Measurement File (NMF), which is produced by adding random noise to tabulated statistics. The NMF is critical to understanding any errors introduced in the data, and performing valid statistical inference on published census data. Unfortunately, the current release format of the NMF is difficult to access and work with. We describe the process we use to transform the NMF into a usable format, and provide recommendations to the bureau for how to release future versions of the NMF. These changes are essential for ensuring transparency of privacy measures and reproducibility of scientific research built on census data.

Keywords: census data, differential privacy, open science

1. Introduction

For the 2020 Decennial Census, the Census Bureau adopted a new Disclosure Avoidance System (DAS) based on differential privacy. The DAS was designed to protect the confidentiality of responses by injecting statistical noise into the tabulations of a confidential individual census data set. The bureau postprocessed the resulting Noisy Measurement File (NMF) to produce the final tabulated statistics. Postprocessing ensured that the final tabular statistics met various requirements, including nonnegative counts, invariant state populations, and consistency across different geographical levels of census hierarchy.

However, as some of the bureau's staff themselves warned, postprocessing also introduces biases in the tabulated statistics (see, e.g., National Academies of Sciences, Engineering, and Medicine, 2020, p. 42).[1] Some of these biases, like inflating the average population of some zero-population areas, were expected; others were discovered by concerned data users and in cases required changes to the DAS (Dwork et al., 2021; Kenny et al., 2021; National Congress of American Indians, 2021; JASON, 2022; Kenny, Kuriwaki, McCartan, Rosenman, et al., 2023; Scariano & Youngs, 2022).

The NMF, which is an intermediate product of the DAS, is an invaluable resource for researchers and other census data users to understand the error introduced by the DAS. The NMF is also essential for performing statistically valid analyses of census data that properly account for DAS-introduced error (Kenny, Kuriwaki, McCartan, Simko & Imai, 2023).

In 2021, however, the bureau did not make the NMF available and only released the final (postprocessed) tabulated statistics. Through an open letter (Dwork et al., 2021), a Freedom of Information Act (FOIA) request,

and subsequent litigation (Phillips v. Census Bureau, 2022), dozens of scholars and data users urged the bureau to release the NMF, which allows one to understand and correct potential biases induced by the postprocessing procedure. Following these developments, in April 2023, the bureau released a demonstration NMF based on the 2010 Census data, and subsequently in June made available the NMF for the 2020 P.L.94-171 redistricting data.[2] Furthermore, in May 2023, the bureau also published the 2020 Demographic and Housing Characteristics File (DHC) and plans to publish its NMF later this year.

We commend the Bureau's decision to provide the NMF, which will help advance social science research, improve policy decisions, and further strengthen the DAS itself. But as the bureau's own chief scientist recently wrote in a court filing, the NMF "in [its] native format would be practically unusable by anyone not intimately familiar with the minutiae of the [DAS] production code base" (Keller, 2023, p. 10). To maximize the benefits of the released NMF, therefore, we believe that the bureau must substantially improve the way in which the NMF is formatted and released (McCartan et al., 2023).

Below, we first explain several obstacles many end users are likely to face when accessing, processing, and using the released NMF for statistical analyses. We then provide recommendations for how the Census Bureau could release the NMF going forward in a manner and format that would address these issues.

2. The Census Noisy Measurement File

After collecting census questionnaires from the U.S. population, the Census Bureau fills in missing responses using other data sources and imputation methods. The resulting confidential data set is called the Census Edited File (CEF), which has a row for every person in the United States. To convert the CEF into statistics that can be reported publicly, the DAS proceeds in two steps, as illustrated in Figure 1. The process described in this section is a simplification of the full workings of the DAS; further details may be found in Abowd et al. (2022).
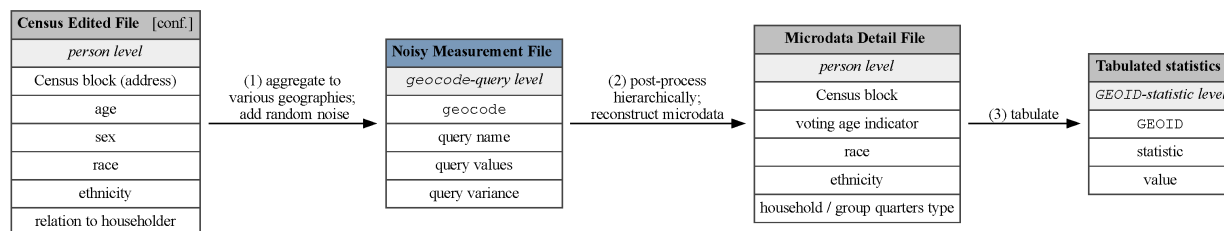

Figure 1. How the Census Bureau produces differentially private census tabulations for person-level redistricting (P.L. 94-171) data, starting from the confidential Census Edited File. Step (1) provides the differential privacy guarantee. Step (2) is the core of the TopDown Algorithm, detailed in . Together, steps (1) and (2) comprise the Disclosure Avoidance System (DAS).

