U.S. Census Bureau Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Population Association of America
April 12, 2023

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session I: Welcome, Objectives, and Introductions

Sallie Ann Keller
Associate Director for Research and
Methodology and Chief Scientist

United States® Census Bureau

The Formal Privacy Guidance on Using and Interpreting Data and Estimates (FP-GUIDE) Initiative

Mission Statement: To identify and develop methods, resources, and guidance to assist data users in effectively using differential privacy-adjusted data in statistical and demographic analysis.

FP-GUIDE Objectives

1. Identify major statistical and demographic use cases for which resources and guidance should be developed.
2. Research and develop statistical methods for incorporating total DP-error (bias and variance) into the methods for the identified use cases.
3. Generate measures of total DP-error for the 2020 Census Redistricting Data (P.L. 94-171) Summary File and Demographic and Housing Characteristics File (DHC) data products.
4. Develop guidance and communications materials to support data users in incorporating these statistical methods into their analyses.
5. Catalyze a longer-term research program on using DP-adjusted data involving external academic partners and data users.

Today, we will be sharing the results of some of the FP-GUIDE initiative's preliminary research and we'll introduce you to some tools and resources to help you as you use DP-adjusted census data.

United States® Census Bureau

Today's Agenda

8:00am – 8:15am	Welcome and Introductions
8:15am – 8:30am	Data Product Updates
8:30am – 9:00am	An Overview of the 2020 Census Disclosure Avoidance System
9:00am – 9:25am	Detailed Summary Metrics
9:25am – 9:30am	Disclosure Avoidance Briefs: New Directions Based on Recommendations
9:30am – 10:30am	Constructing Uncertainty Estimates for TDA-based Data Products
10:30am – 10:45am	Break
10:45am – 11:45am	Using the Noisy Measurement File
11:45am – 12:00pm	Concluding Discussion

United States® Census Bureau

Who are we?

- Sallie Ann Keller
- Robert Ashmead
- Michael Hawes
- Cynthia Davis Hollingsworth
- Beth Jarosz
- Alexandra Krause
- Roberto Ramirez
- Matthew Spence

Who are you?

Please tell us:

- Your name
- Your organization
- One thing you hope to get out of today’s workshop

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session II: Data Product Updates

Alexandra Krause and Roberto Ramirez
Population Division

United States® Census Bureau

Background
2020 Census Data Products

Alexandra Krause
U.S. Census Bureau

United States® Census Bureau

2020 Census Data Products

Released

Apportionment
April 26, 2021

Redistricting File
(Public Law 94-171)
August 12, 2021
September 16, 2021

Demographic Profile

Demographic and Housing Characteristics File (DHC)
Planned May 2023

Detailed DHC-A
Planned Sept 2023

Detailed DHC-B
Release Date TBD

Supplemental DHC (S-DHC)
Release Date TBD

Future Effort

Privacy Protected Microdata File (PPMF)

Special Tabulations

United States® Census Bureau

Apportionment Release

- Apportionment is the process of dividing the 435 memberships, or seats, in the U.S. House of Representatives among the 50 states. At the conclusion of each decennial census, the results are used to calculate the number of seats to which each state is entitled.

- **Results were released on April 26, 2021**
- **Subjects include:**
  - Resident population
  - Overseas population
  - Apportionment population
- **Geography:** 50 states, the District of Columbia (DC), and Puerto Rico
- **Disclosure avoidance:** Results do not undergo disclosure avoidance

United States® Census Bureau

Redistricting File (Public Law 94-171)

- Public Law 94-171 directs the Census Bureau to provide data to the governors and legislative leadership in each of the 50 states for redistricting purposes. This product is the first file released that includes demographic and housing characteristics.

- **Results were released on August 12, 2021 (Summary Files) and September 16, 2021 (data.census.gov)**
- **Subjects include:**
  - Voting age
  - Race
  - Hispanic or Latino origin
  - Housing occupancy
  - Group quarters (GQ) population by major GQ type
- **Lowest level of geography:** Census Block
- **Disclosure avoidance:** Differentially private TopDown Algorithm (TDA)

United States® Census Bureau

Demographic Profile

- This product will provide select demographic and housing characteristics about local communities in a streamlined, easy to use format.

- **Expected release date:** May 2023
- **Subjects include:**
  - Sex by 5-year age groups
  - Median age by sex
  - Race
  - Hispanic or Latino origin
  - Relationship to householder
  - GQ population
  - Household type
  - Housing occupancy
  - Housing tenure
- **Lowest level of geography:** Tract
- **Disclosure avoidance:** Differentially private TDA

United States® Census Bureau

Demographic and Housing Characteristics File (DHC)

- The DHC will include many of the demographic and housing tables previously included in 2010 Summary File 1 (2010 SF1). Some tables are repeated by race and ethnicity.

- **Expected release date:** May 2023
- **Subjects include:**
  - Sex by single year-of-age
  - Hispanic or Latino origin of householder by race of householder
  - GQ population by sex by age
  - Relationship by age for population under 18 years
  - Household type by relationship and presence of people of specific ages
  - Multigenerational households
  - Family type by presence of children
  - Tenure by household size
  - Tenure by household type by age of householder
  - Vacancy Status
- **Lowest level of geography:** Varies with many tables at Census Block
- **Disclosure avoidance:** Differentially private TDA

United States® Census Bureau

Detailed Demographic and Housing Characteristics File A (Detailed DHC-A)

- Detailed DHC-A includes population counts repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribal and village population groups
- **Expected release date:** Sept 2023
- **Subjects are repeated by detailed racial and ethnic groups:**
  - Total population
  - Sex by Age for Selected Age Categories
- **Proposed levels of geography:** Nation, State, County, Tract, Place, AIANNH areas
- **Disclosure avoidance:** Differentially private SafeTab-P algorithm

