Some Lessons from the 2020 U.S. Census Disclosure Avoidance System

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau

Computing Community Consortium, INFORMS, ACM SIGAI

Artificial Intelligence/Operations Research Workshop II

Panel C: Robustness/Privacy, Tuesday, August 16, 2022, 3:30pm

United States® Census Bureau

The views expressed in this talk are my own and not those of the U.S. Census Bureau. DMS Project ID: P-7502798. DRB Clearance numbers: CBDRB-FY20-DSEP-001, CBDRB-FY22-DSEP-003, CBDRB-FY22-DSEP-004.

Acknowledgements

Co-authors on the HDSR paper: Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, Pavel Zhuravlev
The 2020 Census Disclosure Avoidance System TopDown Algorithm · Special Issue 2: Differential Privacy for the 2020 U.S. Census (mit.edu)

Annual Review of Statistics paper (with Michael Hawes): [2206.03524] Confidentiality Protection in the 2020 US Census of Population and Housing (arxiv.org)

This presentation also includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators from the following Census Bureau divisions and outside organizations: ADCOM, ADDC, ADRM, CED, CEDDA, CEDSCI, CES, CSRM, DCMD, DITD, ESMD, GEO, POP, TAB, CDF, Econometrica Inc., Galois, Knexus Research Corp, MITRE, NLT, TI, and Tumult Labs.

I also acknowledge and greatly appreciate the ongoing feedback we have received from external stakeholder groups that has contributed to the design and improvement of the Disclosure Avoidance System.

United States® Census Bureau

Bottom Line Up Front:

Going from suppression to differential privacy is much easier than going from publishing all the microdata to differential privacy.

United States® Census Bureau

Translation:

2020 Census data clients had accuracy expectations that modern privacy protection can't support (2010 Census basically released all the microdata, although not intentionally).

United States® Census Bureau

Forecast:

AI applications, particularly in industry, are going to face the same conundrum. Advertising executives are not going to like the privacy-protected models. (Conventional AI applications are inherently disclosive.)

United States® Census Bureau

Start | Base Map | Selection | Results

Distance/Direction Analysis

Work to Home

Display Settings

Labor Market Segment Filter: All Workers

Year: 2019

Map Controls

Color Key

Thermal Overlay

Point Overlay

Selection Outline

Identify | Zoom to Selection

Clear Overlays | Animate Overlays

Report/Map Outputs

Detailed Report

Export Geography

Print Chart/Map

Legends

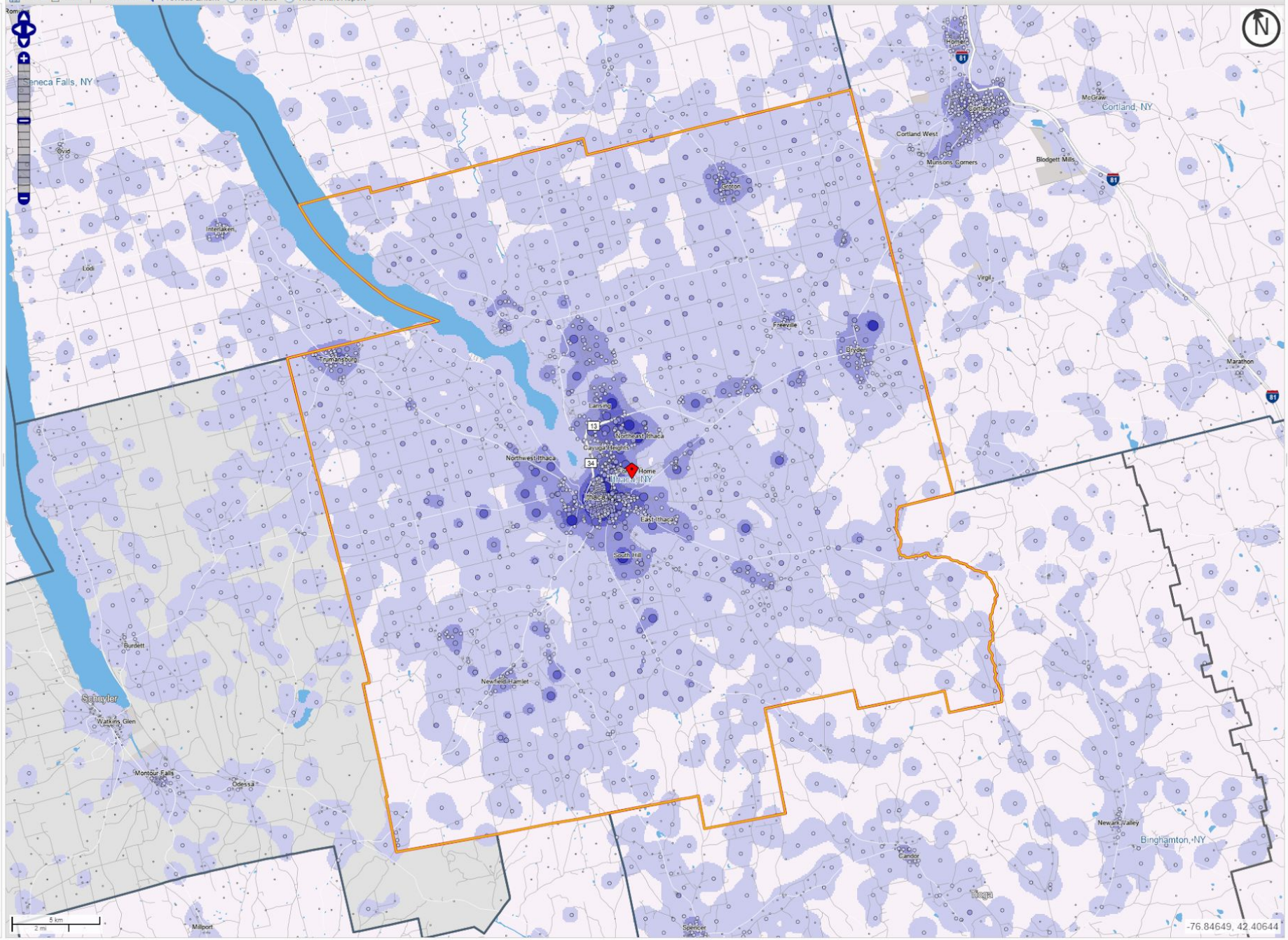
- 5 - 97 Jobs/Sq.Mile
- 98 - 376 Jobs/Sq.Mile
- 377 - 840 Jobs/Sq.Mile
- 841 - 1,490 Jobs/Sq.Mile
- 1,491 - 2,326 Jobs/Sq.Mile
- 1 - 3 Jobs
- 4 - 19 Jobs
- 20 - 64 Jobs
- 65 - 152 Jobs
- 153 - 296 Jobs
- Analysis Selection

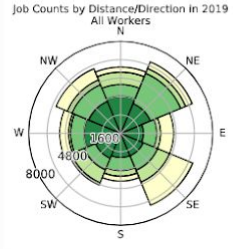
Analysis Settings

Analysis Type	Distance/Direction
Selection area as	Work
Year(s)	2019
Job Type	Primary Jobs
Selection Area	Ithaca, NY from Metropolitan/Micropolitan Areas (CBSA)
Selected Census Blocks	3,082
Analysis Generation Date	04/28/2022 04:50 - OnTheMap 6.8.1
Code Revision	f9358819d46a60bb8905203651 6a1c8fe8bbbeac
LODES Data Version	20211018_1647

Change Settings

Save | Load | Feedback | Previous Extent | Hide Tabs | Hide Chart/Report





View as Radar Chart

Jobs by Distance - Work Census Block to Home Census Block

	2019 Count	Share
Total Primary Jobs	44,266	100.0%
Less than 10 miles	20,790	47.0%
10 to 24 miles	11,346	25.6%
25 to 50 miles	5,330	12.0%
Greater than 50 miles	6,800	15.4%

United States Census Bureau

PSEO

Post-Secondary Employment Outcomes Explorer

What is PSEO? | Download Data | Email Us | Tutorial | Help

Data Type
Earnings | Flows

State
New York

Institution
Binghamton University

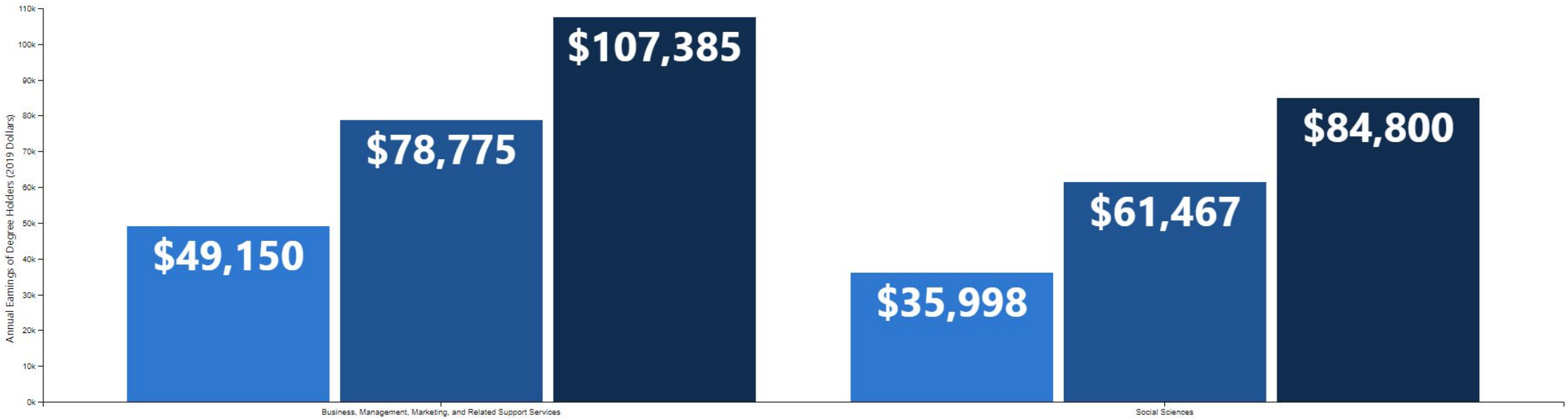
Degree Level
Baccalaureate

Graduation Cohort
All Cohorts

Compare Earnings by
Time | Percent

Earnings-percentile Filter
25 | 50 | 75

Legend
1 Years Postgrad
5 Years Postgrad
10 Years Postgrad



Program Specificity
General | Detailed

Sort Program List by
Size | Name

Clear Selected

Filter Degrees...

All Instructional Programs	Social Sciences	Business, Management, Marketing, and Related Support Services	Psychology	Engineering
Health Professions and Related Programs	Biological and Biomedical Sciences	English Language and Literature/Letters	Computer and Information Sciences and Support Services	Mathematics and Statistics
Visual and Performing Arts	History	Foreign Languages, Literatures, and Linguistics	Philosophy and Religious Studies	Physical Sciences
Natural Resources and Conservation	Area, Ethnic, Cultural, Gender, and Group Studies	Multi/Interdisciplinary Studies		

Some cells are not available due to suppression.

Source: U.S. Census Bureau, Center for Economic Studies, LEHD | Email: CES.PSEO.Feedback@census.gov

United States® Census Bureau

Major data products from the 2020 Census:

- Apportion the House of Representatives
  (April 26, 2021)
- Supply data to all state redistricting offices
  (August 12, 2021)
- Demographic and housing characteristics
  (May 2023)
- Detailed demographic and housing characteristics
  (Part A August 2023; Part B TBD; Supplemental DHC TBD)
- American Indian, Alaska Native, Native Hawaiian data
  (Included in Part A Detailed DHC; August 2023)

For the 2010 Census, this was *more than 150 billion* statistics from 15GB total data.