First, a series of *queries* are made for every geographic unit on the NMF *geographic spine*: states, counties, tracts, optimized block groups, and blocks.[3] Notably, the geographic spine is split between regions inside and

outside of American Indian / Alaska Native (AI/AN) areas. This means that, for example, a county containing a Native reservation is split into two geographic units: the AI/AN area and the remainder of the county. Each of these has queries made separately. These queries are contingency tables for various statistics. One example is the "total population" query, which counts the total number of people in the geographic unit. Another is the "voting age by Hispanic" query, which reports a 2-by-2 table of counts of people in those binary categories.

For the decennial census redistricting data, there are 11 total queries for each geography, including a "detailed" query, which is the full contingency table of voting age, Hispanic status, race, and household type, with 2,016 values. The counts that make up each of these queries have a certain amount of discrete Gaussian noise added by the DAS (Canonne et al., 2020), in accordance with a privacy loss schedule set by the bureau, which is recorded in the `variance` column of the NMF. It is this step that ensures the released statistics meet a certain standard of differential privacy (Bun & Steinke, 2016; Kifer et al., 2022).

Second, the set of noised queries across all geographic units and levels is postprocessed using a multipass optimization routine. The goal of this postprocessing step is to produce a set of counts that is self-consistent across geographic units and different statistics, and respects commonsense constraints like having nonnegative population counts. From these consistent counts, synthetic person-level data can be reconstructed, yielding the Microdata Detail File. Unlike the CEF, the exact format of the Microdata Detail File depends on the particular census product. It is then straightforward to tabulate this microdata to produce traditional census statistics for each geography in the country.

3. Processing the Noisy Measurement File

We identified four key obstacles that make it difficult to directly use the 2020 NMF released in June 2023 (Abowd et al., 2023b):

1. Collecting the documentation necessary to open and parse the NMF;
2. Converting the NMF data files to a familiar format amenable to computer and statistical analysis;
3. Navigating the many-to-many mapping between NMF *queries* and tabulated *statistics*; and
4. Connecting the *geocodes* used in the NMF to traditional census *tabulation geographies*.

This section describes these obstacles and our process for overcoming them.

The NMF was accompanied by sparse documentation that was not alone sufficient to reconstruct the format and coding of the data. The bureau has made the code that implements the DAS publicly available, but this code as well is sparsely documented and highly unlikely to be accessible to nonexperts. Particularly troubling is that the codebook for the variables (Male = 0, Female = 1, etc.) exists only as a set of undocumented Python files in a folder in this code repository. This problem is compounded for statistics that involve cross-tabulations of base variables. For example, a 'sex by Hispanic' table could be reported in the order of (Non-Hispanic male, Non-Hispanic female, Hispanic male, Hispanic female), or it could be reported in the order of (Non-Hispanic