United States® Census Bureau

Detailed Demographic and Housing Characteristics File B (Detailed DHC-B)

- Detailed DHC-B includes household counts repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribal and village population groups

- **Expected release date:** TBD
- **Subjects are repeated by detailed racial and ethnic groups:**
  - Household Type
  - Tenure
- **Proposed levels of geography:** Nation, State, County, Tract, Place, AIANNH areas
- **Disclosure avoidance:** Differentially private SafeTab-H algorithm

United States® Census Bureau

Detailed DHC-B

Adaptive design for Household Type

HOUSEHOLD TYPE (UNIVERSE)
Universe: Households
Total

HOUSEHOLD TYPE (2 CATEGORIES)	
Universe: Households	
Total:	
	Family households
	Nonfamily households

HOUSEHOLD TYPE (6 CATEGORIES)		
Universe: Households		
Total:		
	Family households:	
		Married couple family
		Other family
	Nonfamily households:	
		Householder living alone
		Householder not living alone

HOUSEHOLD TYPE (8 CATEOGIRES)			
Universe: Households			
Total:			
	Family households:		
		Married couple family	
		Other family:	
			Male householder, no spouse present
			Female householder, no spouse present
	Nonfamily households:		
		Householder living alone	
		Householder not living alone	

Detailed DHC-B

Adaptive Design for Tenure

TENURE (UNIVERSE)
Universe: Occupied housing units
Total

TENURE (3 CATEGORIES)	
Universe: Occupied housing units	
Total:	
	Owned with a mortgage or a loan
	Owned free and clear
	Renter occupied

Supplemental Demographic and Housing Characteristics File (S-DHC)

- S-DHC includes counts of children and people in households by certain characteristics, including average tables. Many tables are repeated by race and ethnicity.

- **Expected release date:** TBD
- **Subjects include:**
  - Average household size by age*
  - Household type for the population in households
  - Household type by relationship for the population under 18 years*
  - Population in families by age*
  - Average family size*
  - Family type and age for own children under 18 years
  - Total population in occupied housing units by tenure*
  - Average household size of occupied housing units by tenure*
- **Proposed levels of geography:** Currently being revised from initial proposal
- **Disclosure avoidance:** Differentially private PHSafe algorithm

**Indicates the table is proposed to be repeated by major Hispanic origin and race groups*

United States® Census Bureau

S-DHC Example

HOUSEHOLD TYPE FOR THE POPULATION IN HOUSEHOLDS		
Universe: Population in households		
Total:		
	In married couple household	
		Opposite-sex married couple
		Same-sex married couple
	In cohabiting couple household	
		Opposite-sex cohabiting couple
		Same-sex cohabiting couple
	Male householder, no spouse or partner present:	
		Living alone
		Living with others
	Female householder, no spouse or partner present:	
		Living alone
		Living with others

Questions?

United States® Census Bureau

Update on The 2020 Census Detailed Demographic and Housing Characteristics File A (Detailed DHC-A)

Roberto Ramirez
U.S. Census Bureau

United States® Census Bureau

Differences Between the TopDown and SafeTab-P Differentially Private Algorithms

TopDown **Redistricting Data (P.L. 94-171), DHC, Demographic Profile**	**SafeTab-P** **Detailed DHC-A**
Algorithm produces privacy-protected microdata	Algorithm directly produces privacy-protected tabulations
All geographies aggregate as expected	There is no requirement that geographies aggregate as expected
When aggregating data, the statistical noise generally cancels out and the statistics become *more* accurate	When aggregating data, it generally becomes more variable the more you aggregate
Consistency across data products	Not consistent with other 2020 Census data products
Overall accuracy can be targeted, but the exact levels of accuracy cannot be known in advance *(except in the case of the Redistricting Data (P.L. 94-171) Summary File)*	All margins of error are determined in advance and met 95% of the time
Does not use adaptive design	Uses adaptive design to determine the amount of data provided

United States® Census Bureau

About the Detailed DHC-A

- Subjects repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribes and villages:
  - Total population
  - Sex by age for selected age categories

- Geographic levels included:
  - Nation
  - State
  - County
  - Tracts
  - Places
  - American Indian/Alaska Native/Hawaiian Home Land (AIANNH) areas

