Disclosure Avoidance in the Census Bureau's 2010 Demonstration Data Product

David Van Riper(✉), Tracy Kugler, and Steven Ruggles

Minnesota Population Center, University of Minnesota, Minneapolis, MN 55455, USA
{vanriper,takugler,rugg1001}@umn.edu

Abstract. Producing accurate, usable data while protecting respondent privacy are dual mandates of the US Census Bureau. In 2019, the Census Bureau announced it would use a new disclosure avoidance technique, based on differential privacy, for the 2020 Decennial Census of Population and Housing [19]. Instead of suppressing data or swapping sensitive records, differentially private methods inject noise into counts to protect privacy. Unfortunately, noise injection may also make the data less useful and accurate. This paper describes the differentially private Disclosure Avoidance System (DAS) used to prepare the 2010 Demonstration Data Product (DDP). It describes the policy decisions that underlie the DAS and how the DAS uses those policy decisions to produce differentially private data. Finally, it discusses usability and accuracy issues in the DDP, with a focus on occupied housing unit counts. Occupied housing unit counts in the DDP differed greatly from 2010 Summary File 1 differed greatly, and the paper explains possible sources of the differences.

Keywords: Differential privacy · 2020 US Decennial Census · Accuracy

1 Background

1.1 History of Census Disclosure Avoidance Techniques

Using a disclosure avoidance technique based on differential privacy represents a major break from methods used in prior decennial censuses. From 1970 through 2010, the Bureau used a variety of techniques, including whole table suppression (1970–1980), swapping (1990–2010), and blank and impute (1990–2010), to protect the confidentiality of respondents [24]. To implement these methods, the Bureau identified potentially disclosive variables and then found cells with small counts based on those variables. They would then suppress tables with these

Supported by the Minnesota Population Center (R24 HD041023), funded through grants from the Eunice Kennedy Shriver National Institute for Child Health and Human Development.

J. Domingo-Ferrer and K. Muralidhar (Eds.): PSD 2020, LNCS 12276, pp. 353–368, 2020.
https://doi.org/10.1007/978-3-030-57521-2_25

small counts or swap households matched on key demographic characteristics between geographic units.[1]

Traditional disclosure techniques introduced uncertainty into published data. Whole table suppression withheld information about certain aspects of the population. Swapping introduced error into some counts because households would not match on all demographic characteristics. It is impossible to precisely quantify the error introduced by these methods because the swap rates and key characteristics are confidential, but the Census Bureau concluded that "the impact in terms of introducing error into the estimates was much smaller than errors from sampling, non-response, editing, and imputation" [24, p. 11].

1.2 Reconstruction and Re-identification Attack

Concerned about increased computing power, increased availability of individual-level publicly available databases, and the massive volume of statistics published from decennial censuses, the Bureau executed a reconstruction and re-identification attack on the 2010 decennial census [2,17,23]. Working from the database reconstruction theorem [14],[2] the Bureau reconstructed 308,745,538 microdata records using their published census block and tract tabulations. Each reconstructed record had a census block identifier and values for sex, age, race, and Hispanic ethnicity.[3]

The Bureau then linked the reconstructed records to a commercial database by matching on age,[4] sex, and block ID. Forty-five percent (138,935,492) of the reconstructed records shared the same age, sex, and block ID as a record in the commercial database, which also included names. Finally, the Bureau attempted to link these 138 million records to the confidential microdata on all attributes–census block ID, age, sex, race, Hispanic ethnicity, and name. For 52 million persons, the Census Bureau was able to confirm that the two records referred to the same person. In other words, the Bureau reconstructed microdata that allowed it to match of 17% (52 million out of the 309 million) of the population enumerated in the 2010 decennial census to an external source.

The Census Bureau was able to verify the linkage of the reconstructed microdata and the commercial database because they have access to names via the confidential microdata. As Acting Director of the Census Bureau Ron Jarmin pointed out, an external attacker would not have access to such data; thus, they would not know for sure which 17% of the matches were true [19]. Of course, the

[1] Space constraints prevent us from a complete discussion of the Bureau's disclosure avoidance techniques. Interested readers are directed to McKenna [24,25].

[2] The database reconstruction theorem states that respondent privacy is compromised when too many accurate statistics are published from the confidential data. For the 2010 decennial census, more than 150 billion statistics were published [23].

[3] The Bureau reconstructed microdata from a set of 2010 decennial tables. Tables P1, P6, P7, P9, P11, P12, P12A-I, and P14 for census blocks and PCT12A-N for census tracts were used in the reconstruction [23].

[4] The Bureau linked the two datasets by exact age and by age plus or minus one year. [23].

attacker may have access to other datasets for verification purposes, but those datasets will differ from the confidential microdata.