male, Hispanic male, Non-Hispanic female, Hispanic female). There is no clear documentation as to which of these orders variables are combined in, and the actual order depends on the way the NumPy array library for Python works. Despite these challenges, by manually copying the codebook from the Python code and running numerous experiments on the NMF data, we were able to reconstruct the structure and variable coding of the NMF ourselves.[4]
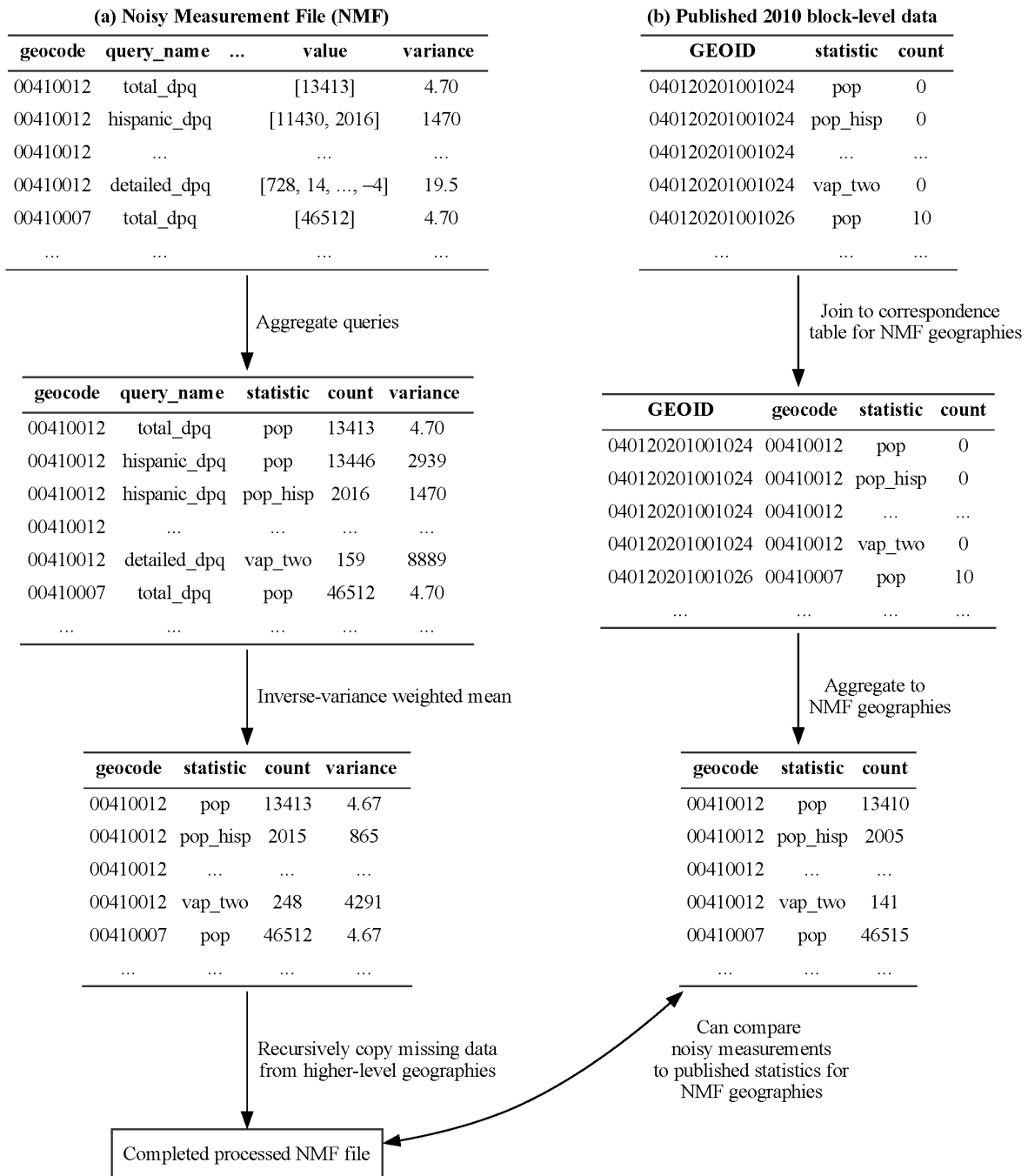
(a) Noisy Measurement File (NMF)

geocode	query_name	...	value	variance
00410012	total_dpq		[13413]	4.70
00410012	hispanic_dpq		[11430, 2016]	1470
00410012	...		...	...
00410012	detailed_dpq		[728, 14, …, −4]	19.5
00410007	total_dpq		[46512]	4.70
...	...		...	...

Aggregate queries

geocode	query_name	statistic	count	variance
00410012	total_dpq	pop	13413	4.70
00410012	hispanic_dpq	pop	13446	2939
00410012	hispanic_dpq	pop_hisp	2016	1470
00410012	...	...	...	...
00410012	detailed_dpq	vap_two	159	8889
00410007	total_dpq	pop	46512	4.70
...	...	...	...	...

Inverse-variance weighted mean

geocode	statistic	count	variance
00410012	pop	13413	4.67
00410012	pop_hisp	2015	865
00410012	...	...	...
00410012	vap_two	248	4291
00410007	pop	46512	4.67
...	...	...	...

Recursively copy missing data from higher-level geographies

Completed processed NMF file

(b) Published 2010 block-level data

GEOID	statistic	count
040120201001024	pop	0
040120201001024	pop_hisp	0
040120201001024	...	...
040120201001024	vap_two	0
040120201001026	pop	10
...	...	...

Join to correspondence table for NMF geographies

GEOID	geocode	statistic	count
040120201001024	00410012	pop	0
040120201001024	00410012	pop_hisp	0
040120201001024	00410012	...	...
040120201001024	00410012	vap_two	0
040120201001026	00410007	pop	10
...	...	...	...