- Planned for release in September 2023

United States® Census Bureau

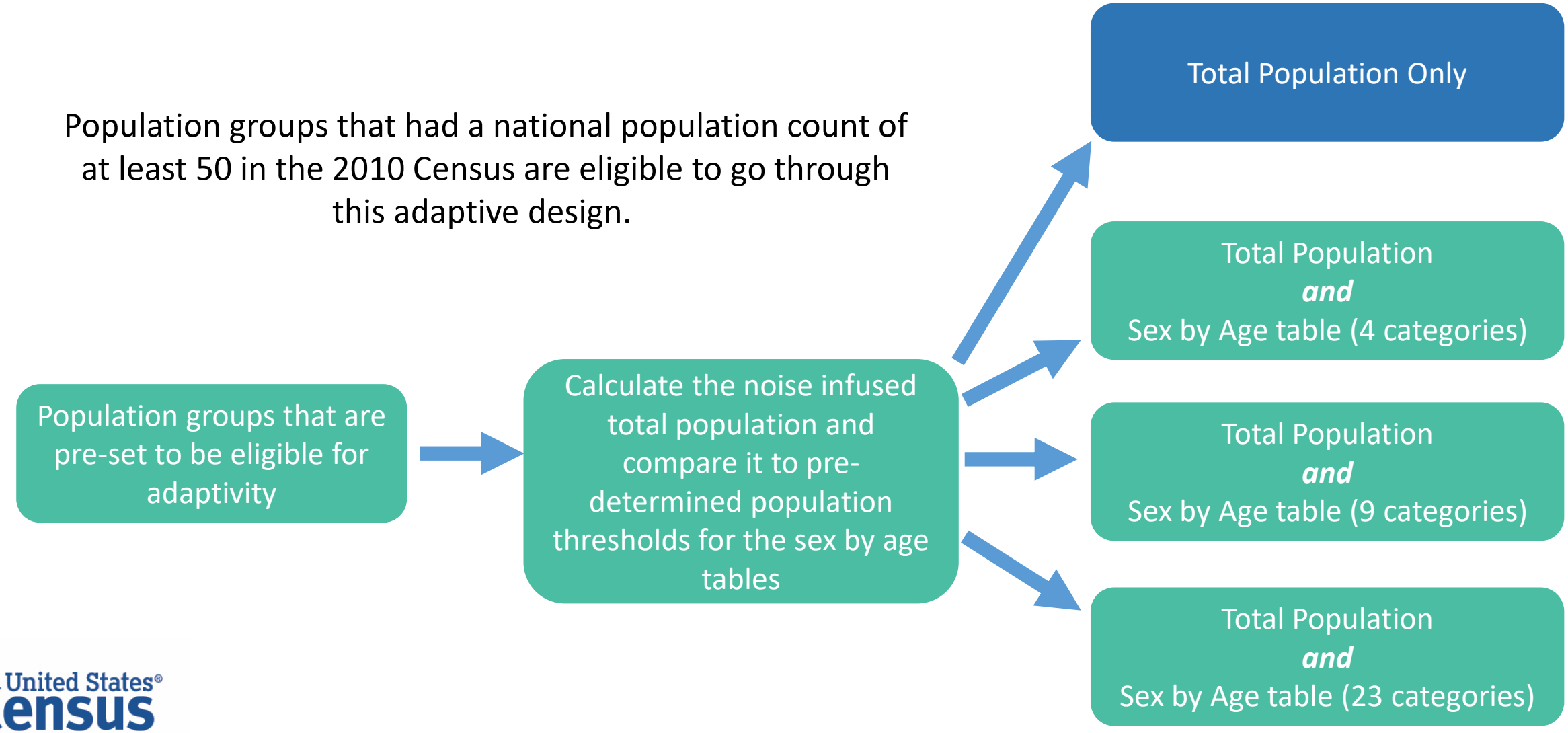
Using Adaptive Design to Produce the Detailed DHC-A

Detailed groups with a national population count less than 50 in the 2010 Census are pre-set to receive total population only in the Detailed DHC-A.

United States® Census Bureau

Using Adaptive Design to Produce the Detailed DHC-A

Population groups that had a national population count of at least 50 in the 2010 Census are eligible to go through this adaptive design.



United States® Census Bureau

Sex by Age
23 categories

Sex x Age(23)		
Age Group	**Male**	**Female**
Under 5 years	146	147
5 to 9 years	130	131
10 to 14 years	109	107
15 to 17 years	79	77
18 and 19 years	68	97
20 years	34	89
21 years	64	76
22 to 24 years	150	151
25 to 29 years	235	265
30 to 34 years	237	302
35 to 39 years	230	374
40 to 44 years	177	356
45 to 49 years	187	372
50 to 54 years	139	275
55 to 59 years	84	151
60 and 61 years	27	53
62 to 64 years	30	60
65 and 66 years	7	15
67 to 69 years	12	27
70 to 74 years	12	25
75 to 79 years	13	12
80 to 84 years	3	7
85 years and over	2	3

Sex by Age
9 categories

Sex x Age(9)		
Age Group	**Male**	**Female**
Under 5 years	146	147
5 to 17 years	318	315
18 to 24 years	316	413
25 to 34 years	472	567
35 to 44 years	407	730
45 to 54 years	326	647
55 to 64 years	141	264
65 to 74 years	31	67
75 years and over	18	22

Sex by Age
4 categories

Sex x Age(4)		
Age Group	**Male**	**Female**
Under 18 years	464	462
18 to 44 years	1,195	1,710
45 to 64 years	467	911
65 years and over	49	85

United States® Census Bureau

Note: These are fictitious data for illustration purposes.

Detailed DHC-A Minimum Noise Infused Population Counts and Margins of Error (MOE) by Geography

	Detailed groups		Regional groups	
Most comprehensive table type produced	**Nation & state (MOE=3)**	**Sub-state & AIANNH (MOE=11)**	**Nation & state (MOE=50)**	**Sub-state (MOE=50)**
Total count only	0-499	22-999	0-4,999	94-4,999
Sex by age – 4 categories	500-999	1,000-4,999	5,000-19,999	5,000-19,999
Sex by age – 9 categories	1,000-6,999	5,000-19,999	20,000-149,999	20,000-149,999
Sex by age – 23 categories	7,000+	20,000+	150,000+	150,000+

Note: The listed population thresholds are applied to the population counts after they have been processed by the approved differential privacy mechanism.
Note: MOE refers to the margin of error

United States® Census Bureau

Overview of Feedback on the Proof of Concept

- Commenters overall pleased to receive tract data. Only two requested more granular geographies (blocks and/or block groups)
- Commenters overall pleased with use of adaptive design and lowered thresholds from 2010

Institution Type	Count
Total	**15**
Non-profit	6
Local government	4
Academic institution	3
State government	1
Federal government	1

United States® Census Bureau

Commenter Recommendations on the Proof of Concept

Recommendations	Count (of 15)
Release detailed guidance on Detailed DHC-A limitations and uses	11
Adjust regional groupings	6
Make margins of error visible in data products	5
Continue/increase outreach	4
Eliminate postprocessing suppression	2
Release new metrics	1
Offer only age tables (not by sex)	1
Change thresholds for sex by age data	1
Increase thresholds	1

United States® Census Bureau

Questions?