1.3 Differential Privacy

For the 2020 Census, the Bureau has adopted disclosure control methods conforming to differential privacy. Differential privacy is a class of methods for introducing noise into published data [15].[5] Differential privacy is best understood as a description of the privacy-protecting properties of the algorithm used to generate published data, rather than a specific algorithm for implementing disclosure control. While differential privacy guarantees risk is below a certain level, it does not guarantee absolute protection from re-identification for all individuals.[6]

While differential privacy is not a specific algorithm, implementations of differential privacy generally follow a pattern of calculating cross-tabulations from "true" data and injecting noise drawn from a statistical distribution into the cells of the cross-tabulation. To illustrate with a simple example, let's say we asked 100 people about their sex and school attendance status and created the cross-tabulation of sex by school attendance (Table 1, confidential panel). For each of the six cells in the cross-tabulation, we draw a random value from a Laplace distribution with a pre-specified scale and add it to the value of the cell (Table 1, diff. private panel).

Table 1. Confidential and differentially private cross-tabulations from a simple survey.

Sex	Never attended	Attending	Attended in past	Total
Confidential				
Male	3	12	33	48
Female	4	17	31	52
Total	7	29	64	100
Diff. private				
Male	$3 - 1 = \mathbf{2}$	$12 + 0 = \mathbf{12}$	$33 + 1 = \mathbf{34}$	$48 + 0 = \mathbf{48}$
Female	$4 + 8 = \mathbf{12}$	$17 + 2 = \mathbf{19}$	$31 - 2 = \mathbf{29}$	$52 + 8 = \mathbf{60}$
Total	$7 + 7 = \mathbf{14}$	$29 + 2 = \mathbf{31}$	$64 - 1 = \mathbf{63}$	$100 + 8 = \mathbf{108}$

[5] Readers interested in learning more about differential privacy are directed to Wood et al. [32] and Reiter [27]. These papers provide a relatively non-technical introduction to the topic. A critique of differential privacy can be found in Bambauer et al. [5].

[6] The Census Bureau executed a reconstruction and re-identification attack on the 2010 Demonstration Data Product, which was generated from a differentially private algorithm. Approximately 5% of the reconstructed microdata records were successfully matched to confidential data. The 5% re-identification rate represents an improvement over the 17% re-identified from the 2010 decennial census data, but it still represents approximately 15 million census respondents [21].

Three key points from this simple example are worth noting. First, the noise introduced into each cell is independent of the original value of the cell. Therefore, it is possible to introduce relatively large noise values into relatively small cells. In the example, the number of females who had never attended school tripled from four to twelve in the noisy data. Second, introducing noise perturbs not only the values of the cells in the cross-tabulation but the size of the overall population. In the example, our original sample included 100 individuals, but the differentially private data described 108 synthetic records. Finally, though not illustrated in this example, it is possible to introduce noise such that cell values become negative. If a data producer wishes to maintain the total population count and avoid negative values in published data, they must use a post-processing algorithm that enforces total population as an invariant and non-negativity.

2 2010 Demonstration Data Product

The Census Bureau released the 2010 Demonstration Data Product (DDP) in October 2019 to help users examine impacts of the new disclosure avoidance algorithm on decennial census data [9].[7] The DDP consists of two datasets - the Public Law 94-171 Redistricting (PL 94-171) dataset and a partial version of the Demographic and Housing Characteristics (DHC) dataset.[8] Each dataset contains multiple tables for multiple geographic levels. Details about the tables and geographic levels are available in the technical documentation for the 2010 DDP [10].

The Bureau generated the DDP by running the 2010 Census Edited File (CEF)[9] through its Disclosure Avoidance System (DAS). The DAS takes in a data file and a set of parameters detailing the privacy loss budget, its allocation, and invariants and constraints. It then injects noise drawn from particular two-sided geometric distributions into the counts for geographic units. After the noise injection, the DAS uses a "Top-Down Algorithm" developed by Census to construct differentially private tabulations for specified geographic levels that are internally consistent and satisfy other invariants and constraints. Finally, it

[7] The 2010 Demonstration Data Product was the Census Bureau's third dataset produced by the Disclosure Avoidance System (DAS). The DAS consists of the Bureau's differentially private algorithm and the post-processing routines required to enforce constraints. The first dataset contained tabulations from the 2018 Census Test enumeration phase, carried out in Providence County, Rhode Island. The second dataset consists of multiple runs of the DAS over the 1940 complete-count census microdata from IPUMS. Details are available in [3,22].

[8] The Demographic and Housing Characteristics dataset is the replacement for Summary File 1.

[9] All decennial census products, except for Congressional apportionment counts, are derived from the Census Edited File (CEF). The CEF is produced through a series of imputations and allocations that fill in missing data from individual census returns and resolve inconsistencies. Readers interested in a more detailed discussion of the CEF production are directed to pages 10–11 of boyd [6].

generates the Microdata Detail File (MDF), which is differentially private. The Bureau's tabulation system reads in the MDF and constructs the PL94-171 and DHC tables for specified geographic levels. The DDP consists of these PL94-171 and DHC tables.