Aggregate to NMF geographies

geocode	statistic	count
00410012	pop	13410
00410012	pop_hisp	2005
00410012	...	...
00410012	vap_two	141
00410007	pop	46515
...	...	...

Can compare noisy measurements to published statistics for NMF geographies

Figure 2. (a) Our workflow for processing the Noisy Measurement File into a usable form, illustrated as samples from the NMF data table as it is transformed, and **(b)** the necessary workflow for processing published statistics so that they are directly comparable with noisy measurements. The _dpq suffix is used by the bureau to indicate a differentially private query (as opposed to a traditional tabulation statistic).

The NMF is made available as a series of large Apache Parquet files, a flexible archival storage format that underpins much of modern big data infrastructure. This flexibility allows the bureau to store the NMF in a non-

rectangular format: each row of the NMF contains normal records like the query name and variance, but also a variable-length vector containing the query results (collapsed from a multidimensional contingency table). This nested data structure unfortunately cannot be directly analyzed using standard tools designed for rectangular data.

As described in [Section 2](#), at this stage each row in the NMF corresponds to a single query run for a particular geographic unit (see Figure 2(a), top panel). Each query's result can be aggregated to produce an estimate of various traditional census statistics. For example, the 'voting age by Hispanic' query result for some geography g is stored in a single row of the NMF as the following vector with 4 elements:

$$m_g = (\overbrace{1050}^{\text{Non-Hisp. children}}, \underbrace{204}_{\text{Hisp. children}}, \overbrace{10050}^{\text{Non-Hisp. adults}}, \underbrace{1812}_{\text{Hisp. adults}})^{\top}.$$

This vector, which represents a 2×2 table, can be summed along each dimension to produce (noisy) estimates of the voting age population or the Hispanic population, or all its entries can be summed to estimate the total population of the geographic unit.

Thus, each traditional tabulated statistic, like total or voting-age population, has multiple noisy 'versions' spread across multiple queries. For example, the Hispanic population can be estimated not just by the 'Hispanic' query, but by the 'voting age by Hispanic' query, the 'race by Hispanic' query, the 'voting age by race by Hispanic' query, and the detailed query. For further analysis, data users must map the queries to statistics of interest, possibly combining information across multiple queries.

To address this challenge, we developed a series of aggregation matrices which, when multiplied by a vector of noisy query values, produces a set of tabulated statistics. For example, the aggregation matrix for the 'voting age by Hispanic' query shown above would be applied as follows:

$$\overbrace{A}^{\text{Agg. matrix}} \underbrace{m_g}_{\text{Query result}} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1050 \\ 204 \\ 10050 \\ 1812 \end{pmatrix} = (\overbrace{13116}^{\text{Pop. estimate}}, \overbrace{11862}^{\text{Hispanic estimate}}, \underbrace{2016}_{\text{VAP estimate}}, \underbrace{1812}_{\text{Hisp. VAP estimate}})^{\top}.$$

Applying these aggregation matrices to each of the queries produces multiple noisy estimates of each of the desired tabulated statistics. Applying them again to the variance of each of the queries (which is included in the NMF) calculates the variance of each of these estimates, since each entry in m_g is independent of the others. The output of this step is shown in the second panel of [Figure 2(a)](#). Then, to integrate these estimates, we take an inverse-variance weighted mean, which produces a minimum-variance estimator of each (true) tabulated statistic.[5] The output of this step is shown in the third panel of [Figure 2(a)](#).

Mapping queries to statistics yields data that look similar to the tabulated tables traditionally released by the bureau: each row corresponds to a tabulated statistic for a particular geographic unit (along with the error variance). But the similarity is only superficial—the geographic units in the NMF, identified by numeric *geocodes*, do not directly correspond to traditional census tabulation geographies, identified by so-called *GEOIDs*. Each census block with nonzero population has a unique associated geocode. Geocodes are combinations of traditional census information contained in *GEOIDs* (e.g. FIPS codes) and information on TopDown processing steps (e.g. an indicator for whether the geography is on the AI/AN spine or not). A full explanation is provided in Kenny, Kuriwaki, McCartan, Simko & Imai (2023, Table 1), and original detail is available in the NMF documentation (Abowd et al., 2023a, see, e.g. footnotes 14–15).