United States® Census Bureau

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session III: An Overview of the 2020 Census Disclosure Avoidance System

Michael Hawes
Research and Methodology

United States® Census Bureau

Disclosure Avoidance for the 2020 Census

The 2020 Census improves on the noise injection methods of the 1990-2010 Censuses by employing a mathematical framework known as Differential Privacy (DP) to assess and quantify disclosure risk and confidentiality protection.

Every individual that is reflected in a particular statistic contributes towards that statistic's value.

Every statistic that you publish "leaks" a small amount of private information.

DP as a framework allows you to assess each individual's contribution to the statistic, and to measure (and thus, limit) how much information about them will leak.

United States® Census Bureau

The 2020 Census Disclosure Avoidance System (DAS)

TopDown Algorithm (TDA)

Produces privacy-protected microdata (Microdata Detail File) that is ingested by Decennial tabulation system

- Redistricting Data (P.L. 94-171) Summary File
- Demographic Profile
- Demographic and Housing Characteristics File (DHC)
- Congressional District Summary Files

SafeTab PHSafe

Produce privacy-protected tabulations

- Detailed DHC-A
- Detailed DHC-B
- Supplemental DHC

United States® Census Bureau

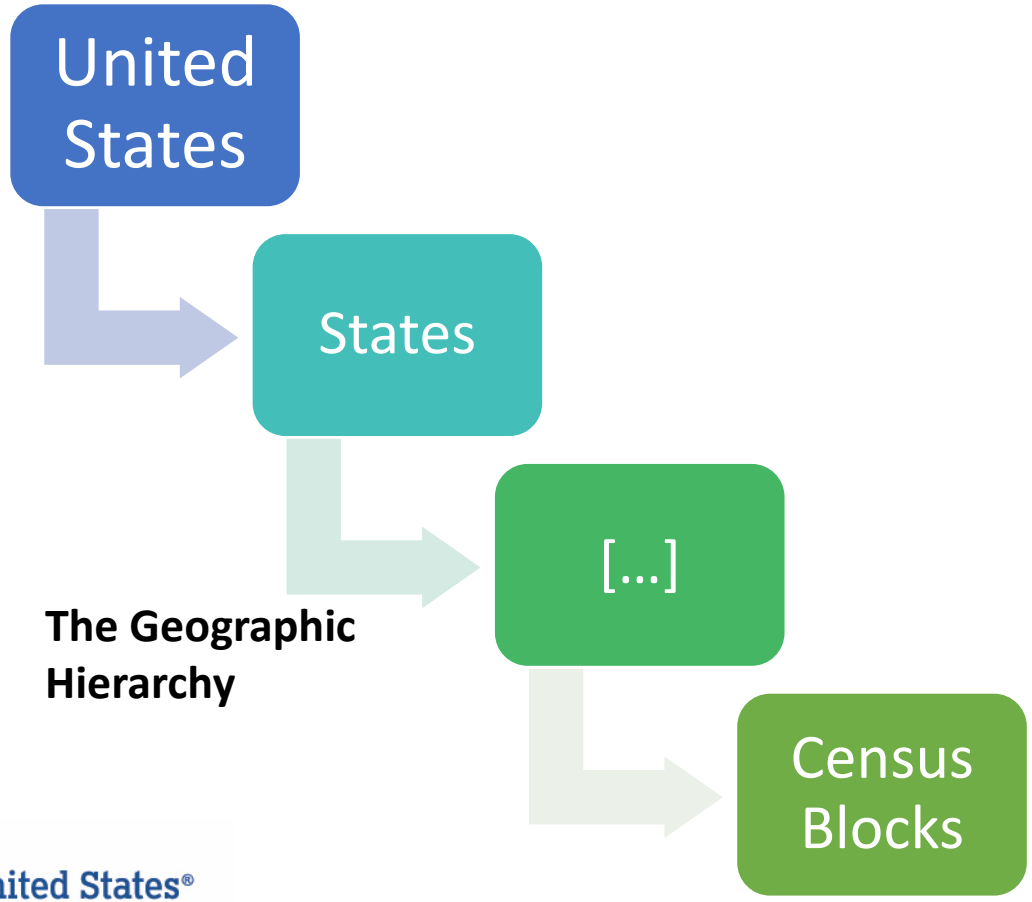
The TopDown Algorithm

*A histogram, in this context, is a tabular representation of the microdata with counts of records for each possible combination of values for each attribute in the microdata.