United States® Census Bureau

Reconstructing the 2010 Census-I

- The 2010 Census collected information on the age, sex, race, ethnicity, and relationship (to householder) status for 308,745,538 million individuals. (about 1.5 billion confidential data points; Garfinkel et al. 2019)
- The 2010 Census data products released over 150 billion statistics
- Internal Census Bureau research confirms that the confidential 2010 Census microdata can be accurately reconstructed from the publicly released tabulations
- This means that all the tabulation variables for 100% of the person records on the confidential data file can be accurately reproduced from the published tabulations
- Based on Dinur and Nissim (2003) and Dwork and Yekhanin (2008)
- *A violation of the 2010 Census contemporaneous disclosure avoidance standards for 2010 Census microdata files*

United States® Census Bureau

Reconstructing the 2010 Census-II

- A violation of the 2010 Census contemporaneous disclosure avoidance standards for 2010 Census microdata
  - The reconstructed microdata are not a sample; there is one record for every person enumerated in the 2010 Census, and the geographic identifier on that record is always correct (matches the geographic identifier on the confidential input file—the Hundred-percent Detail File, which was swapped)
  - The reconstructed microdata have geography identifiers with an average population of 29 (50, if only occupied blocks are counted)
  - The reconstructed microdata have U.S.-level demographic cells (race, ethnicity) with fewer than 10,000 persons
- The standards for releasing microdata from the 2010 Census required (McKenna, 2019)
  - Sample (10% rate was used)
  - Restrict geographic identifiers to areas with at least 100,000 persons (Public-use Microdata Areas)
  - Collapse demographic categories until the national population in 1-way marginals contains at least 10,000 persons
  - The standards for tabular data permitted universe files, block geography, and low-U.S. population demographic groups (McKenna, 2018) on the assumption that microdata reconstruction was infeasible
- These are the reason the Data Stewardship Executive Policy Committee instructed the 2020 Census not to use swapping as the main protection for 2020 Census products from the reconstruction evidence alone: swapping plus aggregation did not protect the 2010 Census confidential microdata properly

United States® Census Bureau

Reconstructing the 2010 Census: What did we find?

Table 1 Agreement Rates (Reconstruction to CEF) by Block Size

Block Size	Total	1-9	10-49	50-99	100-249	250-499	500-999	1,000+
Agreement	91.8%	74.0%	93.0%	93.1%	92.1%	91.3%	90.6%	91.5%

DRB clearance number CBDRB-FY22-DSEP-004; Source: Hawes (2022).

- Block, sex, age (exact/binned in 38 categories), race (OMB 63 categories), and ethnicity were reconstructed:
  - Exactly for 91.8% of the population
  - Exactly 74.0% in the smallest population blocks, but 93.0% in blocks with 10-49 people and 93.1% in blocks with 50-99 people
  - An external user can confirm that these solutions correspond to the exact record in the confidential data for 65% of all blocks using only the published data because there is provably one and only one reconstruction possible in these blocks. That user can identify population uniques on any combination of reconstructed variables.

United States® Census Bureau

This is one of the principal failures of the 2010 tabular disclosure avoidance methodology — swapping provided protection for households deemed "at risk," primarily those in blocks with small populations, whereas for the for the entire 2010 Census 57% of the persons are population uniques on the basis of block, sex, age (in years), race (OMB 63 categories), and ethnicity. Furthermore, 44% are population uniques on block, age and sex. Aggregation provided no additional protection for most blocks.

United States® Census Bureau

Table 5
Distribution of Population and Population Uniques by Block Population Size

Block Population Bin	Number of Blocks in Bin	2010 Census Population in Bin	Cumulative Population	Percent of Population in Bin	Cumulative Percent of Population	Population Uniques (block, sex, age) in Bin	Percent of (block, sex, age) Uniques in Bin
TOTAL	11,078,297	308,745,538				135,432,888	43.87%
0	4,871,270	0	0	0.00%	0.00%		
1-9	1,823,665	8,069,681	8,069,681	2.61%	2.61%	7,670,927	95.06%
10-49	2,671,753	67,597,683	75,667,364	21.89%	24.51%	53,435,603	79.05%
50-99	994,513	69,073,496	144,740,860	22.37%	46.88%	40,561,372	58.72%
100-249	540,455	80,020,916	224,761,776	25.92%	72.80%	27,258,556	34.06%
250-499	126,344	42,911,477	267,673,253	13.90%	86.70%	5,297,867	12.35%
500-999	40,492	27,028,992	294,702,245	8.75%	95.45%	1,051,924	3.89%
1000+	9,805	14,043,293	308,745,538	4.55%	100.00%	156,639	1.12%

DRB clearance number CBDRB-FY21-DSEP-003

United States® Census Bureau

Table 2 Reidentification rates for population uniques

Match file	Universe	Putative rate[a]	Confirmed rate[b]	Precision[c]
Commercial	All data defined persons[d]	60.2%	24.8%	41.2%
	Population uniques[e]	23.1%	21.8%	94.6%
CEF	All data defined persons[d]	97.0%	75.5%	77.8%
	Population uniques[e]	93.1%	87.2%	93.6%

[a]The number of records that agree on block, sex, and age (exact/binned), divided by the total number of records in the universe. [b]The number of records that agree on PIK (the Census Bureau's internal person identifier), block, sex, age (exact/binned), race, and ethnicity, divided by the total number of records in the universe. [c]The number of confirmed reidentifications [records that agree on PIK, block, sex, age (exact/binned), race, and ethnicity] divided by the number of putative reidentifications [records that agree on block, sex, and age (exact/binned)]. [d]All individuals with a unique PIK identifier within the block (276 million persons for the 2010 Census). [e]All data defined individuals who are unique in their block on sex and exact/binned age. DRB clearance number CBDRB-FY22-DSEP-004; Data are from Abowd et al. (under review) released in Hawes (2022). Abbreviations: CEF, Census Edited File; DRB, Disclosure Review Board; PIK, Protected Identification Key.

United States® Census Bureau

Table 3 Reidentification rates for population uniques of the block's modal and non-modal races

Match file	Population uniques[a]	Putative rate	Confirmed rate	Precision
Commercial	All population uniques	23.1%	21.8%	94.6%
	Of the modal race	25.3%	24.2%	95.3%
	Of the nonmodal races	13.7%	12.2%	89.2%
CEF	All population uniques	93.1%	87.2%	93.6%
	Of the modal race	94.8%	91.3%	96.3%
	Of the nonmodal races	86.2%	70.2%	81.5%

[a]Individuals who are unique in their block on sex and exact/binned age. DRB clearance number CBDRB-FY22-DSEP-004. Data are from Abowd et al. (under review) released in Hawes (2022).
Abbreviations: CEF, Census Edited File; DRB, Disclosure Review Board.

- *This is not a statistical use, and both the Census Act (13 U.S. Code §§ 8(b) & 9) and CIPSEA (44 U.S. Code § 3561(11) 'Statistical Purpose') clearly prohibit releasing data that support not-statistical uses.*

United States® Census Bureau

All 2020 Census Publications

- Will all be processed by a collection of differentially private algorithms (Dwork et al. 2006a, 2006b; Dwork 2006) using the zero-Concentrated DP privacy-loss accounting framework (Bun and Steinke 2016) implemented with the discrete Gaussian mechanism (Canonne et al. 2020, 2021)
- Using a total privacy-loss budget set as policy, not hard-wired, determined by the Data Stewardship Executive Policy Committee
- Production code base, technical documents, and extensive demonstration products based on the 2010 Census confidential data have all been released to the public
- More information: https://www.census.gov/newsroom/blogs/research-matters/2019/10/balancing_privacyan.html

United States® Census Bureau

TopDown Algorithm System Requirements

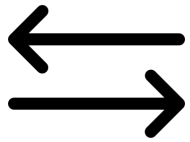
- The 2020 Disclosure Avoidance System's TopDown Algorithm (TDA) implemented formal privacy protections for the P. L. 94-171 Redistricting Data Summary File
- Planned for use in the Demographic Profiles, Demographic and Housing Characteristics (DHC), and Special Tabulations of the 2020 Census
- TDA system requirements include:
  - Input/output specifications
  - Invariants
  - Edit constraints and structural zeros
  - Tunable utility/accuracy for pre-specified tabulations
  - Privacy-loss budget asymptotic consistency
  - Transparency

United States® Census Bureau

What is a histogram?

Record ID	Block	Race	…	Sex
1	1001	Black	…	Male
2	1001	Black	…	Male
3	1001	Asian	…	Female
4	1001	Asian	…	Female
5	1001	Black	…	Male
6	1001	AIAN	…	Female
7	1001	AIAN	…	Male
8	1001	Black	…	Female
9	1001	Black	…	Female

Microdata: One record per respondent



Attribute Combination (Block/Race/…/Sex)	# of Records
1001/AIAN/…/Male	1
1001/AIAN/…/Female	1
1001/Asian/…/Male	0
1001/Asian/…/Female	2
1001/Black/…/Male	3
1001/Black/…/Female	2
…	…

Histogram: Record count for each unique combination of attributes (including location), equivalent to the fully saturated contingency table, vectorized, and with structural zeros removed or imposed by constraint

United States® Census Bureau

Noisy Measurements

- TDA allocates shares of the total privacy-loss budget by geographic level and by query
- Each query of the confidential data will have noise added to its answer
- The noise is taken from a probability distribution with mean=0, and variance determined by the share of the privacy-loss budget allocated to that query at that geographic level
- These noisy measurements are independent of each other, and can include negative values, hence the need for post-processing

United States® Census Bureau

19

Zero-Concentrated Differential Privacy (zCDP)

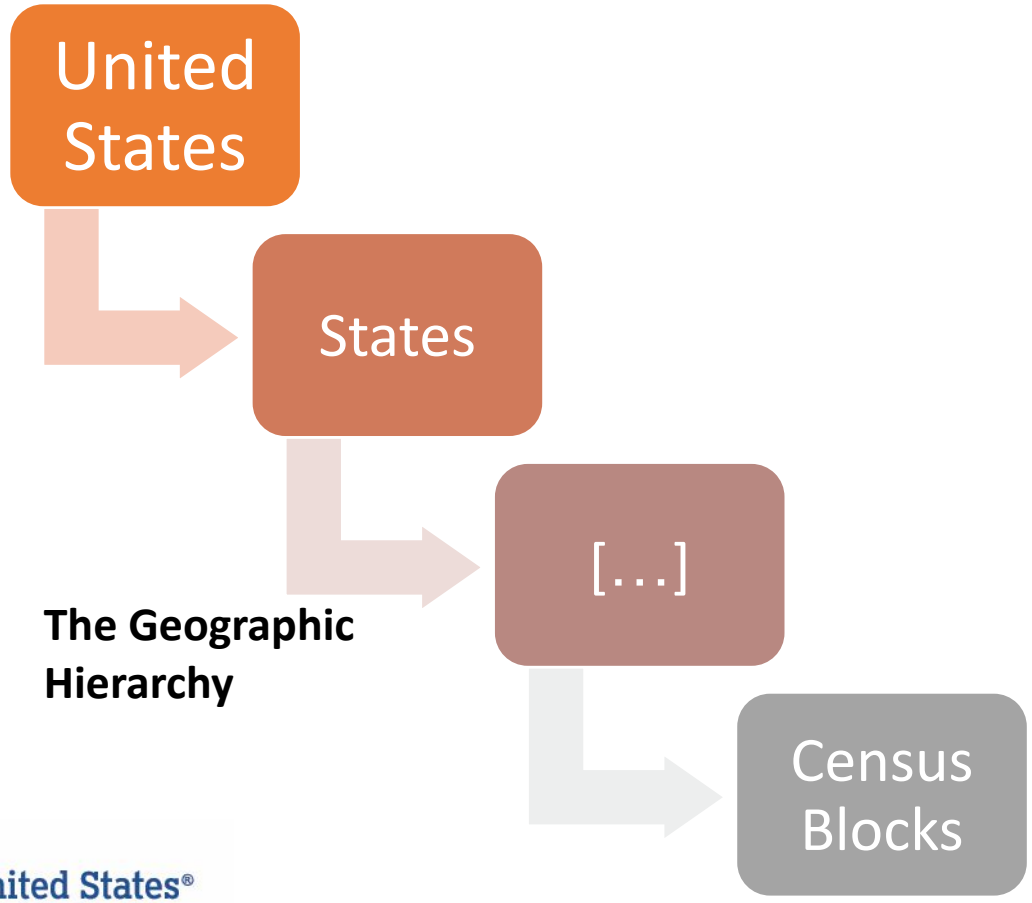
- Privacy-loss parameter: ρ (Bun and Steinke 2016)
- ρ-based privacy-loss budgets can be converted to any single point along a continuum of (ε,δ) pairs. Analysis of the privacy protection afforded by a ρ budget should use the entire continuum, not a single (ε,δ) point. Some formulas provide tighter bounds on the (ε,δ) curve implied by a particular value of ρ. TDA uses this one:

$$\varepsilon = \rho + 2\sqrt{-\rho \log_e \delta}$$

- Noise distribution: discrete Gaussian (Canonne et al. 2020, 2021)
- The expected variance of any noisy measurement can be estimated by knowing the total privacy-loss budget and the share of ρ allocated to that query at that geographic level [see Appendix B of Abowd et al. (2022) technical paper]