While blocks correspond to blocks, every other geography aside from the country as a whole is different (Cumings-Menon et al., 2022). Part of this difference is due to separating the geographic spine for regions inside and outside of AI/AN areas as mentioned above. This decision means that states, counties, tracts, and optimized block groups are split between AI/AN regions and must be recombined for many analyses. Traditional block groups are also replaced with 'optimized block groups,' which are designed to minimize DAS-induced error for specific off-spine geographies such as census places. Finally, any geography with neither housing units nor group quarters facilities is removed from the NMF.[6]

An additional complication is that when a geographic unit contains only a single subunit (for example, a tract containing only one block group), the subunit's privacy loss budget is reallocated to the higher level unit, and the subunit is removed from the NMF (see Abowd et al., 2022). This creates 'holes' in the NMF and substantially increases the difficulty of analyzing the NMF.

Thankfully, a full listing of (populated) block geocodes exists in auxiliary constraint files released alongside the NMF. Since the NMF block-level geocodes can be split into two pieces, with the second piece corresponding to the block's tabulation GEOID, this allows us to build a correspondence table that links every census block to its traditional tabulation geographies and NMF-specific geographies. Finally, by recursing down the geographic hierarchy and identifying geographies that are missing in the NMF but present in the auxiliary data, we can copy data from higher geographic levels and fill the 'holes' in the NMF.

The correspondence tables linking NMF blocks to traditional tabulation geographies are only partially satisfying. They allow analysts to aggregate any block-level data, such as the final published census data (the right table in Figure 1) to both NMF and traditional geographies. Yet, because noisy measurements are not consistent across geographic levels—the sum of tract populations does not equal the county population, in general—there are many ways to aggregate NMF data to larger geographies. The simplest way—aggregating NMF block-level data—also produces far noisier estimates, which may be practically unusable for many purposes.

A better way involves aggregating noisy estimates from the largest possible NMF geographies. For example, to produce an NMF-based estimate of a school district's population, we would add up the NMF estimates of all the tracts contained fully within the district, then all the optimized block groups that are contained fully within the remainder, and then finally the remaining blocks. If applied to all school districts in the country, this process will produce a minimum-variance unbiased estimates of the true tabulated populations, under the constraint that each of the district estimates remain statistically independent of the others and are built out of only geographies contained in each district. This approach is used in [Kenny, Kuriwaki, McCartan, Simko & Imai (2023)](). Further research is warranted, however, in establishing minimum-variance estimators under other constraints, and evaluating the tradeoffs between different estimation approaches.

For now, we recommend that most researchers interested in comparing NMF data to published tabulations should proceed by aggregating published block data to NMF geographies (rather than the other way around). This is illustrated in [Figure 2(b)]().

4. Recommendations for the Census Bureau

Based on our experience obtaining the NMF and building tools to load, format, and analyze it, we offer the following recommendations that will significantly improve the usability of the NMF. Our recommendations are focused on facilitating primary use cases of the NMF: understanding the DAS-introduced error in published statistics and performing statistically valid inference on those data.

1. Data access and documentation

   (a) Host the raw NMF files on an easy-to-use, publicly accessible website. The NMFs were first posted to the Globus service, which required users to create an account and install their third-party software locally in order to download any data, even the documentation. Some of the NMFs have been released on other platforms, such as the Harvard Dataverse; others have not. Ideally, users could easily (and programmatically) download the individual Parquet files that make up the NMF from the bureau's own website.[7]

   (b) Consider un-nesting the NMF file before publication so that the data are in a rectangular format, with each query's histogram bins fully labeled. In addition to making the data much more amenable to standard data processing tools, the inclusion of labels would additionally obviate the need for users to manually copy values from codebooks.

   (c) In the event that further reformatting and aggregation is performed by the bureau on the NMF (see below for specific recommendations), provide application programming interface (API) access to these reformatted data, similar to the API for existing Decennial Census tabulations or American Community Survey (ACS) results.

   (d) Produce centralized NMF documentation and make it available separately from the data. This

documentation should explain the high-level structure of the NMF and its relation to published decennial statistics and tabulation geographies. The documentation should also contain sufficient detail and code books needed to properly read and format the file. The existing documentation provided alongside the raw NMF is around 8 pages long and, while technical, does not contain enough detail on its own to allow researchers to work with the NMF.

2. Data parsing and formatting

(a) Fill in NMF geographies missing due to privacy budget reallocation with data copied from higher level geographies. While the original NMF reflects the structure of the geographic spine used as input to the DAS, without this filling-in, researchers are unable to use the NMF to evaluate bias and noise or make statistically valid inferences from privacy-protected data.