For complete details see: 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*. (June) https://doi.org/10.1162/99608f92.529e3cb9

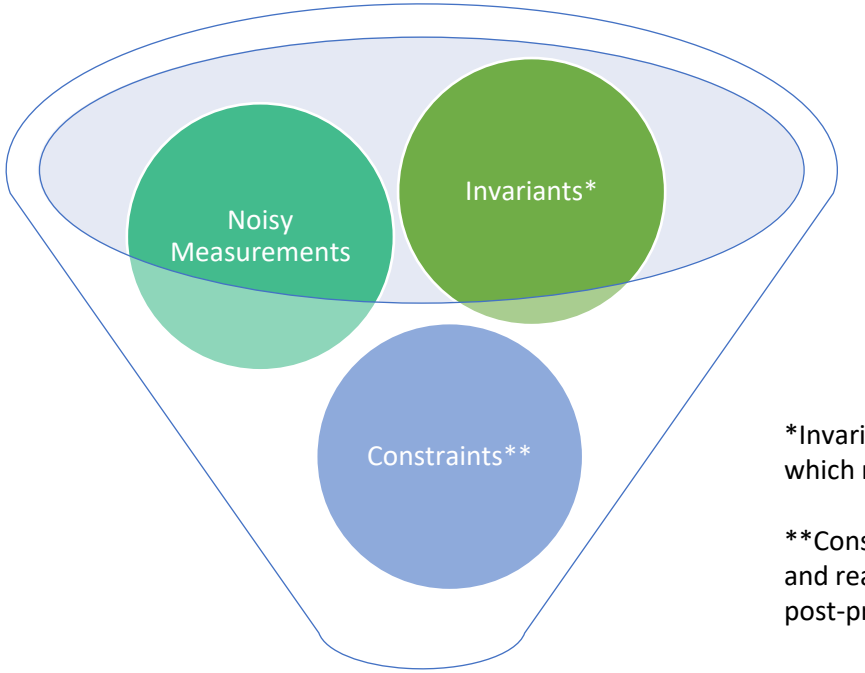
United States® Census Bureau

The TopDown Algorithm



The Geographic Hierarchy

At each geographic level:



*Invariants are counts to which no noise is added.

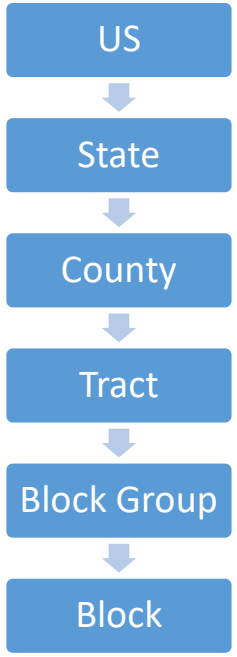
**Constraints are consistency and reasonableness rules the post-processing must impose.

Internally consistent histogram

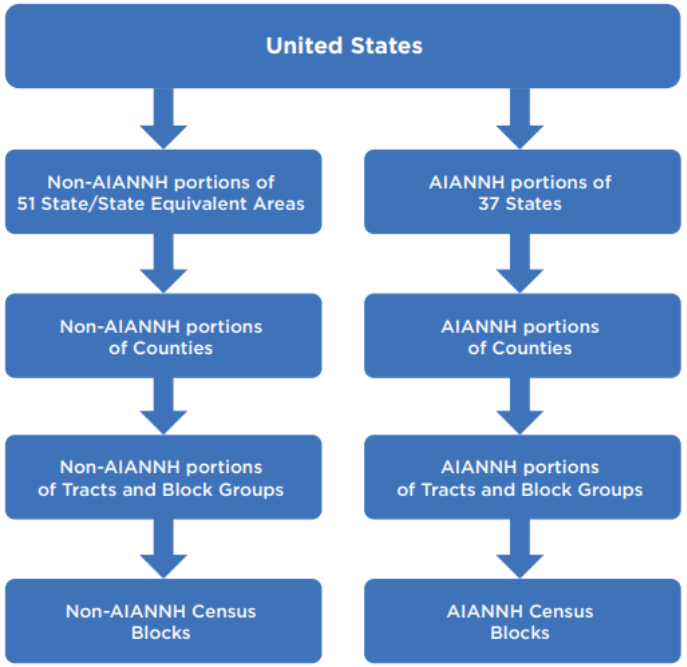
United States® Census Bureau

The Geographic Hierarchy ("Spine")

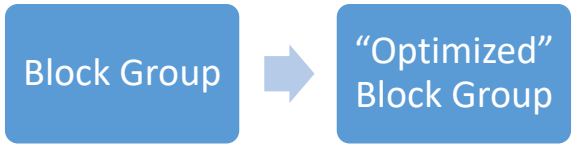
Standard Spine Tabulation Hierarchy



TDA's American Indian/Alaska Native/Native Hawaiian (AIANNH) Spine for Redistricting



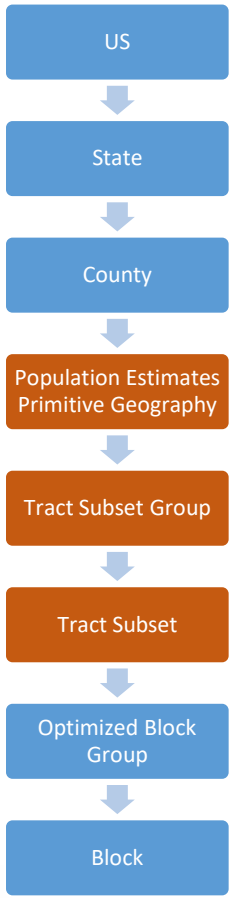
Geographic Optimization
(for Redistricting Data)



Reconfigured TDA's definition of block groups to optimize accuracy in statistics for certain types of geographies, including Minor Civil Divisions, Places, and individual AIANNH areas. Optimized block groups are used inside TDA. Tabulation block groups from the standard hierarchy are used for all published data tables.

United States® Census Bureau

Population Estimates Primitive Geographies



Population Estimates Primitive Geographies are the most granular geographic areas that are required in order to derive tables for every geography for which official Population Estimates are produced.

The Population Estimates Primitive Geographies form a complete, mutually exclusive partition of the U.S.

Tract Subsets are defined as the intersection of Population Estimates Primitive Geographies with census tabulation tracts.

Tract Subset Groups are defined as the union of multiple tract subsets that are all within the same Population Estimates primitive geography.