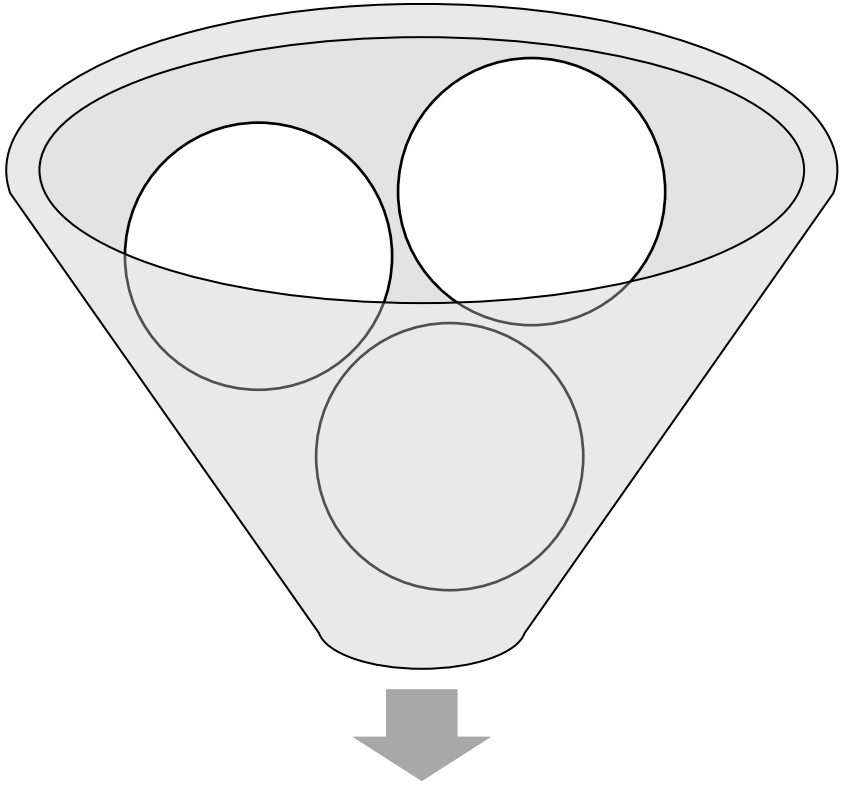
United States® Census Bureau

The TopDown Approach



The Geographic Hierarchy

At each geographic level:



Internally consistent histogram

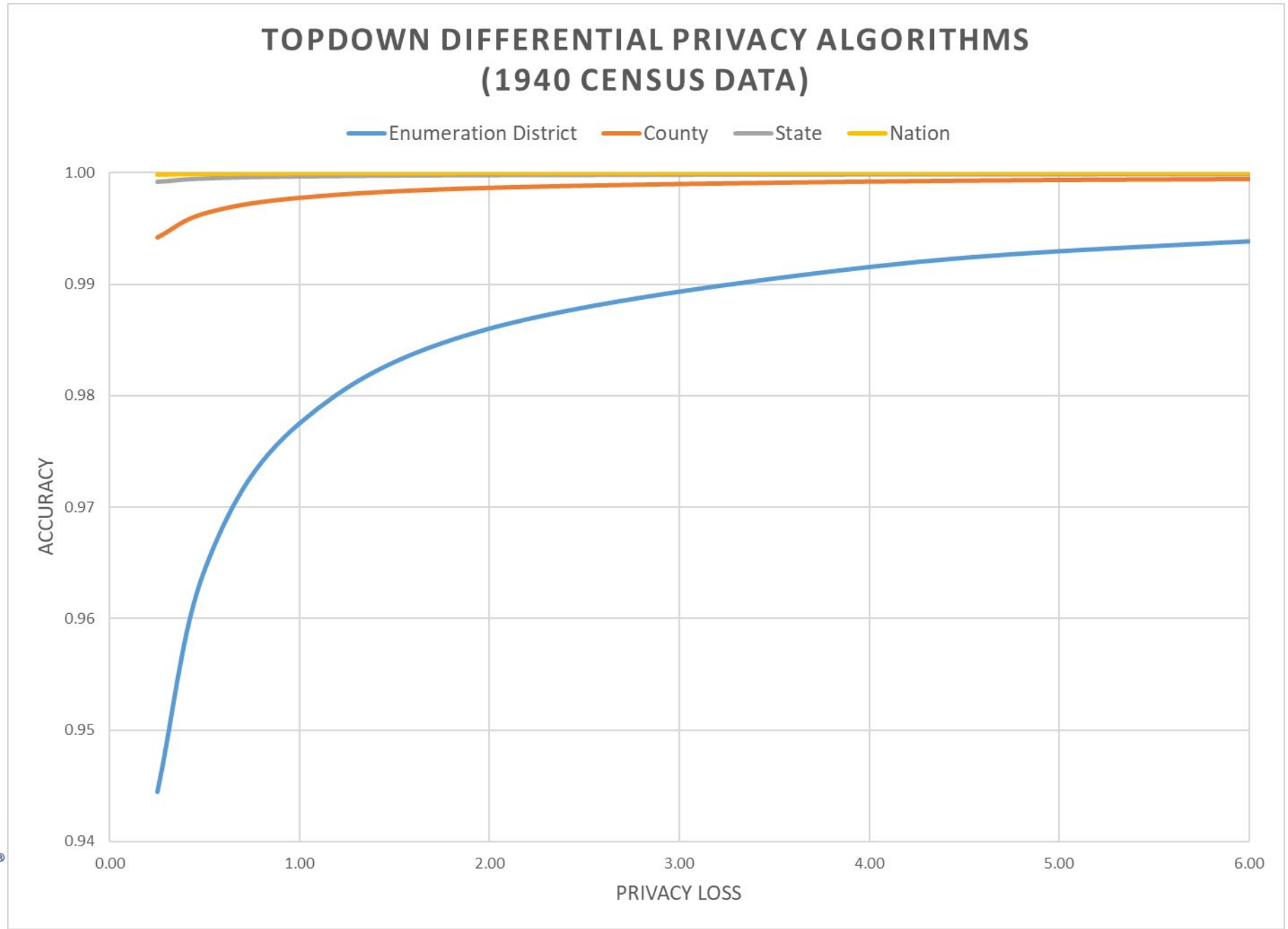
United States® Census Bureau

Naïve Method: BottomUp or Block-by-Block

- Apply differential privacy algorithms to the most detailed level of geography
- Build all geographic aggregates from those components as a post-processing

United States® Census Bureau

United States® Census Bureau



United States® Census Bureau

Benefits of TDA Compared to Block-by-block

- TDA is in stark contrast with naïve alternatives (e.g., block-by-block or bottom-up)
- TDA disclosure-limitation error does not increase with number of contained Census blocks in the geographic entity (on spine)
- TDA yields increasing relative accuracy as the population being measured increases (in general), and increased count accuracy compared to block-by-block
- TDA “borrows strength” from upper geographic levels to improve count accuracy at lower geographic levels (e.g., for sparsity)

United States® Census Bureau

If you feed TDA 16.6 billion differentially private measurements (23 trillion for DHC), it will do a good job that completely satisfies no one.

(This was predicted in Abowd and Schmutte 2019.)

United States® Census Bureau

Accurate, but to whom?

- DAS operates under interpretable formal privacy guarantees, given privacy-loss budgets
- Accuracy properties depend upon the output metric (use case)
- Distinct groups of data users will have a particular analyses they wish to be accurate
- Tuning accuracy for a given analysis can reduce accuracy for other analyses
- Policy makers must consider reasonable overall accuracy metrics for privacy tradeoffs

United States® Census Bureau

Deep Dive: Redistricting Data

- Legislative districts for politically defined entities of arbitrary size
- Must be (approximately) equal populations in each district
- Districts must be consistent with Section 2 scrutiny under the 1965 Voting Rights Act
  - Large minority populations cannot be clustered into a few districts
  - Majority-minority districts (approximately 50%+ minority population) must be drawn when feasible
- Focus statistics: total population, ratio largest race/ethnic population to total population

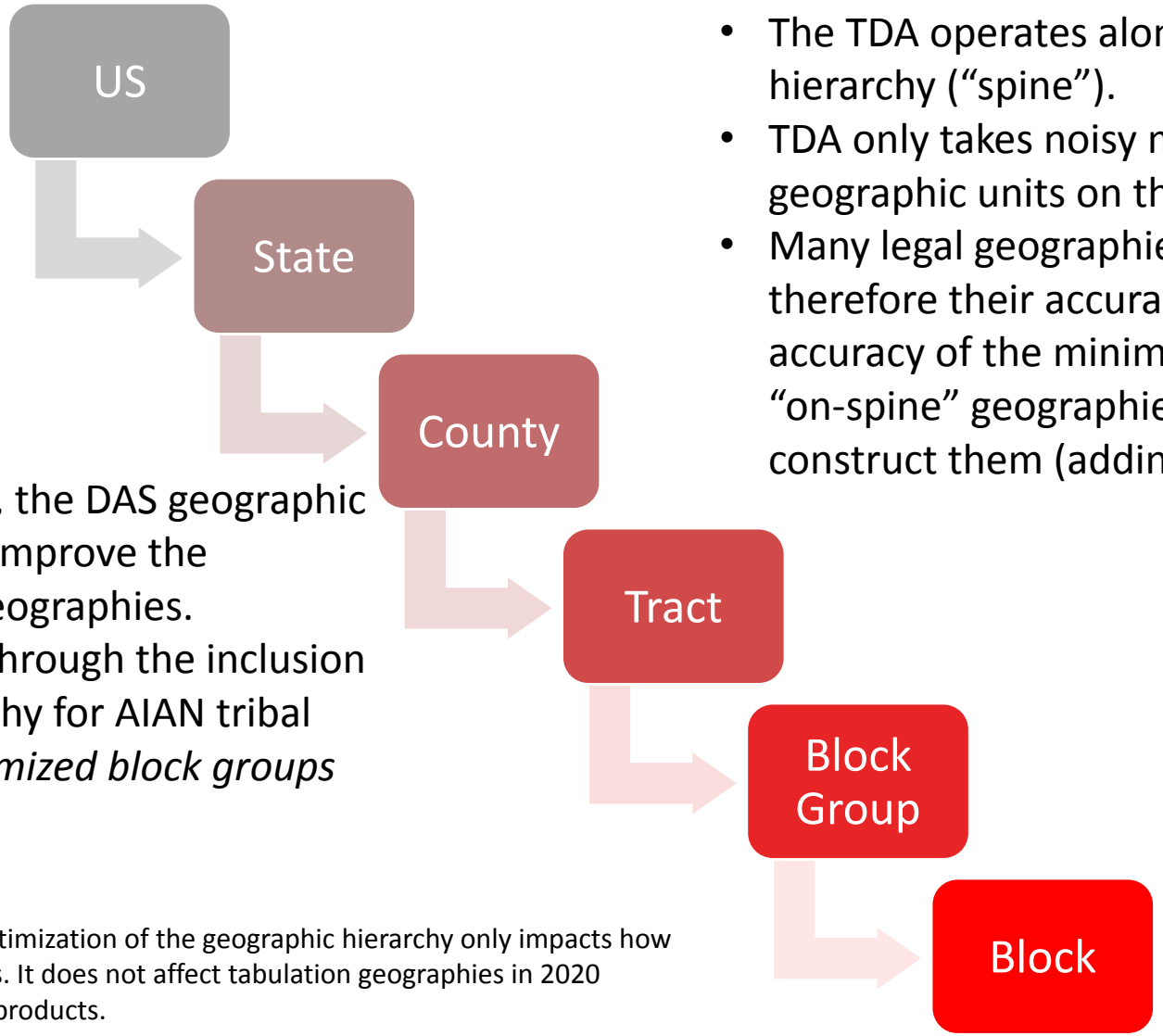
United States® Census Bureau

Multi-pass Post-processing

- The sparsity of many queries (i.e., prevalence of zeros and small counts) has the potential to introduce bias in TDA's post-processing
- To address the sparsity issue, TDA processing is performed in a series of passes
- At certain geographic levels, the algorithm constructs histograms for a subset of queries in a series of passes for that level, constraining the histogram for each pass to be consistent with the histogram produced in the prior pass
- Example for the P.L. 94-171 Redistricting Data Summary File:
  - Pass 1: Total Population
  - Pass 2: Remaining tabulations supporting P.L. 94-171 Redistricting Data