(b) In the event that the un-nested NMF is too large for users to work with, release code in standard data analysis languages (R, Python, Stata, etc.) that demonstrates how to load the NMF in and attach labels to the noisy queries.

3. Connecting NMF queries to tabulated statistics

(a) Provide aggregation specifications that link queries to traditional tabulation statistics in the redistricting or DHC files. These specifications should be designed as tables that can be joined to the NMF and then summarized to produce the necessary tabulations.

(b) Produce an additional version of the NMF for which these aggregations have already performed, and the results possibly combined using inverse-variance weighted means, so that each statistic has a single minimum-variance estimate. As described in [Section 3](), this aggregation and weighting can be done independently across geographies, preserving the key statistical advantage of the NMF data—independent, unbiased noise—while making it directly usable for downstream tasks. Aggregation-and-weighting also significantly reduces the size of the data, allowing the output to be hosted by the bureau and processed more easily by most end users. Other than the differences in the geographic spine, this data product could look much like the ACS tabulations, which are accompanied by a margin of error estimate.

4. Connecting the geocodes used in the NMF to traditional census tabulation geographies

(a) Provide full block assignment files (BAFs) for NMF geographies, including blocks with no housing units.

(b) Produce shapefiles for the NMF geographies so that they can be used directly in analyses, if desired, and compared to geographies that are of interest to end users such as school or voting districts.

(c) Provide generalized geography assignment files that describe the best way to build each traditional tabulation geography (on- and off-spine) out of NMF geographies, following the approach discussed in [Section 3](). Such generalized assignment files would enable researchers to have independent, unbiased noisy measurements of statistics of interest for their desired geographies, which are also minimum-variance.[8]

If implemented, these recommendations would produce a NMF that is much more accessible for and familiar to data users, enabling them to focus on developing methods to analyze noisy census measurements, and answering substantive research questions.

5. Conclusion

We commend the Census Bureau for their decision to release the Noisy Measurement File, which has the potential to allow census data users to understand and account for the DAS-introduced error for a wide array of analyses. In addition, the NMF demonstration data release offered a potential opportunity to discuss improvements to data usability *before* a wide range of census stakeholders begin to build new data pipelines to incorporate the NMF into their work.[9] Unfortunately, little change was made in the final release of the NMF to overcome the obstacles we identified in the initial version of this article.[10]

The NMF, as currently released, is difficult to process and use for the kinds of tasks that will allow practitioners to evaluate and correct for DAS-introduced error in their analyses. Fortunately, many of these difficulties are logistical and not due to scientific necessity. We believe that adopting our recommendations will result in a more accessible and useful NMF for many census data users. The Census Bureau, as the developer of the DAS and of course the principal expert on census data, is best-positioned to make these changes and produce a data product that is both statistically sound and practically useful. Our recommendations are consistent with the bureau's own stated goals around data usability, such as their pioneering Enterprise Data Lake (EDL) initiative to modernize data collection, processing, and dissemination.[11]

Data releases from the U.S. Census Bureau serve as the backbone for a great deal of scientific analyses and policy decisions. We hope these discussions around data usability, accuracy, and privacy in the census can continue to serve as a useful precedent for data providers and data users in the future.

References

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*, (Special Issue 2). https://doi.org/10.1162/99608f92.529e3cb9

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2023a). 2010 Census Production Settings Redistricting Data (P.L. 94-171) Demonstration Noisy Measurement File (2023-04-03). *Harvard Dataverse*.

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2023b). 2020 Census

Redistricting Data (P.L. 94-171) Noisy Measurement File, Version 1. *Harvard Dataverse*, DOI: 10.7910/DVN/5LAVKV.

Bun, M.. & Steinke, T. (2016). *Concentrated differential privacy: Simplifications, extensions, and lower bounds*. ArXiv. https://doi.org/10.48550/arXiv.1605.02065

Canonne, C. L., Kamath, G., & Steinke, T. (2020). The discrete Gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33, 15676–15688.

Cumings-Menon, R., Abowd, J. M., Ashmead, R., Kifer, D., Leclerc, P., Ocker, J., Ratcliffe, M., & Zhuravlev, P. (2022). *Geographic spines in the 2020 Census Disclosure Avoidance System*. ArXiv. https://doi.org/10.48550/arXiv.2203.16654

Dwork, C., Greenwood, R., & King, G. (2021). Letter to U.S. Census Bureau: Request for release of "noisy measurements file" by September 30 along with redistricting data products. Public letter, available at https://tinyurl.com/yahq2lz8

Hawes, M. B. (2020). Implementing differential privacy: Seven lessons from the 2020 United States Census. *Harvard Data Science Review*, *2*(2).https://doi.org/10.1162/99608f92.353c6f99

JASON (2022). *Consistency of Data Products and Formal Privacy Methods for the 2020 Census (JSR-21-02, 2022 January 11)*. (Tech. report). The MITRE Corporation.