The 2010 Summary File 1 and PL94-171 datasets were also tabulated from the 2010 CEF and contained the same set of geographic levels and units as the DDP. By publishing the same set of tabulations for the same set of geographic units based on the same input data, the Bureau facilitated comparisons between a dataset produced using traditional disclosure avoidance techniques (Summary File 1 and PL94-171) and one produced using a differentially private algorithm (DDP). Data users could compare statistics derived from both products and determine whether the differentially private data would fit their needs.[10]

2.1 Policy Decisions

Disclosure control algorithms require parameters that control the amount of noise, suppression, or swapping applied to the input data. Values for these parameters have impacts on the quality and accuracy of the output data, and it is critical that data users understand both the significance of the parameters and their possible range of values. This section will discuss the Top-Down Algorithm's parameters and their values that were used to generate the DDP.

Global Privacy Loss Budget. The global privacy-loss budget (PLB), usually denoted by the Greek letter ϵ, establishes the trade-off between the privacy afforded to Census respondents and the accuracy of the published data. Values for ϵ range from essentially 0 to infinity, with 0 representing perfect privacy/no accuracy and infinity representing no privacy/perfect accuracy.[11] Once the global PLB is established, it can then be spent by allocating fractions to particular geographic levels and queries. Geographic levels or queries that receive larger fractions will be more accurate, and levels or queries that receive smaller fractions or no specific allocation will be less accurate.

For the DDP, the Census Bureau's Data Stewardship Executive Policy Committee set the global PLB to 6.0, allocating 4.0 to person tables and 2.0 to household tables. The person and household PLBs were then allocated to combinations of geographic levels and queries [12,16]. The geographic level-query allocations ultimately determine the magnitude of the noise injected into counts.

[10] The National Academies of Sciences, Engineering, and Medicine's Committee on National Statistics (CNStat) hosted a 2-day workshop on December 11–12, 2019. Census Bureau staff members presented details of the algorithm used to create the DDP. Census data users presented results from analyses that compared the 2010 DDP with 2010 Summary File 1 and PL94-171 data products. Privacy experts discussed issues surrounding the decennial census and potential harms of re-identification. Videos and slides from the workshop are available at https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518.

[11] Technically, ϵ must be greater than 0. If ϵ was zero, then no data would be published.

Geographic Levels. If we think of the cross-tabulations into which noise is injected as a set of rows and columns, the geographic levels define the rows. Each row in a cross-tabulation is a geographic unit within a geographic level (e.g., Minnesota is a geographic unit in the State geographic level). Seven geographic levels in the Census Bureau's hierarchy [8] received direct allocations of the PLB. The nation and state levels received 20% each, and the remaining five levels (county, census tract group,[12] census tract, block group, block) received 12% each. These allocations are the same for the person and household tables.

For geographic levels that receive no direct allocation of the PLB, they accumulate accuracy from the units that comprise them. Levels created from census blocks will inherit accuracy of census blocks. Within a particular level, units with larger populations will be more accurate than units with smaller populations.

Queries. If geographic levels define the rows of a cross-tabulation, then queries define the columns. Queries are essentially combinations of demographic or housing variables, and the PLB is allocated to these queries. Queries receiving a larger fraction of the PLB will be more accurate.

The DAS defines two types of queries. "Detailed" queries consist of all unique combinations of variables, and "DP queries" are specific combinations of variables. The "detailed" queries allow the Bureau to reconstruct the underlying microdata, and "DP queries" allow policy makers to select the statistics and relationships that will be more accurate in the published data.

Queries defined in the DAS do not have a one-to-one relationship with the tables published in the 2010 DDP. The queries are used in the noise injection and optimization processes, and the published tables are created from the synthetic microdata created by those processes. Categories in the published tables can and will differ from those used in the queries.

The Census Bureau designed seven queries to support the production of the person tables and six queries to support the production of the household tables in the 2010 DDP. Each query received a direct allocation of the PLB. The *voting age * Hispanic origin * race * citizenship*[13] query received the largest allocation (50%) of the person PLB. The *sex of householder * Hispanic origin of householder * race of householder * household type* and the *Hispanic origin*

[12] The census tract group is not a standard unit in the Census Bureau's geographic hierarchy. It was created specifically for the DAS to control the number of child units for each county. The census tract group consists of all census tracts with the same first four digits of their code (e.g., tract group 1001 consists of tracts 1001.01 and 1001.02). The DDP does not include data for tract groups.