United States® Census Bureau

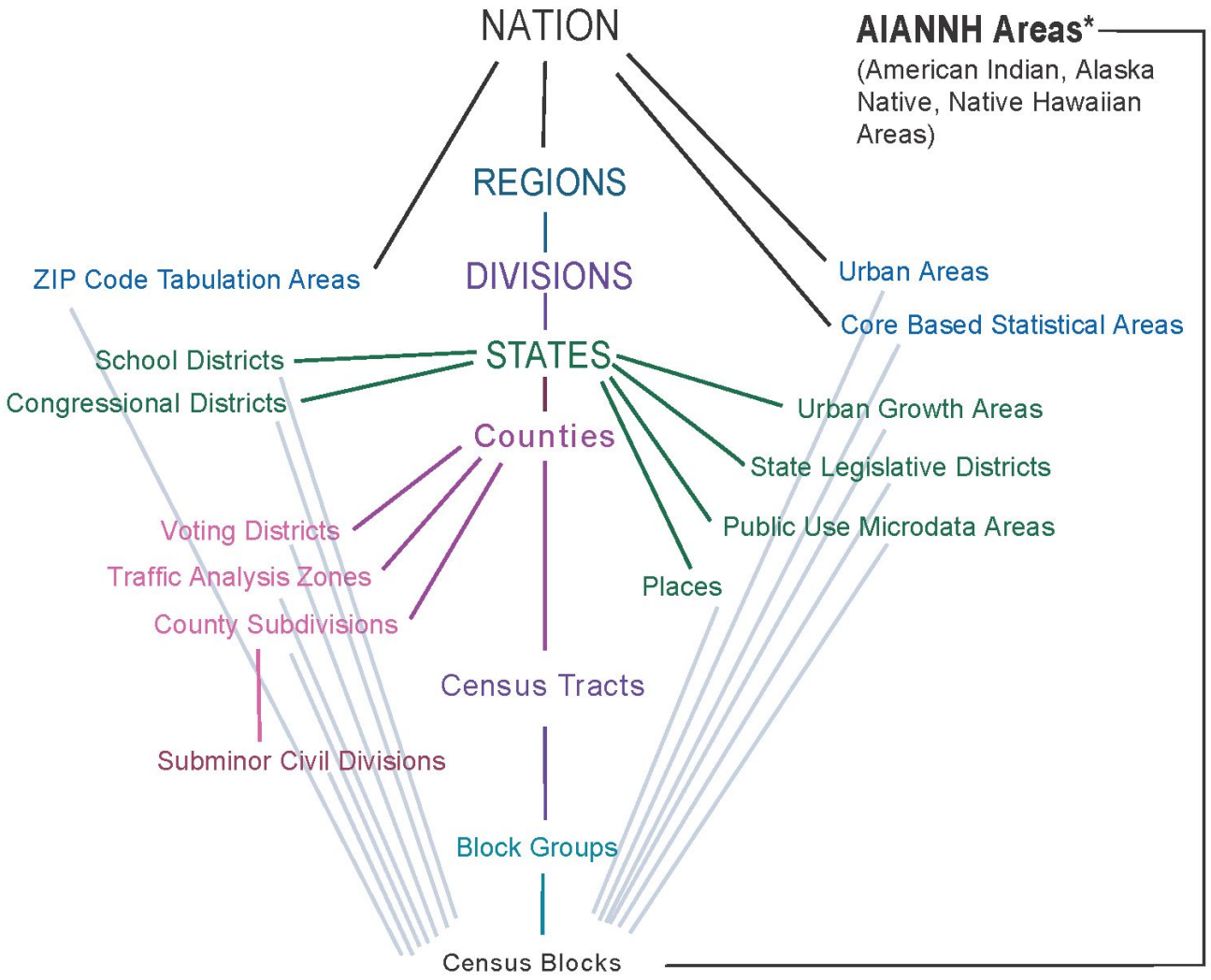
Tabulation Geographic Hierarchy



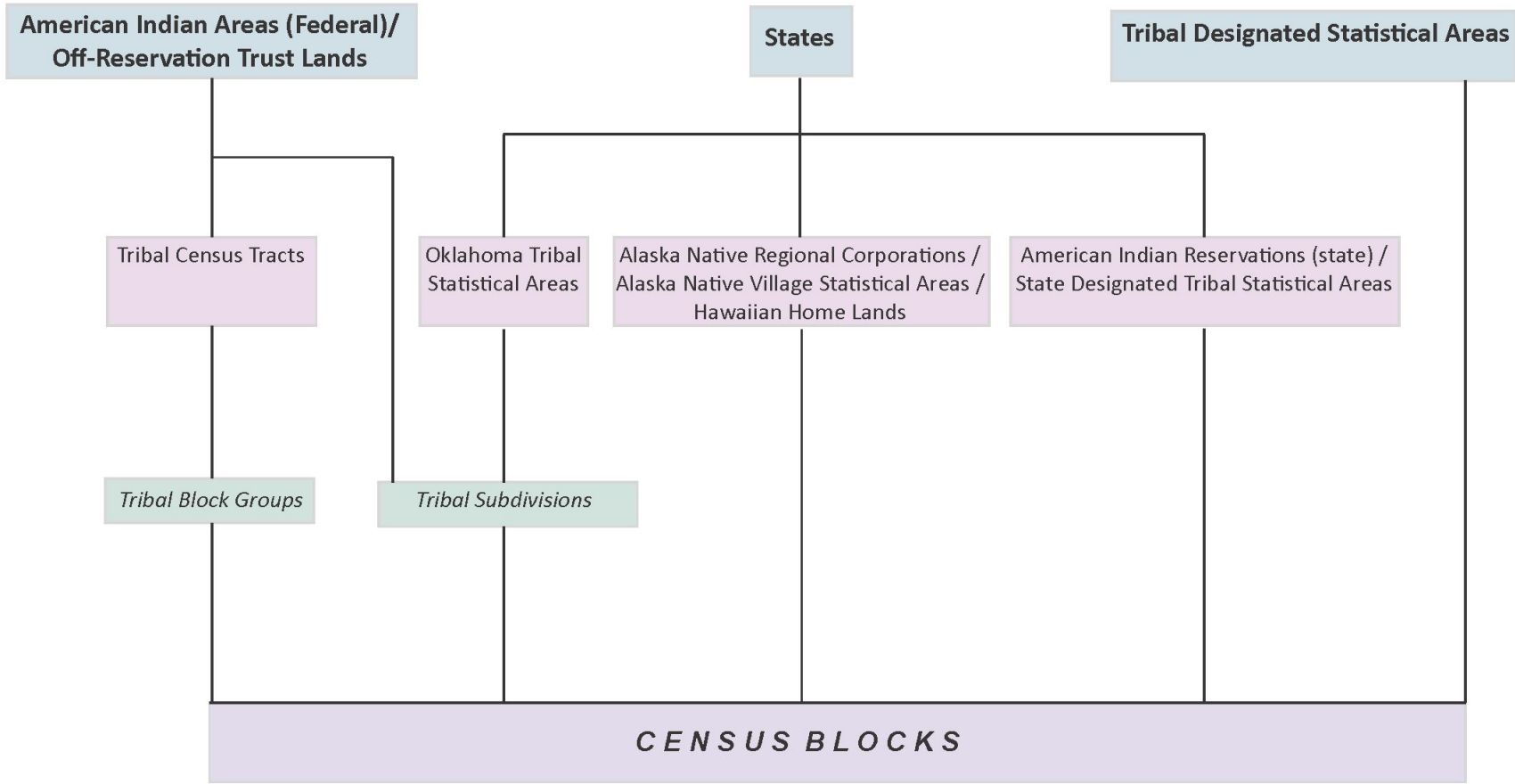
- The TDA operates along a geographic hierarchy ("spine").
- TDA only takes noisy measurements for geographic units on the hierarchy.
- Many legal geographies are "off-spine," therefore their accuracy is impacted by the accuracy of the minimum number of "on-spine" geographies that can be used to construct them (adding or subtracting).

- To address this challenge, the DAS geographic hierarchy was altered to improve the accuracy of "off-spine" geographies.
- This was done primarily through the inclusion of a branch of the hierarchy for AIAN tribal areas and the use of *optimized block groups (not shown).*

Note: The optimization of the geographic hierarchy only impacts how TDA operates. It does not affect tabulation geographies in 2020 Census data products.


NATION
REGIONS
DIVISIONS
STATES
Counties
Census Tracts
Block Groups
Census Blocks
ZIP Code Tabulation Areas
School Districts
Congressional Districts
Voting Districts
Traffic Analysis Zones
County Subdivisions
Subminor Civil Divisions
AIANNH Areas*
(American Indian, Alaska Native, Native Hawaiian Areas)
Urban Areas
Core Based Statistical Areas
Urban Growth Areas
State Legislative Districts
Public Use Microdata Areas
Places

Hierarchy of American Indian, Alaska Native, and Native Hawaiian Areas



United States® Census Bureau

Senate District 48

In 2010, this district's race/ethnicity breakdown was:

	#	Percent
Total:	292,870	100.0%
White:	268,818	91.8%
Black:	6,268	2.1%
Hispanic:	9,013	3.1%
Asian:	2,501	0.9%
All Other:	6,270	2.1%

Source: LATFOR, based on 2010 Census modified with prison population counts

The population eligible to vote (citizens of voting age) is estimated as follows:

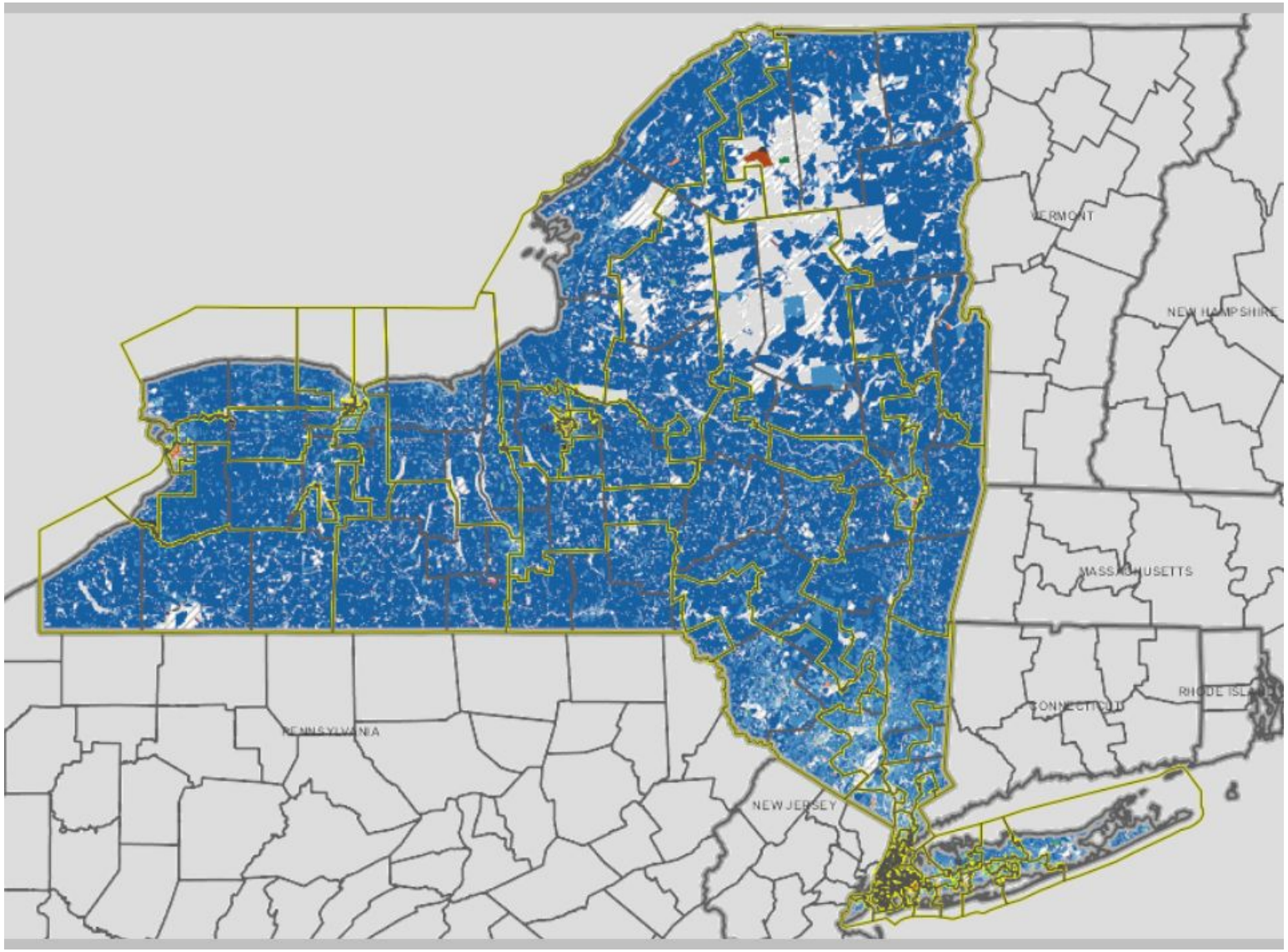
	#	Percent
Citizen voting age population (CVAP):	219,620	100.0%
White CVAP:	203,505	92.7%
Black CVAP:	6,144	2.8%
Hispanic CVAP:	5,968	2.7%
Asian CVAP:	1,142	0.5%
All Other CVAP:	2,861	1.3%

Source: Special Census Bureau tabulation from the 2006-10 5-Year American Community Survey

See the list of candidates for Senate District 48

Compare old and new district maps

Like 105 Share

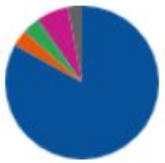


United States® Census Bureau

Assembly District 125

In 2010, this district's race/ethnicity breakdown was:

	#	Percent
Total:	133,580	100.0%
White:	111,052	83.1%
Black:	4,448	3.3%
Hispanic:	5,079	3.8%
Asian:	9,001	6.7%
All Other:	4,000	3.0%



Source: LATFOR, based on 2010 Census modified with prison population counts

The population eligible to vote (citizens of voting age) is estimated as follows:

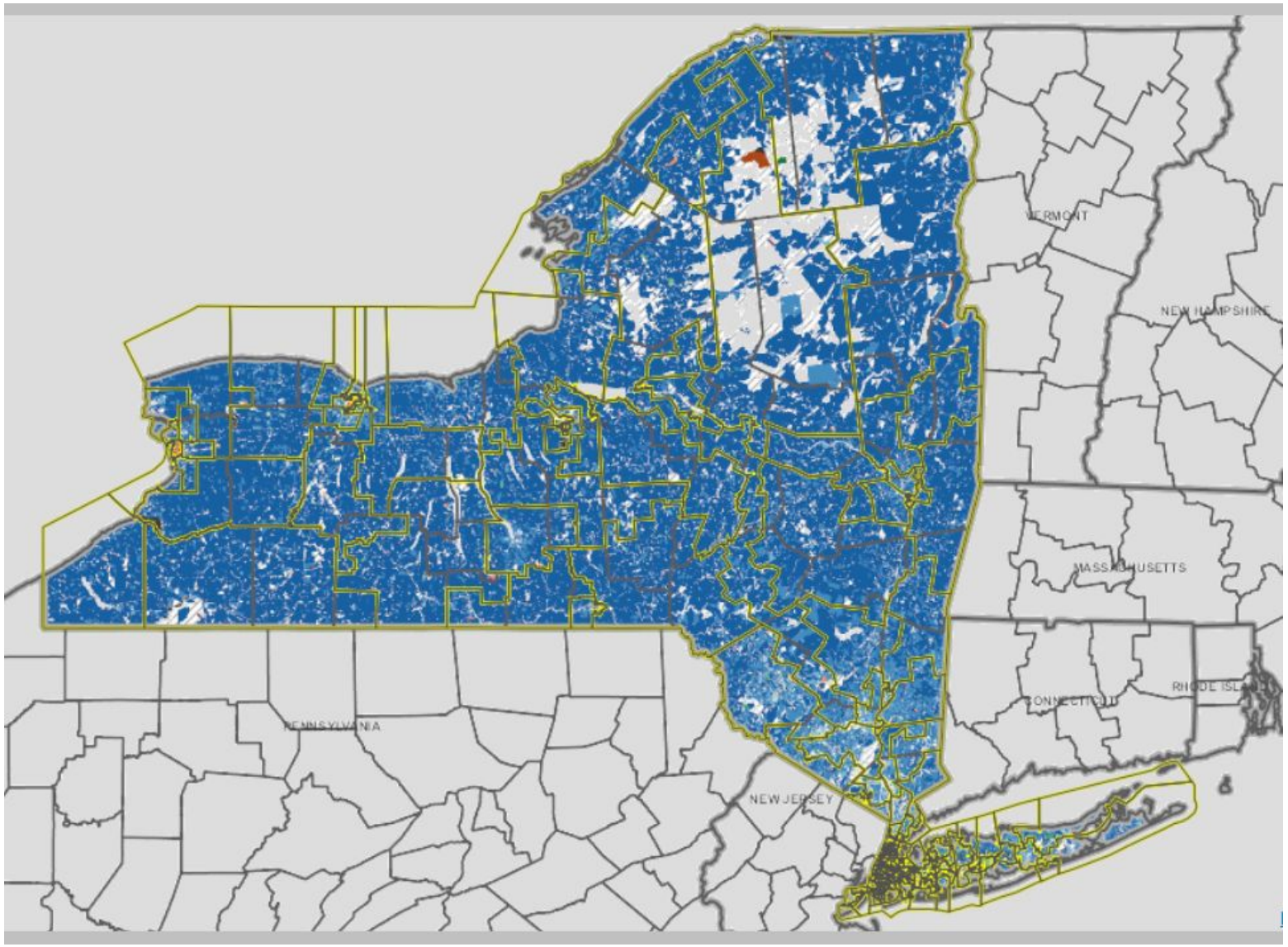
	#	Percent
Citizen voting age population (CVAP):	100,790	100.0%
White CVAP:	89,235	88.5%
Black CVAP:	2,927	2.9%
Hispanic CVAP:	3,221	3.2%
Asian CVAP:	3,710	3.7%
All Other CVAP:	1,697	1.7%

Source: Special Census Bureau tabulation from the 2006-10 5-Year American Community Survey

See the list of candidates for Assembly District 125

Compare old and new district maps

Like 105 Share



United States® Census Bureau

How to reconcile these statistics

- Construct error metrics of the form

$$Pr[|TDA - CEF| \leq \alpha] \geq 1 - \beta$$

- Less than α error with probability at least 1-β for a target minimum population
- Statistical interpretation: absolute differences (=RMSE differences) greater than α are outside the 1-β confidence interval
- A single statistic can be used to tune the redistricting application

$$\frac{Population\ of\ Largest\ Race\ or\ Ethnic\ Group}{Total\ Population}$$

- Calculated for the TopDown Algorithm (TDA) output and the 2020 Census (CEF)
- Implemented successfully for the production code release
- *In the production data: minimum population of 200 to 249 for political areas and 450 to 499 for block groups to achieve 95% accuracy (α = 0.05) at least 95% of the time (β = 0.05) See Wright and Irimata (2021)*

United States® Census Bureau

What do the redistricting data do?

- Total differentially private measurements (queries): 16.6 billion
- Global $\rho = 2.63$ [$(\varepsilon, \delta) = (18.19, 10^{-10})$ and infinitely many other pairs] U.S. persons and housing units
- Total block-level tables 29.4 million
- Total block-level statistics 3.4 billion
- Total independent block-level statistics 1.5 billion
- Accuracy of populations and largest race/ethnic group fit for redistricting and Voting Rights Act scrutiny for populations of at least 200-249, which is much smaller than legal entity subject to VRA

United States® Census Bureau

Figure 2. Mean Absolute Error of the County Total Population among the Least Populous Counties (Population Under 1,000) by Demonstration Data Product Vintage

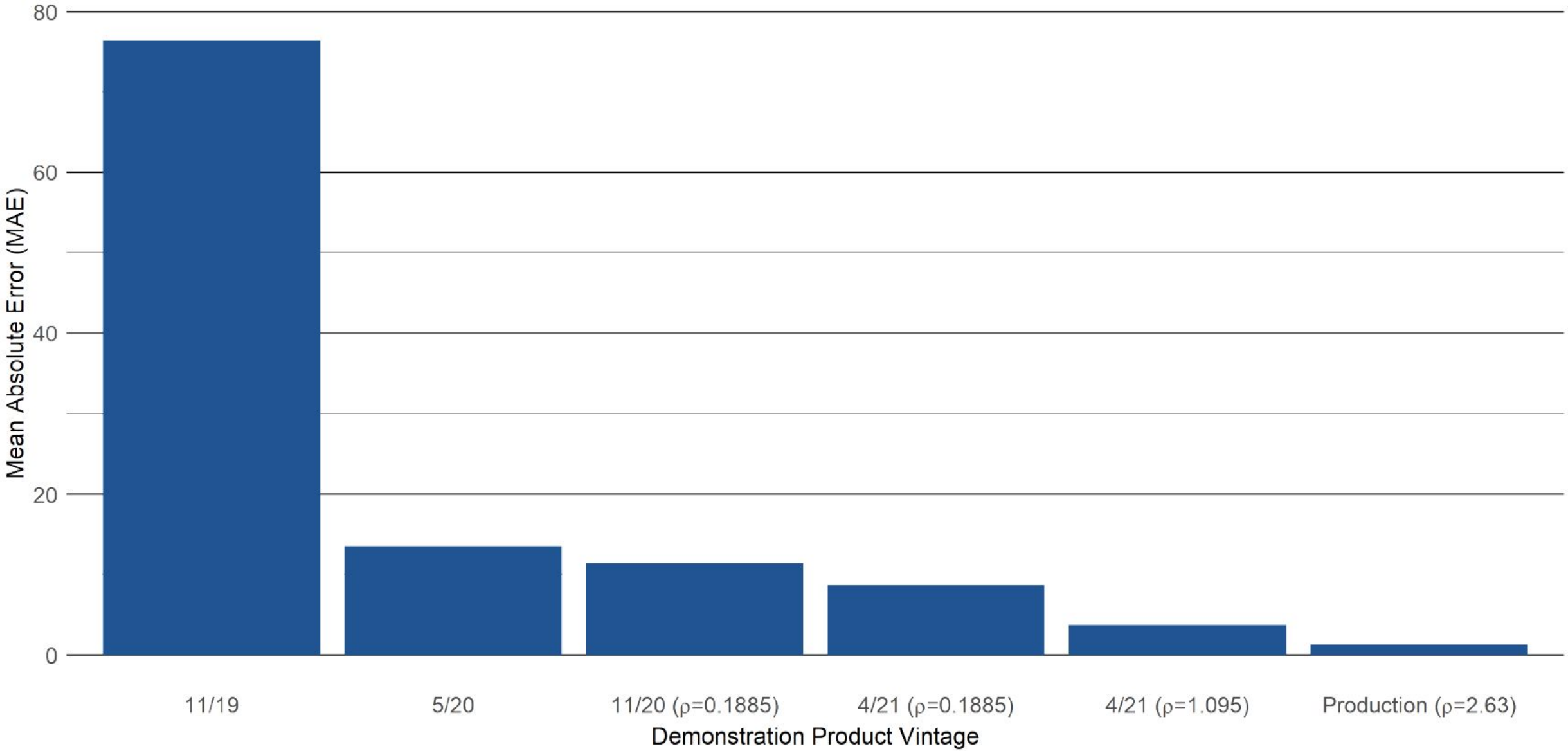


Figure 3. Mean Absolute Error of the Total Population for Federal American Indian Reservation/Off-Reservation Trust Lands by Demonstration Data Product Vintage