Keller, S. A. (2023). Declaration of Sallie Ann Keller. Docket Entry 23, Appendix C, Phillips *v.* Census Bureau, Case 1:2022cv09304, US District Court for the Southern District of New York. Docket available at https://www.courtlistener.com/docket/65646969/phillips-v-united-states-bureau-of-the-census/

Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T., & Imai, K. (2021). The impact of the US Census Disclosure Avoidance System on redistricting and voting rights analysis. *Science Advances*, *7*(41), 1–17.

Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T., & Imai, K. (2023). Comment: The essential role of policy evaluation for the 2020 Census Disclosure Avoidance System. *Harvard Data Science Review*, (Special Issue 2), 1–16. https://doi.org/10.1162/99608f92.abc2c765

Kenny, C. T., Kuriwaki, S., McCartan, C., Simko, T., & Imai, K. (2023). *Evaluating bias and noise induced by the U.S. Census Bureau's privacy protection methods*. ArXiv, https://doi.org/10.48550/arXiv.2306.07521

Kifer, D., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Leclerc, P., Machanavajjhala, A., Sexton, W., & Zhuravlev, P. (2022). *Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census*. ArXiv. https://doi.org/10.48550/arXiv.2209.03310

McCartan, C., Simko, T., & Imai, K. (2023). Researchers need better access to US census data. *Science*, *380*(6648), 902–903. https://doi.org/10.1126/science.adi7004

National Academies of Sciences, Engineering, and Medicine (2020). *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*. National Academies Press.

National Congress of American Indians (2021). Letter to Dr. Ron S. Jarmin from Dante Desiderio, Chief Executive Officer. Public letter.

Phillips v. Census Bureau (2022). Phillips v. Census Bureau. 1:2022cv09304, US District Court for the Southern District of New York.

Scariano, J., & Youngs, I. (2022). *Balancing utility versus privacy in the 2020 Census: Sentiments from data users*. SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4089888

Footnotes

1. See also the transcript of a Bureau webinar: https://www2.census.gov/about/training-workshops/2021/2021-05-13-das-transcript.pdf. ↵

2. Prior to these developments, members of the DAS team mentioned the possibility of releasing the NMF (Hawes, 2020; National Academies of Sciences, Engineering, and Medicine, 2020, see, e.g. Section 11.2) . See also https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_197520.pdf and https://www.usenix.org/sites/default/files/conference/protected-files/enigma2021_slides_garfinkel.pdf. JASON (2022) also made a recommendation to release the NMF. Nevertheless, the bureau refused to release the NMF during litigation, claiming it was "confidential" and could lead to individuals being identified (see the relevant FOIA denial: https://storage.courtlistener.com/recap/gov.uscourts.nysd.588699/gov.uscourts.nysd.588699.14.2.pdf) before changing its course during litigation settlement in April 2023; see https://storage.courtlistener.com/recap/gov.uscourts.nysd.588699/gov.uscourts.nysd.588699.20.0.pdf. ↵

3. The geographic spine is used for the DHC file is different from the one used for the P.L.94-171 redistricting data. ↵

4. The code is available at https://github.com/uscensusbureau/DAS_2020_Redistricting_Production_Code ↵

5. This induces some correlation between noisy measurements *within* each geography, but maintains independence between geographies and across geographic levels. Where this correlation is statistically unhelpful, aggregation matrices can be built that only take the minimum-variance query for each statistic, sacrificing lower variance for ensuring completely independent statistics. ↩

6. Formally, the bureau imposes a *structural zero* requirement that blocks with no housing units report zero population. ↩

7. The Census Bureau had previously cited the size of the files as a reason to use a third-party data storage solution, but as the bureau's chief scientist recently admitted, the large size was due to a poor choice of file format . ↩

8. Specifically, this would ensure that resulting estimates are minimum-variance, subject to the constraint that they are built only out of NMF geographies fully contained within tabulation geographies. This in turn means that estimates be statistically independent across geographies. ↩

9. The Census Bureau has announced multiple 2020 NMF data releases. https://www.census.gov/newsroom/press-releases/2023/2020-census-data-products-schedule-updates.html ↩