[13] At the time the DAS for the 2010 Demonstration Data Product was designed, the Census Bureau assumed the citizenship question would be included on the 2020 Decennial Census questionnaire. Even though the US Supreme Court ruled in favor of the plaintiffs and removed the question, the Bureau did not have time to remove the citizenship variable from the DAS. No actual citizenship data was used to create the 2010 DDP; instead, the Bureau imputed citizenship status for records in the CEF [10].

*of householder * race of householder * household size * household type* queries received the largest allocations (25% each) of the household PLB. A list of all DDP queries and their allocations can be found in Appendix A.

Invariants and Constraints. Invariants and constraints play key roles in the DAS, particularly in post-processing routines applied to the noisy counts. Invariants are counts computed directly from the CEF into which no noise is injected. Constraints control the types and range of values in the final tabulations.

While the Bureau has not finalized the invariants for the 2020 Decennial Census, they did establish four invariants for the DDP. Total population is invariant at the state-level, and total housing units, total group quarters facilities, and total group quarters facilities by type are invariant at the census block-level.[14] These same four counts were invariant at the *census block-level* in the 2010 Decennial Census. Additionally, voting age population and occupied housing units (i.e., households) were invariant at the census block-level [1].

Constraints are the set of rules that the data produced by the DAS must follow. For the DDP, constraints included non-negativity, integer, and hierarchical consistency constraints. The non-negativity and integer constraints require all counts to be positive integer values. The hierarchical consistency constraint imposes consistency among geographic and attribute hierarchies. For geographic hierarchies, counts for child units must sum to the counts of the parent unit. For attribute hierarchies, counts for child attributes must sum to the counts of the parent attribute. If we sum the counts of occupied and vacant housing units for a given geographic unit, the sum must equal the published total housing unit count for the same unit.

2.2 2010 DDP Disclosure Avoidance System (DAS)

The DAS that generated the 2010 DDP consists of three steps: generating counts from the CEF, injecting noise, and post-processing to satisfy constraints.

Generating Counts. The first step in the DAS produces counts from the CEF. The DAS consumes the CEF, the queries, and the geographic levels and creates a set of histograms - one for each combination of query and geographic level. The cells each histogram contain the counts of a particular set of categories (e.g, 40 year-old females) for a given geographic unit. The number of cells in these histograms may be massive, particularly at the census block level, and the counts in the cells may be small or even zero.

Noise Injection. The second step in the DAS injects noise into the cell counts generated in the first step. These "noisy counts" are, by definition, differentially private, but they may not satisfy invariants or constraints set by policy. The noise injection step is accomplished through the following sub-steps.

[14] The geographic level associated with an invariant is the lowest level at which the invariant holds. All geographic levels composed of the lowest level will also be invariant.

*Compute ϵ for Each Geographic Level * Query Combination.* Policy decisions establish the global privacy loss budget and the fractional allocations of that PLB to specific geographic levels and queries. The DAS computes the ϵ value for each histogram as the product of the corresponding query and geographic allocations from the PLB.

Compute the Scale Parameter for the Statistical Distribution. Noise-injection values are generated by randomly drawing a value from a statistical distribution. The shape of the distribution is controlled by the scale parameter calculated in Eq. (1).

$$\frac{2}{\epsilon} = s \tag{1}$$

For this sub-step, ϵ is the histogram-specific value computed in the previous sub-step. The numerator is the sensitivity of the query, which is always 2 for histograms [13].[15] Scale parameters are shown in Table 2. Nation and state parameters are in column $Scale_{nation}$, and parameters for the other five geographic levels are in column $Scale_{county}$. Larger scale parameters represent distributions with higher variances, which yield potentially larger noise values.

Generate and Inject Random Noise into Each Histogram Cell. The final step in the noise injection process is to actually draw random values from the statistical distribution for a given *geographic level * query* combination and inject those values into the histogram cells computed from the CEF in Step 1. The scale parameter computed in the previous step determines the shape of a particular distribution.[16]

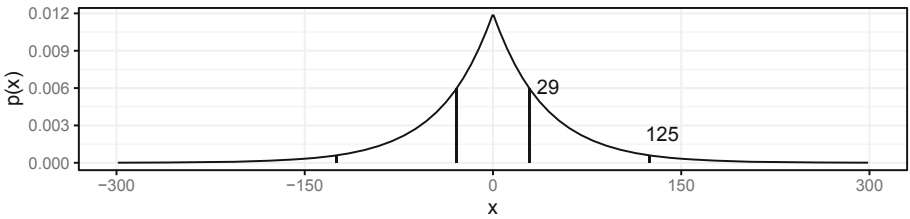

Fig. 1. Noise distributions used for the detailed histogram at the county, tract group, tract, block group, and block geographic level. The scale parameter for this distribution is 41.67.

[15] Sensitivity is the value by which a query changes if we make a single modification to the database. Histogram queries have a sensitivity of 2 - if we increase the count in a cell by 1, we must decrease the count in another cell by 1.