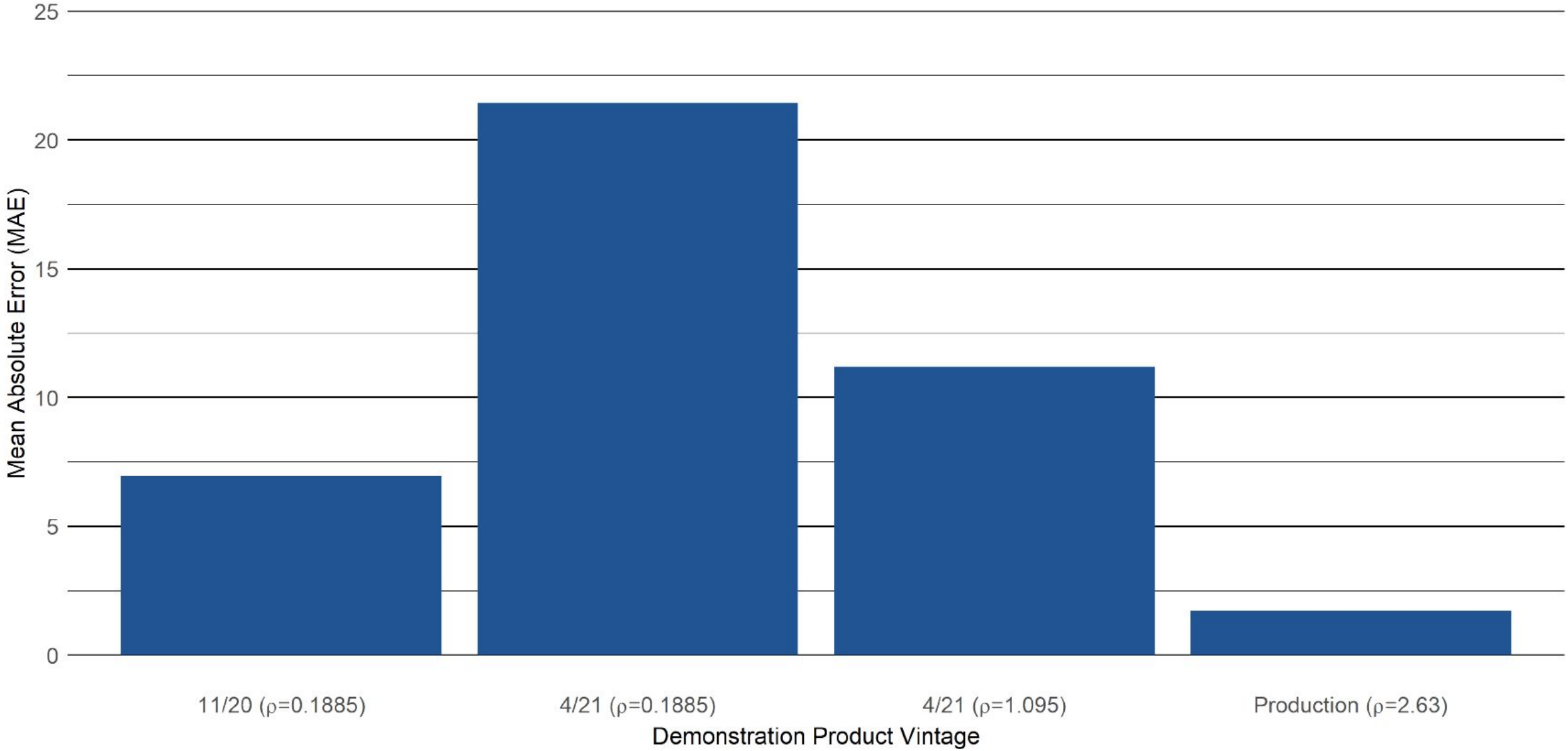


Figure 4. Mean Absolute Error of the Total Population among All Incorporated Places by Demonstration Data Product Vintage

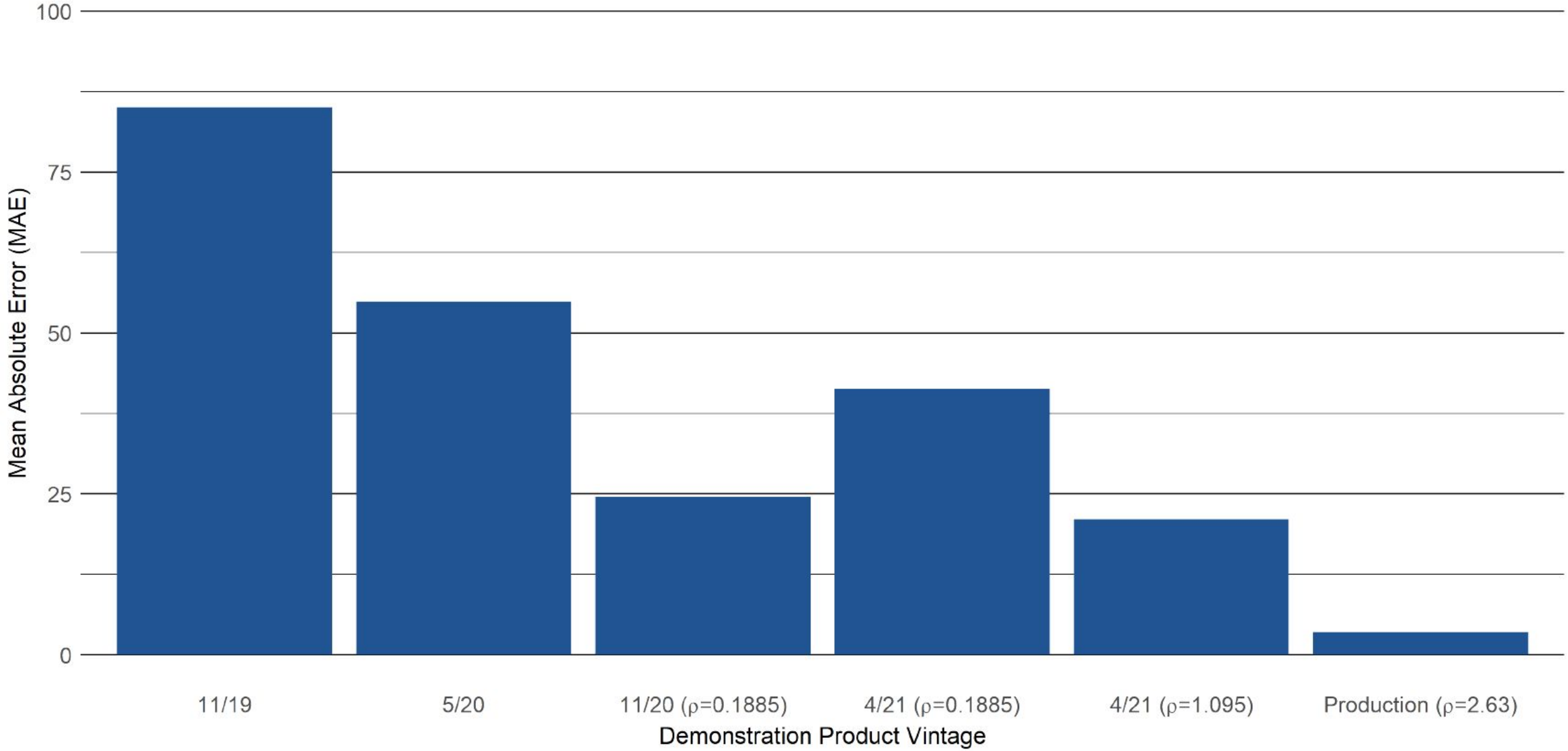
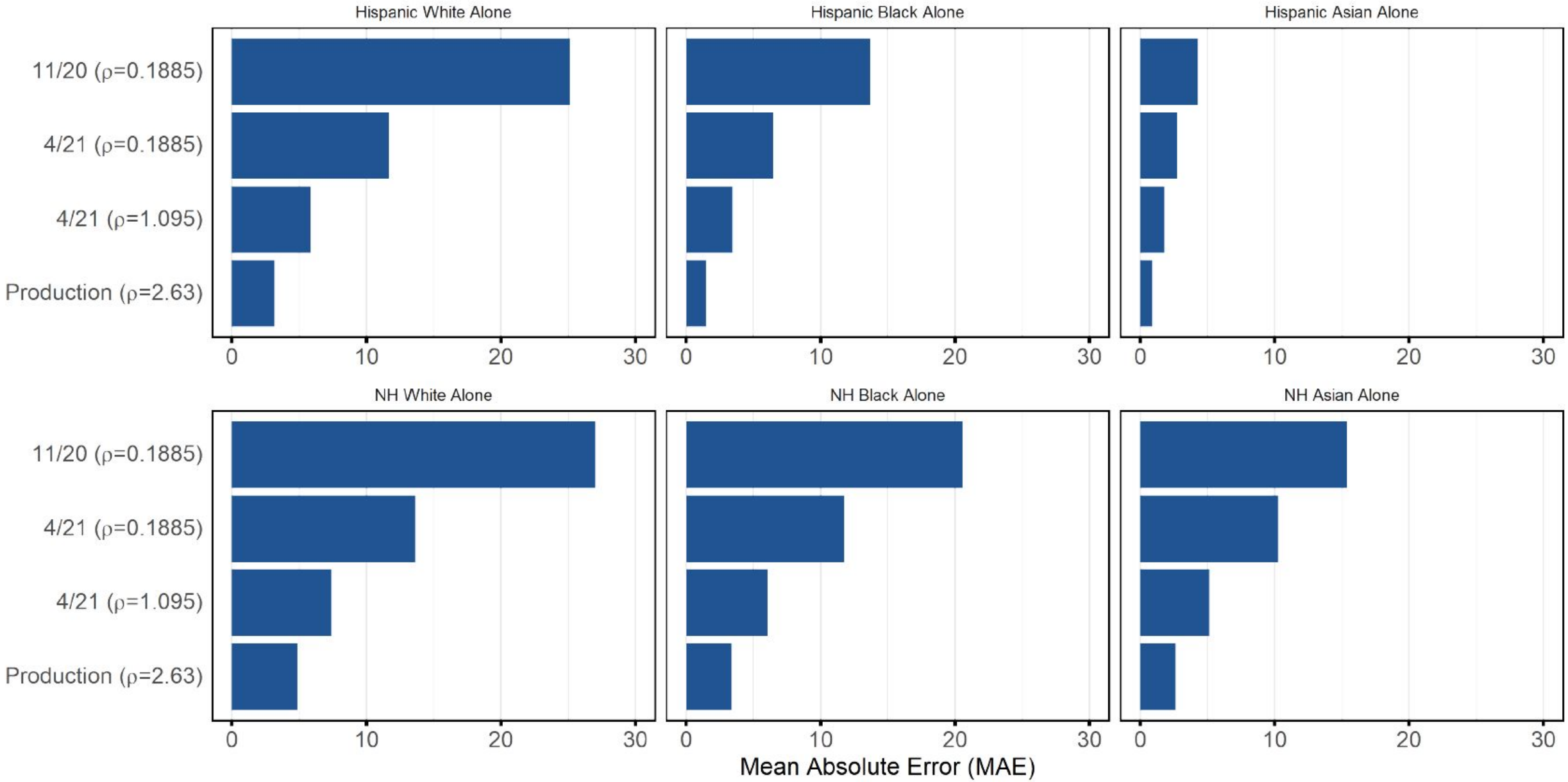


Figure 5. Mean Absolute Error of the Total Population among Tracts for Hispanic x Race Alone Populations by Demonstration Data Product Vintage