For the DHC, the TDA Geographic Hierarchy was further modified to include “Population Estimates Primitive Geographies” on the spine.

United States® Census Bureau

Queries and Privacy-loss Budget Allocation

Global *rho*	2.56
Global *epsilon*	17.90
delta	10^{-10}

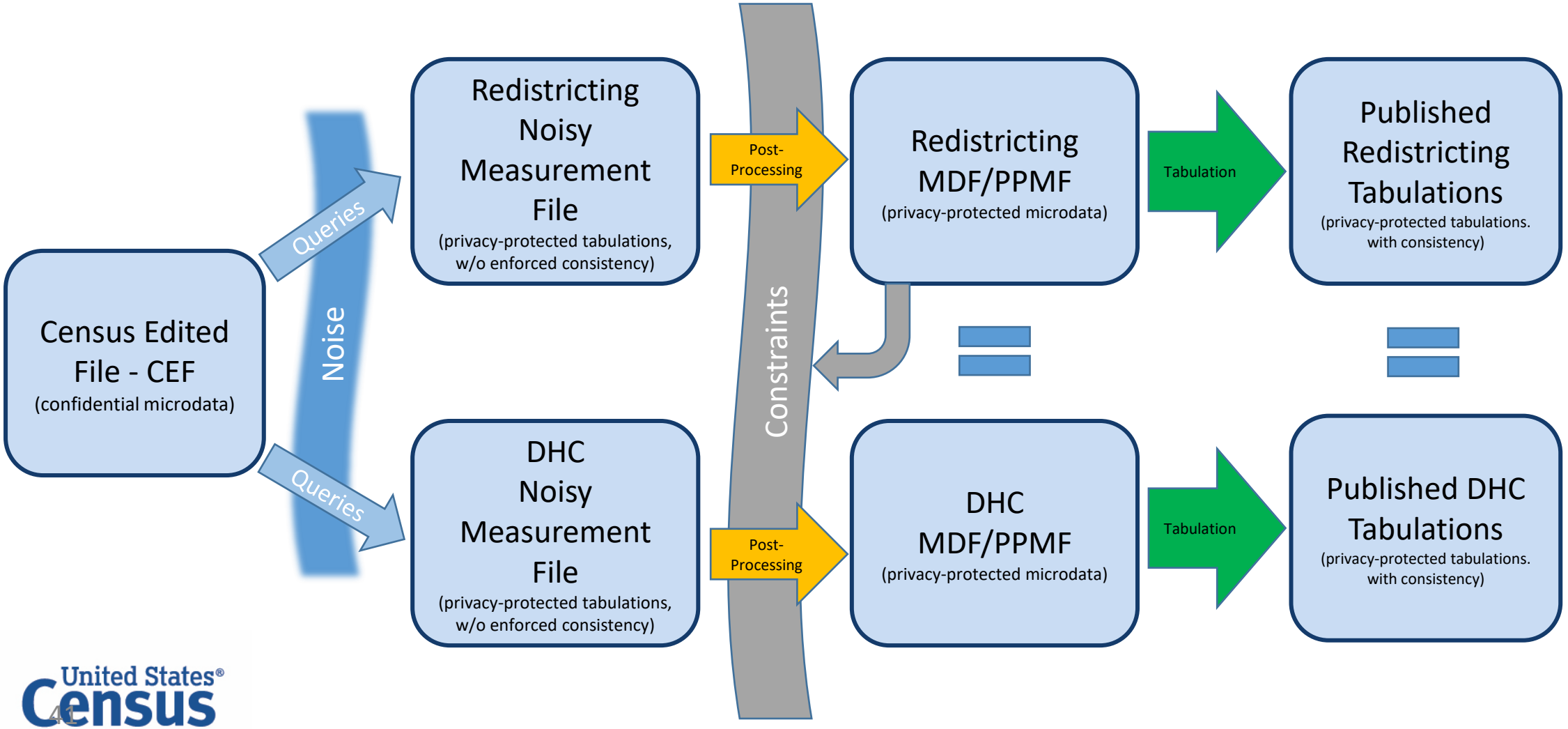
	***rho* Allocation by Geographic Level**
US	2.54%
State	35.13%
County	10.91%
Tract	16.76%
Optimized Block Group*	30.64%
Block	4.03%

Production settings for the 2020 Census Redistricting Data (P.L. 94-171) Summary File (Persons tables P1-P5)

Query	**Per Query *rho* Allocation by Geographic Level**					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)	[redacted]	32.35%	8.32%	6.40%	12.75%	0.00%
CENRACE (63 cells)	0.03%	0.05%	0.03%	0.03%	0.02%	0.01%
HISPANIC (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHINSTLEVELS (3 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHGQ (8 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HISPANIC*CENRACE (126 cells)	0.08%	0.10%	0.07%	7.90%	7.89%	0.02%
VOTINGAGE*CENRACE (126 cells)	0.08%	0.10%	0.07%	0.08%	0.07%	0.02%
VOTINGAGE*HISPANIC (4 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE*HISPANIC*CENRACE (252 cells)	0.27%	0.29%	0.27%	0.27%	0.18%	0.07%
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	1.99%	1.97%	2.01%	1.97%	9.63%	3.88%

United States® Census Bureau

Noisy Measurement Files (NMFs), Privacy-Protected Microdata Files (PPMFs), Published Tabulations