[16] Two types of distributions - the two-tailed geometric and the Laplace - are typically used to achieve differential privacy. The two-tailed geometric distribution is used when integers are required, and the Laplace distribution is used when real numbers are required. Source code for the 2010 DDP includes functions for both types of distributions [11].

Figure 1 depicts the Laplace distributions used for the detailed histogram at the county, tract group, tract, block group and block geographic levels. The labeled vertical lines illustrate the 75th and 97.5th percentiles. Fifty percent of all random draws will fall between 29 and −29, and 95% of all random draws will range from 125 to −125.

Post-processing. The output of step two is a series of noisy, differentially private, histograms, one for each *geographic level * query* combination. The raw noise-injected histograms are not integer-valued and may include negative values. Furthermore, because noise is injected independently into each histogram, they are inconsistent with each other, both within and across geographic levels. The histograms also do not satisfy invariants such as total population. Finally, the set of queries and geographic levels used for noise injection does not match the set of cross-tabulations and geographic levels desired for publication. In order to produce the final dataset, the Census conducts a series of optimization steps that ultimately generate synthetic microdata, which can then be tabulated to create tables for publication.

The Census refers to their post-processing algorithm as a "Top-Down Algorithm" because it starts at the national level and works down the geographic hierarchy, successively generating data for finer geographic levels. A diagram depicting the flow of data through noise injection and optimization can be found in Appendix B.

Generate National-Level Detailed Histogram. The first post-processing step produces an optimized version of the national-level detailed histogram. The detailed histogram is essentially a cross-tabulation of all possible categories of all the variables, which fully defines synthetic microdata.

The national-level detailed histogram is generated by solving an optimization problem to minimize the differences between the detailed histogram and the set of noisy histograms [22].[17] The optimization problem includes constraints to enforce invariants, non-negativity, integer values, and implied constraints to maintain consistency between person and household tables.

Generate Detailed Histograms for Lower Geographic Levels. At each subsequent geographic level, a similar optimization problem is solved, minimizing the differences between that level's set of noisy histograms and the output detailed histogram for that level, while matching invariants and meeting other constraints. For these lower geographic levels, the optimization problem also includes an additional constraint in the form of the detailed histogram from the parent level. Effectively, the lower-level optimization problems assign geographic identifiers to the synthetic microdata defined by the national-level detailed histogram in such a way as to best match the noisy lower-level queries.

[17] The optimization problem is actually solved in two stages, one to enforce non-negativity and optimize over the set of queries and a second stage to produce an integer-valued detailed histogram.

Generate Final Synthetic Microdata and Cross-Tabulations for Publication. The final detailed histogram is at the block level. This histogram is then transformed into synthetic microdata records, each with a full set of characteristics and a block identifier, constituting the Microdata Detail File (MDF). The Bureau's tabulation system then reads in the MDF and constructs the PL94-171 and DHC tables for specified geographic levels.

3 Data Usability Insights from the DDP

Publication of the 2010 Demonstration Data product allowed data users to compare statistics from the DDP with those from Summary File 1 (SF1), which was produced using traditional disclosure avoidance techniques. In doing so, users discovered several problematic aspects with respect to the accuracy of the DDP [4,20,26,28–31]. Many users have concluded that if 2020 decennial census data were published based on the DAS as implemented in producing the DDP it would be unusable for their needs. Below, we explore several of the issues that seem to contribute to these data usability problems by examining the particularly surprising inaccuracies in occupancy rates in the DDP.[18]

The DDP data showed 310 of the 3,221 counties in the United States and Puerto Rico as having occupancy rates of 100% (i.e., no vacant housing units).[19] In the SF1 data, no counties had 100% occupancy rates. These discrepancies represent known error in the DDP data because counts of both total housing units and occupied housing units were invariant to the block level in SF1. These errors can be traced to particular aspects of the DAS algorithm design and policy decisions, including the design of the household queries and allocation of the PLB among them and the invariants and constraints applied during the optimization process.

3.1 DDP Calculation of Occupancy Rates

First, it is important to understand how occupancy rates were calculated for the DDP. The Census Bureau's Data Stewardship Executive Policy (DSEP) committee made two critical recommendations that impacted the occupied housing unit counts: (1) the count of housing units be invariant for all geographic levels in the census hierarchy; (2) the count of occupied housing units will be subject to disclosure avoidance [16]. Housing unit characteristics were not directly included in the DAS queries, but the count of occupied housing units is, by definition, equal to the count of households. The count of vacant housing units is the difference between total housing units and households.

[18] The Census Bureau fielded so many questions about occupancy rates that they added a question to their FAQ [7]. The answer mentions that Census would look into the issue and post answers or updates. As of 2020-05-22, no answers or updates have been posted.