Addressing Other Biases

	April 2021 PPMF		Production Settings	
	Diversity Quintile	**Mean Difference In Total Population**	**Diversity Quintile**	**Mean Difference In Total Population**
Block Groups	0 – Least Diverse	5.04	0 – Least Diverse	-0.375
	1	4.24	1	1.009
	2	0.99	2	0.997
	3	-2.21	3	-0.303
	4 – Most Diverse	-8.07	4 – Most Diverse	-1.352
Tracts	0 – Least Diverse	15.95	0 – Least Diverse	0.029
	1	11.15	1	0.045
	2	3.01	2	0.000
	3	-6.17	3	-0.020
	4 – Most Diverse	-23.94	4 – Most Diverse	-0.053

United States Census Bureau

Block-Level Inconsistencies Due to DAS-induced Uncertainty

Inconsistency	April 2021 ρ=1.095 Count of Blocks	Production Settings ρ=2.63 Count of Blocks
Occupied Housing Units > Household Population	203,519	303,984
Zero Occupied Housing Units; > 0 Household Population	674,598	505,840
Zero Household Population; > 0 Occupied Housing Units	77,947	148,836
Everyone in Block Under 18	90,534	163,884
> 10 Persons Per Household	87,342	121,376

United States® Census Bureau

Privacy-loss Budget Allocation (by geographic level)

Privacy-loss Budget Allocation 2021-06-08 Person Tables (Production Settings) United States	
Global ρ	2.56
Global ε (incl. units)	18.19
delta	10^{-10}

	ρ Allocation by Geographic Level
US	104/4099
State	1440/4099
County	447/4099
Tract	687/4099
Optimized Block Group*	1256/4099
Block	165/4099

Privacy-loss Budget Allocation 2021-06-08 Units Tables (Production Settings) United States	
Global ρ	0.07

	ρ Allocation by Geographic Level
US	1/205
State	1/205
County	7/82
Tract	364/1025
Optimized Block Group*	1759/4100
Block	99/820

Census Bureau

Privacy-loss Budget Allocation (by query)

Query	Per Query ρ Allocation by Geographic Level					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		3773/4097	3126/4097	1567/4102	1705/4099	5/4097
CENRACE (63 cells)	52/4097	6/4097	10/4097	4/2051	3/4099	9/4097
HISPANIC (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHINSTLEVELS (3 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHGQ (8 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HISPANIC*CENRACE (126 cells)	130/4097	12/4097	28/4097	1933/4102	1055/4099	21/4097
VOTINGAGE*CENRACE (126 cells)	130/4097	12/4097	28/4097	10/2051	9/4099	21/4097
VOTINGAGE*HISPANIC (4 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE*HISPANIC*CENRACE (252 cells)	26/241	2/241	101/4097	67/4102	24/4099	71/4097
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	189/241	230/4097	754/4097	241/2051	1288/4099	3945/4097

United States® Census Bureau

Table 4 Accuracy of 2010 Census, enhanced Swap, and DP: mean absolute error (in persons) for age group population counts at the county level

Age group	2010 Census	Enhanced swap	DP
0-17 years	0	256.41	9.84
18-64 years	NA[a]	494.16	12.83
65 years and over	NA[a]	431.37	12.66

[a]Error statistics for the impact of swapping as applied to the published 2010 Census are confidential. The 2010 Census swapping algorithm kept the number of non-voting age individuals (0-17 years) invariant but did inject noise into the age groups within the voting age population. DRB clearance number CBDRB-FY22-DSEP-003. Data are from Devine & Spence (2022). Abbreviations: DP, differential privacy; DRB, Disclosure Review Board; NA, not available.

United States® Census Bureau

Table 5 Reidentification statistics for 2010 Census, enhanced swap, and DP

Reidentification Statistic	2010 Census	Enhanced swap	DP
Putative reidentification rate	97.0%	75.4%	44.4%
Confirmed reidentification rate	75.5%	46.6%	27.4%
Precision rate	77.8%	61.8%	61.7%
Precision for population uniques (nonmodal race)	81.4%	33.4%	24.0%

DRB clearance number CBDRB-FY22-DSEP-004. Data are from Abowd et al. (under review) released in Hawes (2022). External Matching File: Census Edited File. Abbreviations: DP, differential privacy; DRB, Disclosure Review Board.

- Tables 4 and 5 illustrate that TDA is a much more efficient disclosure avoidance mechanism for controlling accuracy and confidentiality than swapping with aggregation, as also shown in Abowd and Schmutte 2019.

United States® Census Bureau

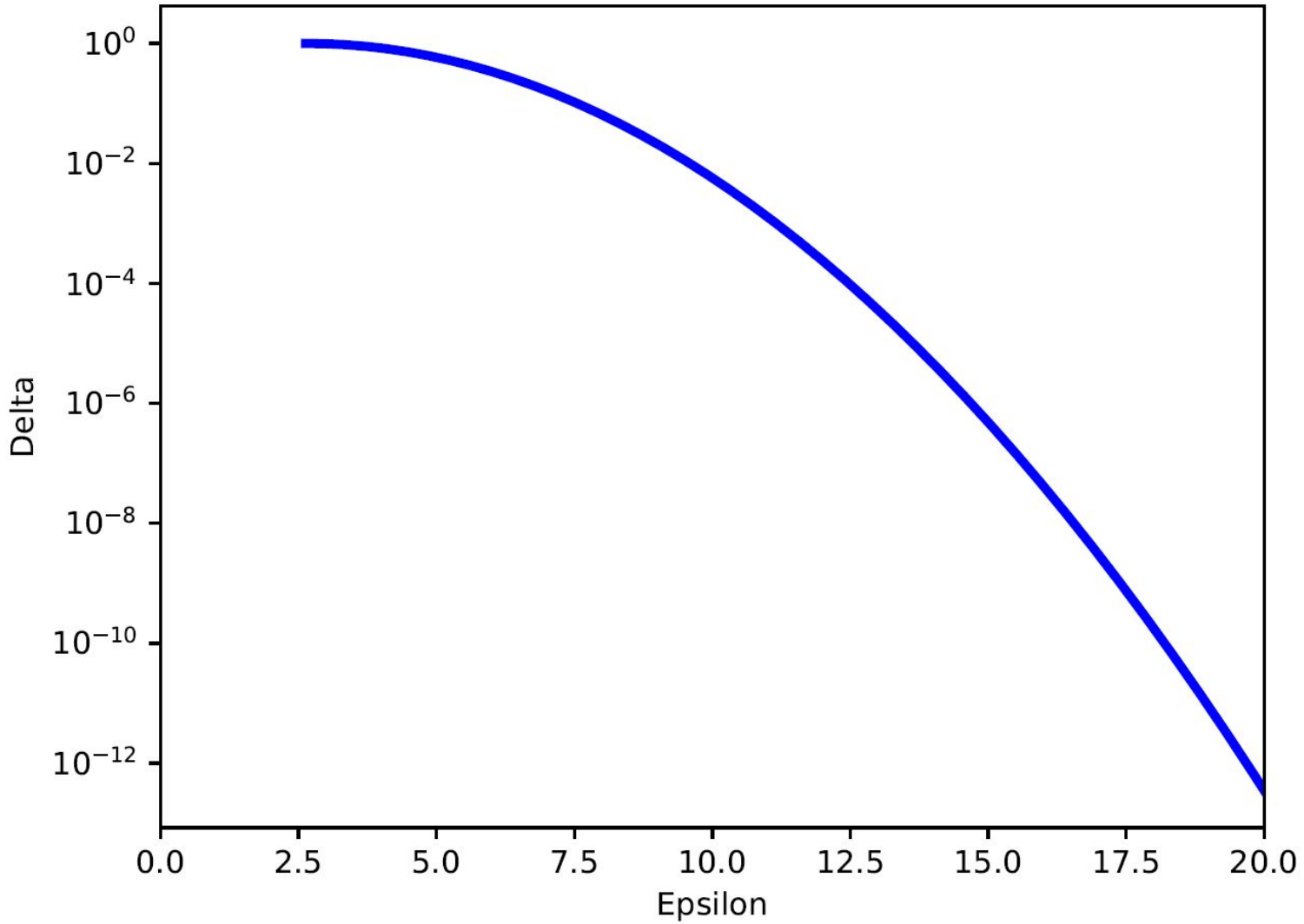


FIGURE 5. Bayesian (ϵ, δ) curve under the semantics of Section 7.4.2 for $\rho = 2.63$.

Source: Kifer et al. In preparation.

Significance Level	Power (Gaussian)	Power (DGM)	zCDP Upper Bound
0.01	0.032	0.032	0.037
0.05	0.12	0.12	0.14
0.10	0.21	0.21	0.24

TABLE 2. **Block within Custom Block Group:** Likelihood ratio test significance level/power tradeoff for block-level queries (1) if Gaussian noise is used, (2) if discrete Gaussian noise is used, (3) guaranteed upper bound if an arbitrary ρ-zCDP mechanism with $\rho = 0.1115$ is used.

$$\sup_{x>1} \frac{\alpha^x(1-\beta)^{1-x} + (1-\alpha)^x\beta^{1-x}}{e^{\rho x(1-x)}} \leq 1$$

where α is the level (probability of a Type I error), β is the probability of a Type II error, and $(1-\beta)$ is the power of the likelihood ratio test for correctly attaching a block-id to a record when block group, age, sex, race and ethnicity are known for zCDP.
H_0: N(0,1/(2ρ)); H_1: N(1,1/(2ρ))

United States® Census Bureau

Source: Kifer et al. In preparation.

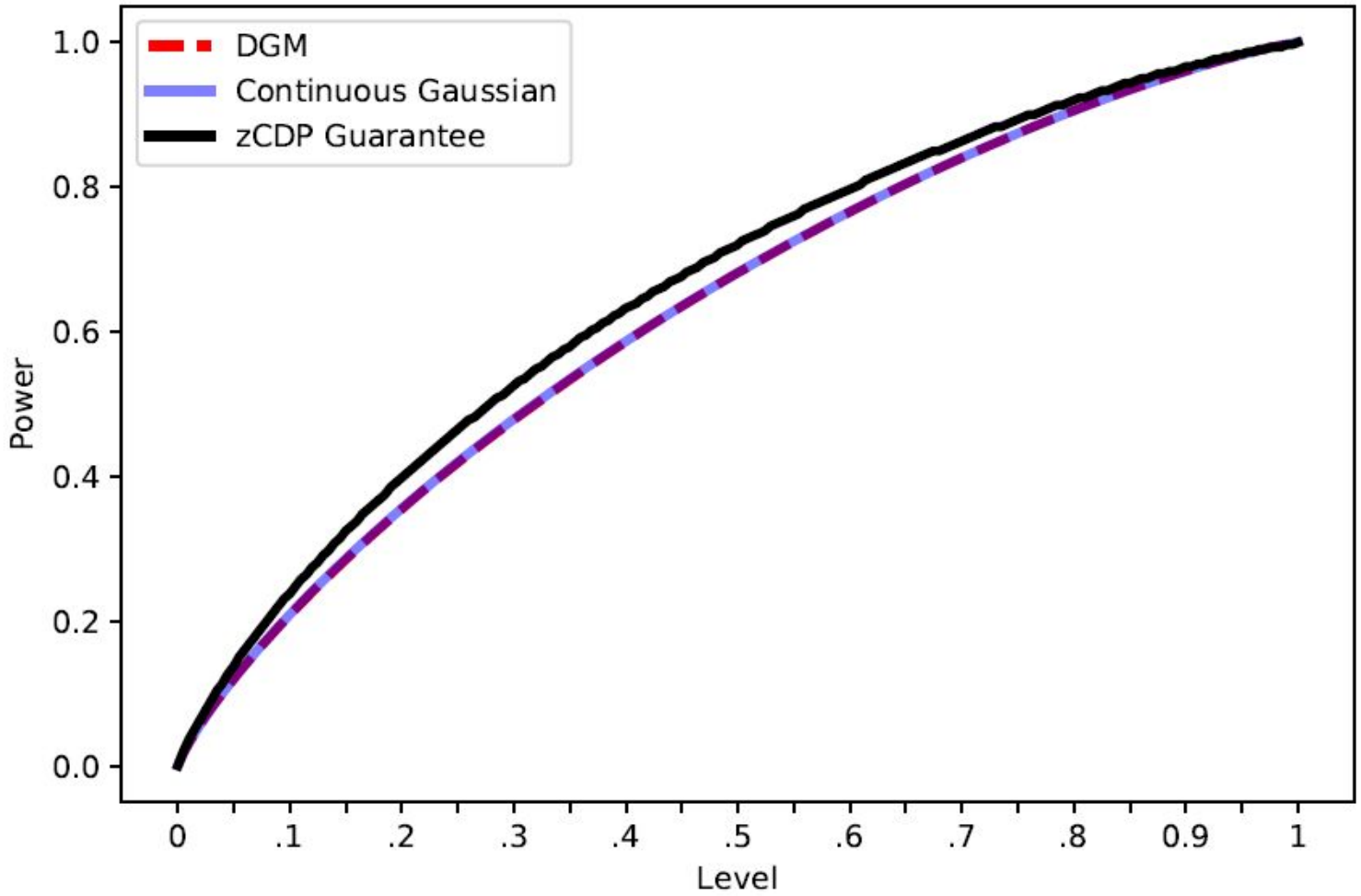


FIGURE 6. **Block within Custom Block Group:** Level (x-axis) vs. power (y-axis) curves for (1) the Gaussian mechanism over block-level queries at production settings for redistricting data ($\rho = 0.1115$), (2) the likelihood ratio test of the discrete Gaussian block-level noisy queries at production settings for redistricting data.