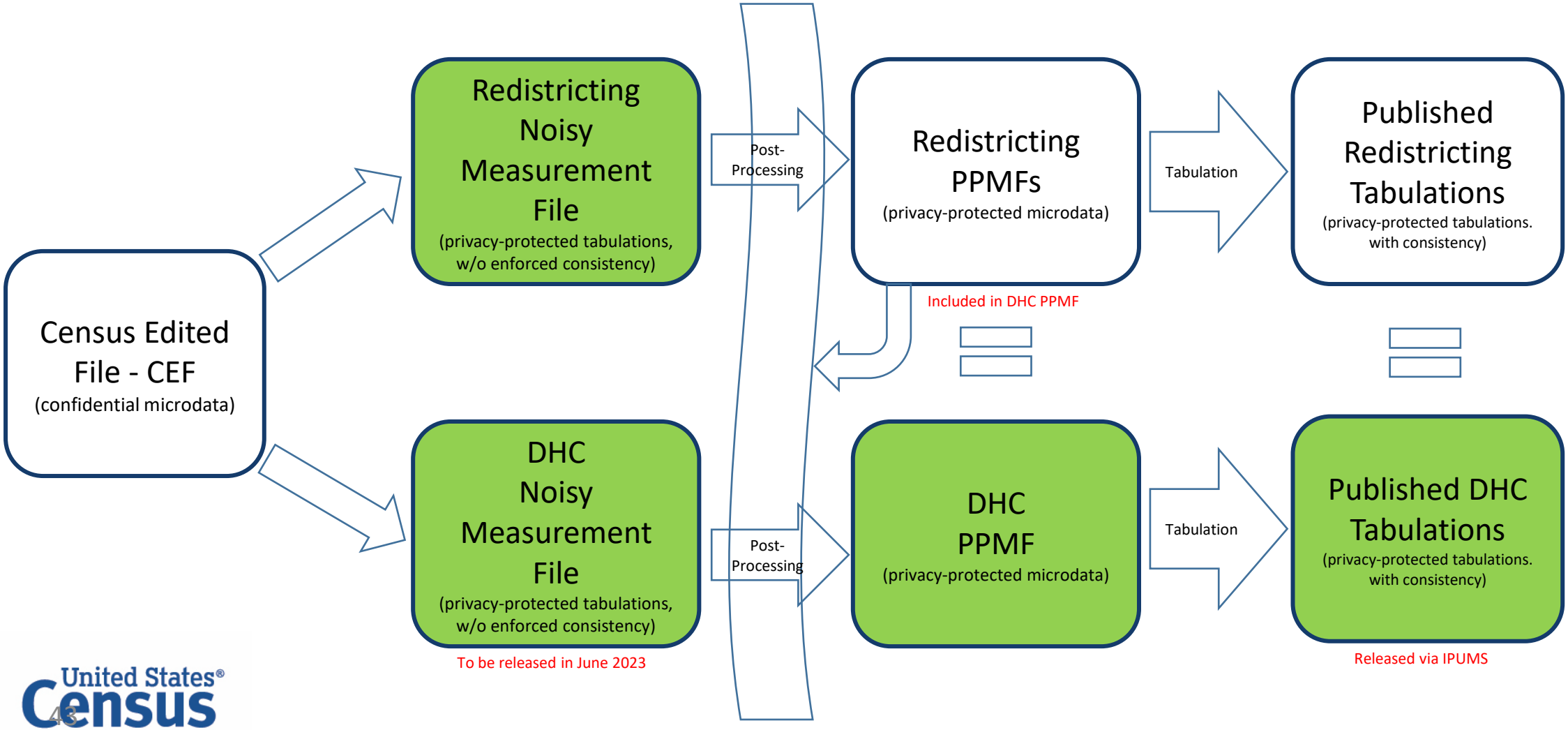
Components of the 2010 Demonstration Data Products Suite – Redistricting and Demographic and Housing Characteristics File – Production Settings (2023-04-03) **(2010 DDPS)**

(2010 Census data processed through the 2020 DAS at production settings)

- 2010 DDPS Fact Sheet
- Detailed Summary Metrics (and Metrics Overview)
- Privacy-Protected Microdata File (PPMF)
- DHC Tabulations (via IPUMS)
- Privacy-loss Budget (PLB) Allocations
- Noisy Measurement File (NMF)

United States® Census Bureau

2010 DDPS NMFs, PPMF, Tabulations

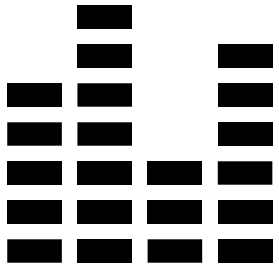


Should I Use the NMF, the PPMF, or the Tabulations?

- There are two sources of error in the published statistics (PPMF and Tabulations):

Differentially private noise

- Unbiased
- Known distribution
- Reflected in the noisy measurements

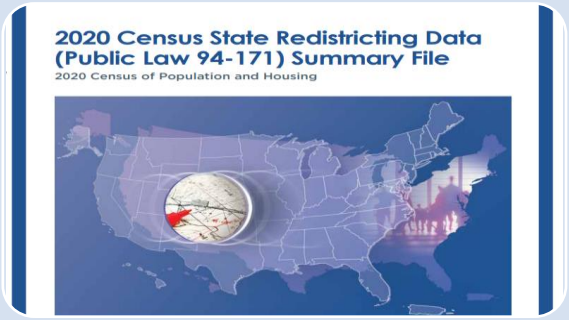


Post-processing

- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

United States® Census Bureau

Should I Use the NMF, the PPMF, or the Tabulations?