[19] Readers interested in learning more about the discrepancy should watch Beth Jarosz' presentation at the December 2019 CNStat workshop on the 2010 Demonstration Data Product [20].

3.2 Household Query Design and PLB Allocation

The combination of the design of the household queries and PLB allocations is likely to have resulted in a low signal-to-noise ratio in the noisy household histograms. Counts of households are, by definition, equal to or less than counts of persons. (In 2010 the national average household size was 2.68.) Except for the detailed histogram, the household histograms generally include many more cells than the person histograms. This means that CEF counts in many of the household histogram cells are quite small. In the SF1 data, the median count of households in counties is approximately 10,000. As an example, if 10,000 households were distributed evenly over the 384 cells in the *sex of householder * household type * elderly* histogram, each cell would have a count of about 26. Of course, most households will be in the *male-no elderly* cells, so other cells will have even smaller counts. With a scale parameter of 83.33, 50% of the draws of noise from the Laplace distribution for this histogram are expected to have absolute values greater than 57, easily doubling the original count or dropping it below zero. The situation is similar or worse for the other household histograms, meaning the information content of each histogram is largely obscured by the noise injected.

The set of household queries and allocation of PLB over the queries further contributes to the low signal-to-noise ratio. Overall, the household queries received an initial PLB allocation of 2.0, compared to the person queries' allocation of 4.0. No single household query received more than a 0.25 fractional allocation of the 2.0 PLB. This means that none of the counts or relationships are particularly well preserved. Furthermore, most of the household variables appear in just one or two of the queries (see Appendix A for details). Variables that appear in multiple queries provide redundancy in the set of histograms passed into the optimization problem that helps to pull the solution back toward the true values. Without this redundancy, the original signal tends to get lost in the noise.

3.3 Optimization and Constraints

The non-negativity constraint and block-level total housing unit invariant used as constraints in the DAS optimization ultimately result in the observed occupancy rate errors. The non-negativity constraint requires that every cell in the final detailed histogram be non-negative. As described above, many of the cells in the noisy household histograms will be negative, especially for geographic units with smaller numbers of households. Returning these cells to zero effectively adds households to these small places, resulting in positive bias. Dividing counties into quintiles by SF1 household count, counties in the three lowest quintiles consistently have more households in the DDP data than in SF1 (Fig. 2).

The invariant number of housing units down to the block level implies an upper-bound constraint on the number of households. Each geographic unit must have no more households than it has housing units. With the low signal-to-noise ratio in the noisy histograms, especially at the block level, this constraint is the strongest signal present in the optimization problem. Many geographic units therefore receive a number of households equal to the number of housing units,

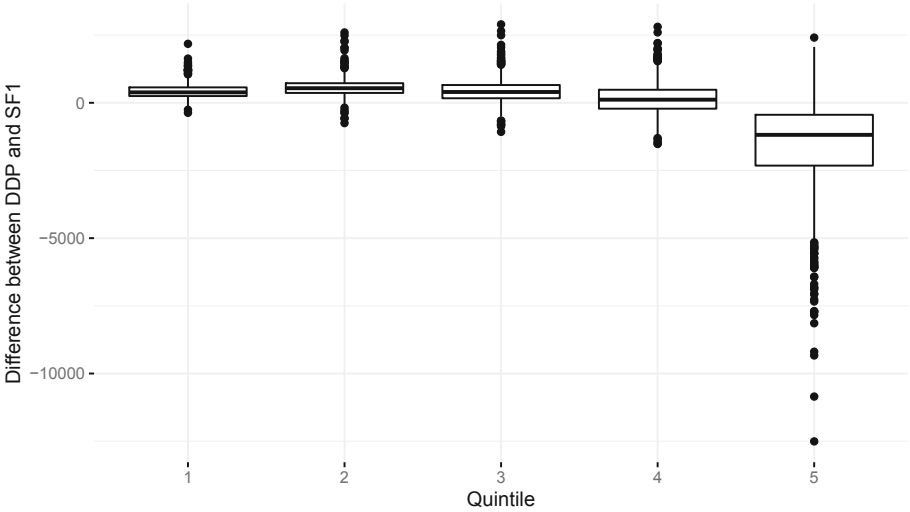

Fig. 2. Boxplots of county-level differences in household counts between DDP and SF1. Quintiles are computed from the invariant SF1 counts.

resulting in 100% occupancy rates. This is especially true for geographic units with smaller numbers of households that are affected by positive bias due to the non-negativity constraint.

While this combination of factors is especially problematic for the occupancy rate results in the DDP, the issues are not limited to this particular case. Researchers analyzing the DDP data in comparison to the SF1 data have found evidence of related issues in many aspects of the data. The issue of scale-independent noise affects all of the millions of cells with small counts in both the person and household histograms, making counts of many population subsets unreliable. The combination of the non-negativity constraint and population invariants consistently leads to bias increasing counts of small subgroups and small geographic units and decreasing counts of larger subgroups and geographic units.