Source: Kifer et al. In preparation.

Privacy protection out of the shadows

- Certain privacy practices for previous censuses depended upon obfuscation
- 2020 DAS demonstration data are the most transparent view into Census Bureau privacy practices ever
- We appreciate and are excited to assess feedback from our external partners

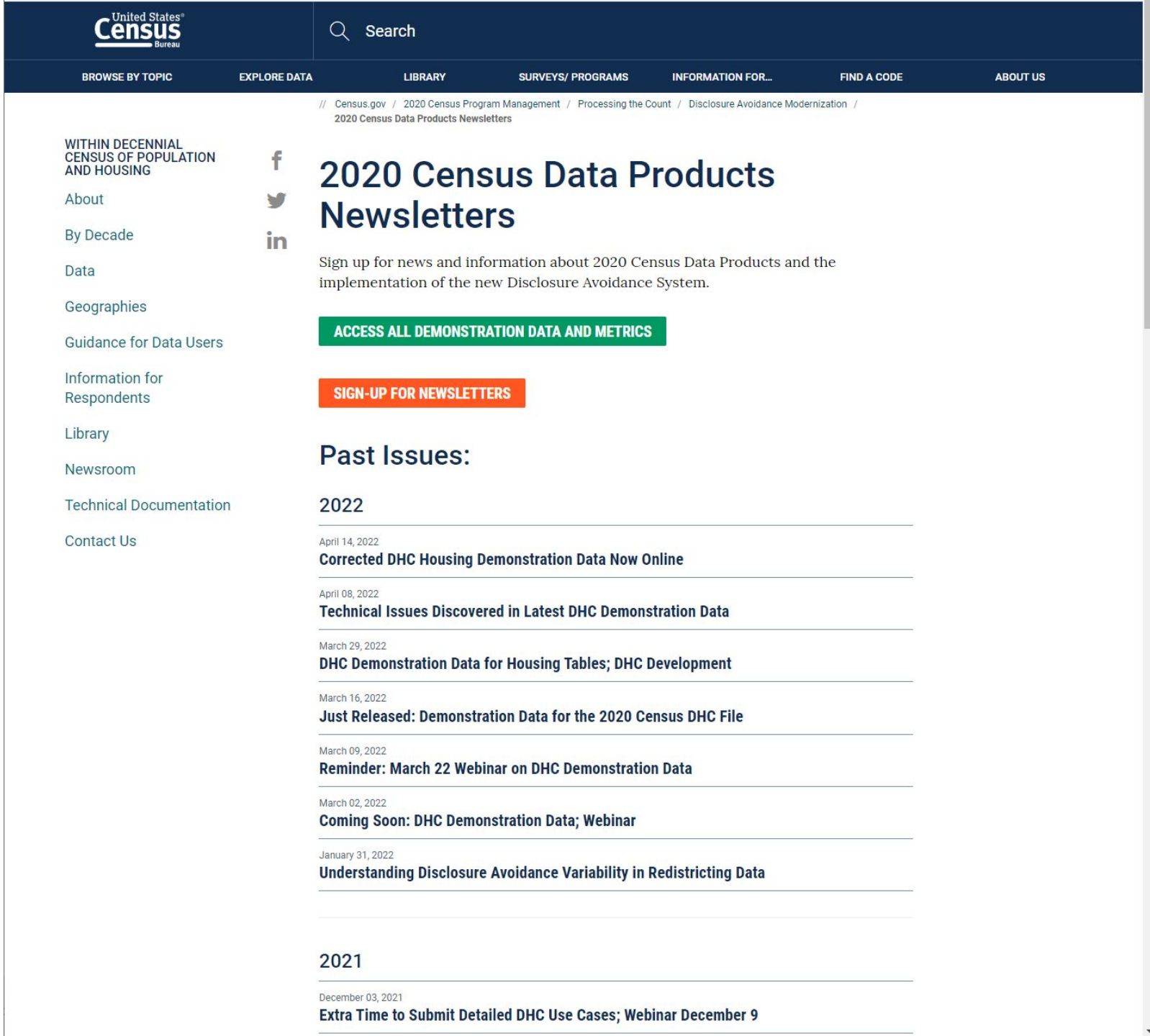
United States® Census Bureau

Stay Informed:

Subscribe to the 2020 Census Data Products Newsletters

*Search “Disclosure Avoidance” at www.census.gov or click the graphic

United States® Census Bureau



Stay Informed:

Visit Our Website

*Search “Disclosure Avoidance” at www.census.gov or click the graphic

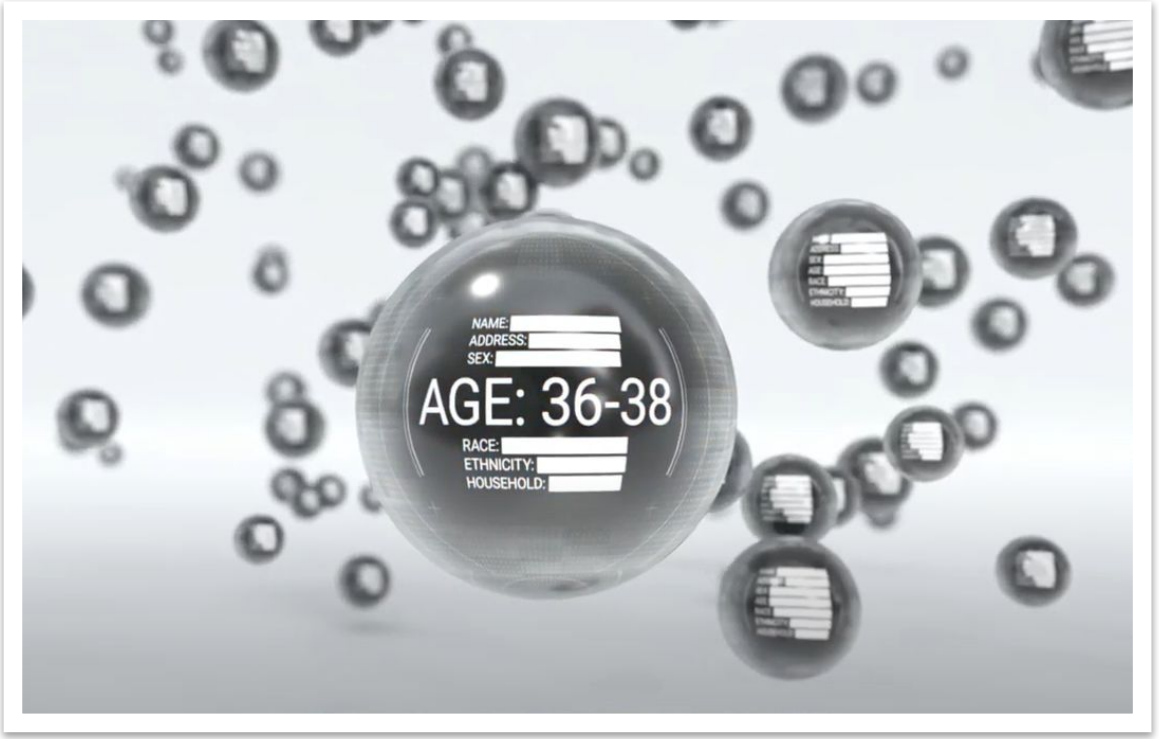
United States® Census Bureau

** Video **

Protecting Privacy in Census Bureau Statistics

*Find it on our website and YouTube Page

Search “Disclosure Avoidance” at www.census.gov or click the graphic



United States® Census Bureau

Selected Additional Resources

- Code: uscensusbureau/DAS_2020_Redistricting_Production_Code: Official release of source code for the Disclosure Avoidance System (DAS) used to protect against the disclosure of individual information based on published statistical summaries. (github.com)
- Technical: *HDSR* The 2020 Census Disclosure Avoidance System TopDown Algorithm
- Updates: Developing the DAS: Demonstration Data and Progress Metrics (census.gov)

United States® Census Bureau

Selected References

- Abowd, John M. and Ian M. Schmutte. 2019. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*, Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627.
- Bun M and Steinke T. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *TCC 2016* 9985:635–658 https://doi.org/10.1007/978-3-662-53641-4_24.
- Canonne C, Kamath G, and Steinke T. 2020. The Discrete Gaussian for Differential Privacy. In Larochelle H, Ranzato M, Hadsell R, Balcan M, and Lin H (eds.) *NeurIPS* pp. 15676—15688 https://proceedings.neurips.cc/paper/2020/file/b53b3a3d6ab90ce0268229151c9bde11-Paper.pdf.
- Canonne C, Kamath G, and Steinke T. 2021. The Discrete Gaussian for Differential Privacy. https://arxiv.org/abs/2004.00010.
- Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.
- Garfinkel S, Abowd J, and Martindale C. 2019. Understanding database reconstruction attacks on public data. \textit{Communications of the ACM} 62,3:46--53 \url{https://doi.org/10.1145/3287287}
- Hawes, Michael. 2022. Reconstruction and Re-identification of the Demographic and Housing Characteristics File (DHC). Presentation to the Census Scientific Advisory Committee, March 17, 2022, https://www2.census.gov/about/partners/cac/sac/meetings/2022-03/presentation-reconstruction-and-reidentification-of-the-dhc.pdf.
- Dwork, Cynthia. 2006. Differential Privacy, *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006), Springer Verlag, 4052*, 1-12, ISBN: 3-540-35907-9.
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006a. in Halevi, S. & Rabin, T. *(Eds.)* Calibrating Noise to Sensitivity in Private Data Analysis *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg*, 265-284, DOI: 10.1007/11681878_14.
- Dwork C, Kenthapadi K, McSherry F, Mironov I, and Naor M. 2006b. Our data, ourselves: Privacy via distributed noise generation. In: Vaudenay S (ed.) *EUROCRYPT* 24:486-503 https://doi.org/10.1007/11761679_29
- Dwork, Cynthia and S. Yekhanin. 2008. New Efficient Attacks on Statistical Disclosure Control Mechanisms. In: Wagner D. (eds) *CRYPTO 2008* Vol 5157 (Berlin: Springer). https://doi.org/10.1007/978-3-540-85174-5_26.
- Daniel Kifer, John Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. In preparation. Bayesian and Frequentist Semantics of Common Variations of Differential Privacy: Applications to the 2020 Census.
- Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd , Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.
- McKenna, Laura. 2019. Disclosure Avoidance Techniques Used for the 1960 Through 2010 Census. https://www.census.gov/library/working-papers/2019/adrm/six-decennial-censuses-da.html.
- McKenna Laura. 2018. Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf.
- Wright ,Tommy and Kyle Irimata. 2021. Empirical Study of Two Aspects of the Topdown Algorithm Output for Redistricting: Reliability & Variability (August 5, 2021 Update) *USCB CSRM Studies Series* Statistics 2021-20 https://www.census.gov/content/dam/Census/library/working-papers/2021/adrm/SSS2021-02.pdf.

United States® Census Bureau

Thank you.

John.Maron.Abowd@census.gov

United States® Census Bureau