10. Some positive changes were made, including publishing copies of some of the NMFs on the Harvard Dataverse and Amazon Web Services Open Data Registry, and publishing copies of the READMEs that accompany the NMFs on the Census Bureau's FTP server. ↩

11. https://www.census.gov/newsroom/blogs/research-matters/2022/10/technology-transformation.html ↩

References

- Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2023). *2010 Census Production Settings Redistricting Data (P.L. 94-171) Demonstration Noisy Measurement File (2023-04-03)*. Harvard Dataverse. ↩
- Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2023b). 2020 Census Redistricting Data (P.L. 94-171) Noisy Measurement File, Version 1. *Harvard Dataverse*, DOI: 10.7910/DVN/5LAVKV.

  ↩
- Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*, (Special Issue 2). https://doi.org/10.1162/99608f92.529e3cb9 ↩

- Bun, M.. & Steinke, T. (2016). *Concentrated differential privacy: Simplifications, extensions, and lower bounds*. ArXiv. https://doi.org/10.48550/arXiv.1605.02065

  ↩

- Canonne, C. L., Kamath, G., & Steinke, T. (2020). The discrete Gaussian for differential privacy. *Advances in Neural Information Processing Systems*, *33*, 15676–15688. ↩

- Cumings-Menon, R., Abowd, J. M., Ashmead, R., Kifer, D., Leclerc, P., Ocker, J., Ratcliffe, M., & Zhuravlev, P. (2022). *Geographic spines in the 2020 Census Disclosure Avoidance System*. ArXiv. https://doi.org/10.48550/arXiv.2203.16654

  ↩

- Dwork, C., Greenwood, R., & King, G. (2021). Letter to U.S. Census Bureau: Request for release of "noisy measurements file" by September 30 along with redistricting data products. Public letter, available at https://tinyurl.com/yahq2lz8

  ↩

- JASON (2022). *Consistency of Data Products and Formal Privacy Methods for the 2020 Census (JSR-21-02, 2022 January 11)*. (Tech. report). The MITRE Corporation.

  ↩

- Keller, S. A. (2023). Declaration of Sallie Ann Keller. Docket Entry 23, Appendix C, Phillips *v.* Census Bureau, Case 1:2022cv09304, US District Court for the Southern District of New York. Docket available at https://www.courtlistener.com/docket/65646969/phillips-v-united-states-bureau-of-the-census/

  ↩

- Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T., & Imai, K. (2021). The impact of the US Census Disclosure Avoidance System on redistricting and voting rights analysis. *Science Advances*, *7*(41), 1–17. ↩

- Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T., & Imai, K. (2023). Comment: The essential role of policy evaluation for the 2020 Census Disclosure Avoidance System. *Harvard Data Science Review*, (Special Issue 2), 1–16. https://doi.org/10.1162/99608f92.abc2c765

  ↩

- Kenny, C. T., Kuriwaki, S., McCartan, C., Simko, T., & Imai, K. (2023). *Evaluating bias and noise induced by the U.S. Census Bureau's privacy protection methods*. ArXiv, https://doi.org/10.48550/arXiv.2306.07521

  ↩

- Kifer, D., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Leclerc, P., Machanavajjhala, A., Sexton, W., & Zhuravlev, P. (2022). *Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census*. ArXiv. https://doi.org/10.48550/arXiv.2209.03310

↩

- McCartan, C., Simko, T., & Imai, K. (2023). Researchers need better access to US census data. *Science, 380*(6648), 902–903. https://doi.org/10.1126/science.adi7004 ↩

- National Academies of Sciences, Engineering, and Medicine. (2020). *2020 census data products: Data needs and privacy considerations: Proceedings of a workshop* (D. L. Cork, C. F. Citro, & N. J. Kirkendall, Eds.). National Academies Press. ↩

- National Congress of American Indians. (2021). *Letter to Dr. Ron S. Jarmin from Dante Desiderio, Chief Executive Officer*. https://www.ncai.org/policy-research-center/research-data/prc-publications/Dr._Ron_S._Jarmin_-_US_Census_Bureau_2020_Census_NCAI-_May_25,_2021.pdf. https://www.ncai.org/policy-research-center/research-data/prc-publications/Dr._Ron_S._Jarmin_-_US_Census_Bureau_2020_Census_NCAI-_May_25,_2021.pdf ↩

- *Phillips v. Census Bureau*. (2022). ↩

- Scariano, J., & Youngs, I. (2022). *Balancing utility versus privacy in the 2020 Census: Sentiments from data users*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4089888 ↩