4 Conclusion

Adopting a disclosure control technique based on differential privacy "marks a sea change for the way that official statistics are produced and published" [18]. It is critical that data users understand this new technique so they can judge whether the published data are fit to use. As the occupancy rate example demonstrated, the algorithm that generated the 2010 Demonstration Data Product Data produced highly problematic and biased data. Users must also be aware of the policy decisions required by the technique so that they may participate effectively in the decision-making process.

A Privacy Loss Budget Allocations

Table 2 lists the 7 person and 6 household queries that received direct allocations of the privacy loss budget. The allocations are shown in the PLB_{frac} column. The bold rows are the queries with the largest PLB allocation.

The $Scale_{nation}$ and $Scale_{county}$ columns list the scale factors used to generate the statistical distributions from which noise injection values are drawn. The $Scale_{nation}$ values are used for the nation and state histograms, and the $Scale_{county}$ values are used for the county, tract group, tract, block group, and block histograms.

The $Hist_{size}$ column lists the number of cells in the particular query. This is the number of cells on each row of the histogram (i.e., for each geographic unit). The value is generated by multiplying together the number of categories for each variable in a query. For example, the *Sex * Age (64 year bins)* query has two categories for *sex* and two categories for *age* giving a histogram size of 4. Category counts for each variable are listed in frequently asked question 11 in [7].

Table 2. Privacy loss budget allocations and scale parameters for 2010 DDP queries.

Query	$Hist_{size}$	PLB_{frac}	$Scale_{nation}$	$Scale_{county}$
Person				
Detailed person	467,712	0.10	25.0	41.67
Household/group quarters	8	0.20	12.5	20.83
Voting age * Hisp * Race * Citizen	**504**	**0.50**	**5.0**	**8.33**
Sex * Age (single year bins)	232	0.05	50.0	83.33
Sex * Age (4 year bins)	58	0.05	50.0	83.33
Sex * Age (16 year bins)	16	0.05	50.0	83.33
Sex * Age (64 year bins)	4	0.05	50.0	83.33
Household				
Detailed household	96,768	0.20	25.0	41.67
$\mathbf{HH_{Hisp}}$ * $\mathbf{HH_{race}}$ * $\mathbf{HH_{size}}$ * $\mathbf{HH_{type}}$	**2,688**	**0.25**	**20.0**	**33.33**
$\mathbf{HH_{sex}}$ * $\mathbf{HH_{Hisp}}$ * $\mathbf{HH_{race}}$ * $\mathbf{HH_{type}}$	**672**	**0.25**	**20.0**	**33.33**
HH_{Hisp} * HH_{type} * HH_{multi}	28	0.10	50.0	83.33
HH_{sex} * HH_{type} * $HH_{elderly}$	384	0.10	50.0	83.33
HH_{sex} * HH_{age} * HH_{type}	432	0.10	50.0	83.33

The variable names in the household queries are described in Table 3.

Table 3. Variable names and descriptions.

Variable name	Variable description
HH_{sex}	Sex of householder
HH_{race}	Race of householder
HH_{Hisp}	Hispanic/Latino origin of householder
HH_{size}	Household size
HH_{type}	Houehold type
HH_{multi}	Presence of three or more generations in household
$HH_{elderly}$	Presence of persons age 60+, 65+, or 75+

B Top-Down Algorithm flow diagram

Figure 3 depicts the flow of data through the noise injection and optimization steps of the Census Bureau's Top-Down Algorithm.

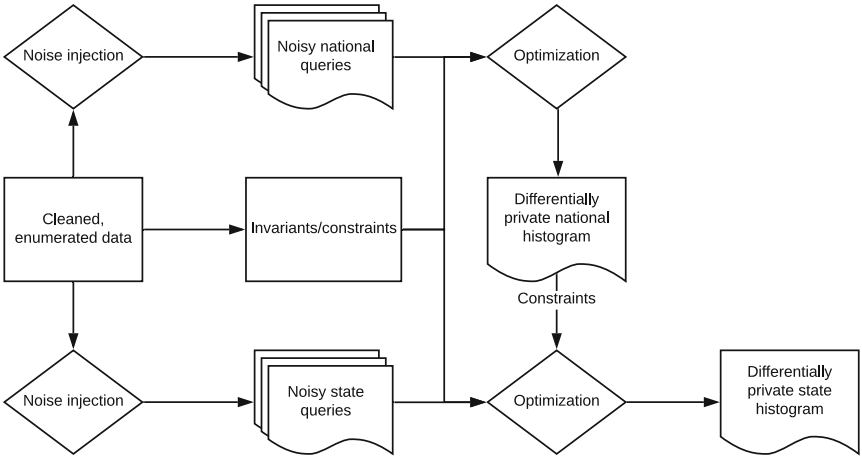

Fig. 3. Census DAS optimization flow diagram.

References

1. Abowd, J.: Disclosure avoidance for block level data and protection of confidentiality in public tabulations, December 2018. https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf

2. Abowd, J.: Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau (2018). https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html
3. Abowd, J., Garfinkel, S.: Disclosure Avoidance and the 2018 Census Test: Release of the Source Code (2019). https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html
4. Akee, R.: Population counts on American Indian Reservations and Alaska Native Villages, with and without the application of differential privacy. In: Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, Washington, DC, December 2019
5. Bambauer, J., Muralidhar, K., Sarathy, R.: Fool's gold: an illustrated critique of differential privacy. Vanderbilt J. Entertain. Technol. Law **16**, 55 (2014)
6. Boyd, D.: Balancing data utility and confidentiality in the 2020 US census. Technical report, Data and Society, New York, NY, December 2019
7. Census Bureau: Frequently Asked Questions for the 2010 Demonstration Data Products. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products/faqs.html
8. Census Bureau: Standard hierarchy of census geographic entities. Technical report, US Census Bureau, Washington DC, July 2010
9. Census Bureau: 2010 Demonstration Data Products (2019). https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html
10. Census Bureau: 2010 demonstration P.L. 94–171 redistricting summary file and demographic and housing demonstration file: technical documentation. Technical report, US Department of Commerce, Washington, DC, October 2019
11. Census Bureau: 2020 Census 2010 Demonstration Data Products Disclosure Avoidance System. US Census Bureau (2019)
12. Census Bureau: 2020 census 2010 demonstration data products disclosure avoidance system: design specification, version 1.4. Technical report, US Census Bureau, Washington DC (2019)
13. Cormode, G.: Building blocks of privacy: differentially private mechanisms. Technical report, Rutgers University, Brunswick, NJ (nd)
14. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. PODS 2003, pp. 202–210. ACM, New York (2003). https://doi.org/10.1145/773153.773173
15. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
16. Fontenot Jr., A.: 2010 demonstration data products - design parameters and global privacy-loss budget. Technical report 2019.25, US Census Bureau, Washington, DC, October 2019
17. Garfinkel, S., Abowd, J.M., Martindale, C.: Understanding database reconstruction attacks on public data. ACM Queue **16**(5), 28–53 (2018)
18. Garfinkel, S.L., Abowd, J.M., Powazek, S.: Issues encountered deploying differential privacy. In: Proceedings of the 2018 Workshop on Privacy in the Electronic Society. WPES 2018, pp. 133–137. ACM, New York (2018). https://doi.org/10.1145/3267323.3268949

19. Jarmin, R.: Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census (2019). https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census_bureau_adopts.html
20. Jarosz, B.: Importance of decennial census for regional planning in California. In: Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, Washington, DC, December 2019
21. Leclerc, P.: The 2020 decennial census topdown disclosure limitation algorithm: a report on the current state of the privacy loss-accuracy trade-off. In: Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, Washington DC, December 2019
22. Leclerc, P.: Guide to the census 2018 end-to-end test disclosure avoidance algorithm and implementation. Technical report, US Census Bureau, Washington DC, July 2019
23. Leclerc, P.: Reconstruction of person level data from data presented in multiple tables. In: Challenges and New Approaches for Protecting Privacy in Federal Statistical Programs: A Workshop, Washington, DC, June 2019
24. McKenna, L.: Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing. Technical report 18–47, US Census Bureau, Washington, DC (2018)
25. McKenna, L.: Disclosure avoidance techniques used for the 1960 through 2010 decennial censuses of population and housing public use microdata samples. Technical report, US Census Bureau, Washington, DC (2019)
26. Nagle, N.: Implications for municipalities and school enrollment statistics. In: Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, Washington, DC, December 2019
27. Reiter, J.P.: Differential privacy and federal data releases. Ann. Rev. Stat. Appl. **6**(1), 85–101 (2019). https://doi.org/10.1146/annurev-statistics-030718-105142
28. Sandberg, E.: Privatized data for Alaska communities. In: Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, Washington, DC, December 2019
29. Santos-Lozada, A.: Differential privacy and mortality rates in the United States. In: Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, Washington, DC, December 2019
30. Santos-Lozada, A.R., Howard, J.T., Verdery, A.M.: How differential privacy will affect our understanding of health disparities in the United States. Proc. Natl. Acad. Sci. (2020). https://doi.org/10.1073/pnas.2003714117
31. Spielman, S., Van Riper, D.: Geographic review of differentially private demonstration data. In: Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, Washington, DC, December 2019
32. Wood, A., et al.: Differential privacy: a primer for a non-technical audience. Vanderbilt J. Entertain. Technol. Law **21**(1), 209–276 (2018)