2020 Census Redistricting and DHC Tabulations

- Official 2020 Census Statistics
- Higher Accuracy (feature of TDA)
- Does include bias due to post-processing

2020 Census PPMF

- 100% microdata file
- Consistent with published tabulations
- Useful for special tabulations and microdata analysis

2020 Census NMF

- Can be used to produce unbiased estimates and confidence intervals
- Can be used to evaluate alternate post-processing mechanisms
- Research product

United States® Census Bureau

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session IV: Detailed Summary Metrics

Matthew Spence
Population Division

United States® Census Bureau

Fitness-for-Use: Detailed Summary Metrics

- One way to understand the anticipated fitness of use – Detailed Summary Metrics, which compare tabulations from 2010 Census data run through Disclosure Avoidance System (DAS) to published 2010 data.
- These comparisons (e.g., absolute value of difference) are averaged across geographies in a geographic level (e.g., counties) to create estimates of accuracy (e.g., Mean Absolute Error or MAE) or bias or to count large outliers.
- We can use these metrics to think about the fitness-for-use of 2020 DHC data, since the same software and settings are used to produce those.

Fitness-for-Use: Detailed Summary Metrics

- Have released **9** versions of Detailed Summary Metrics that show how the DAS and its parameter settings have evolved over time due to internal and external feedback on use cases and data user needs.
- This release showcases improvements that were made between the last demonstration data release (August 2022) and the final code settings for the production run in November 2022.
- Based on these 2010 evaluations, there are acuracy improvements for:
  - Householder race
  - Presence and age of own children
  - Relationship to householder
  - Single year of age for children for counties and school districts
  - Age for persons in group quarters
  - Same-sex married and unmarried partners

Highlights from the Detailed Summary Metrics

- Looking beyond average error (MAE/MAPE):
  - 95th Percentile Absolute Error
  - 95th Percentile Absolute Percent Error

  } 95% of geographies are off by this much (or less).
- Comparing to average populations to provide context

Tenure Numbers are: 2022-03-25 / 2023-04-03 Universe: Occupied Housing Units. Geography: Census Tract

	Mean Population	Mean Absolute Error (MAE)	95th Percentile Absolute Error (95th AE)	Mean Absolute Percent Error (MAPE)	95th Percentile Absolute Percent Error (95th APE)
Owned with a mortgage	725	1.65 / 2.04	4.00 / 5.00	1% / 1%	1% / 1%
Owned free and clear	315	1.65 / 2.05	4.00 / 5.00	2% / 2%	4% / 5%
Renter-occupied	528	1.66 / 2.04	4.00 / 5.00	1% / 1%	3% / 3%

- Acceptable errors even for the 95th percentile case

Updated Metrics on 2010 DDPS Accuracy

Coupled Household Type	Numbers are: 2022-03-25 / 2023-04-03				Universe: Households. Geography: County
	Mean Population	MAE	95th AE	MAPE	95th APE
Opposite-sex married couple household	17,980	5.83/ 4.58	15 / 12	0% / 0%	1% / 1%
Same-sex married couple household	111	3.62/ 2.38	9 / 6	21% / 15%	80% / 60%
Opposite-sex unmarried partner household	2,177	5.12 / 3.68	13 / 10	2% / 1%	8% / 6%
Same-sex unmarried partner household	176	3.61 / 2.29	10 / 6	33% / 24%	200% / 100%

Updated Metrics on 2010 DDPS Accuracy

Presence of Own Children Under 6	2022-03-25 / 2023-04-03				Universe: Households. Geography: Various
	Mean Population	MAE	95th AE	MAPE	95th APE
All counties	4,727	26.47 / 7.41	76 / 18	3% / 2%	12% / 5%
All incorporated places	496	9.67 / 4.37	31 / 13	22% / 18%	100% / 75%
All elementary school districts	479	15.14 / 7.76	51 / 25	17% / 14%	60% / 53%
All secondary school districts	2,036	27.81 / 10.99	100 / 32	7% / 5%	27% / 23%
All unified school districts	1,263	19.62 / 9.60	60 / 27	8% / 6%	25% / 20%

Tenure By Race of Householder	2022-03-25 / 2023-04-03			Universe: Occupied Housing Units. Geography: County	
	Mean Population	MAE	95th AE	MAPE	95th APE
Owner occupied					
White alone	20,187	83.61 / 5.54	258 / 15	2% / 0%	5% / 1%
Black or African American alone	1,992	48.09 / 4.10	157 / 12	238% / 32%	1,200% / 200%
American Indian and Alaska Native alone	162	18.08 / 3.45	60 / 10	81% / 23%	300% / 100%
Asian alone	856	31.60 / 2.96	112 / 9	227% / 35%	948% / 200%
Native Hawaiian and Other Pacific Islander alone	20	5.10 / 1.51	18 / 5	128% / 65%	400% / 200%
Some Other Race alone	629	15.29 / 3.13	53 / 8	66% / 26%	270% / 117%
Two or More Races	332	19.37 / 3.38	63 / 9	47% / 12%	188% / 44%
Renter occupied					
White alone	8,370	57.34 / 4.59	179 / 12	3% / 0%	11% / 2%
Black or African American alone	2,504	33.81 / 3.43	102 / 9	167% / 28%	800% / 200%
American Indian and Alaska Native alone	137	14.67 / 3.06	54 / 8	90% / 29%	375% / 140%
Asian alone	618	17.62 / 2.41	62 / 7	164% / 38%	700% / 200%
Native Hawaiian and Other Pacific Islander alone	26	6.61 / 1.53	27 / 5	146% / 66%	400% / 200%
Some Other Race alone	936	12.75 / 3.04	44 / 8	46% / 21%	200% / 100%
Two or More Races	368	16.26 / 3.08	54 / 8	55% / 16%	214% / 62%

Single Year of Age for Ages 0-15	2022-03-25 / 2023-04-03				Universe: Persons. Geography: County
	Mean Population	MAE	95th AE	MAPE	95th APE
Under 1 year old	1,255	14.62 / 5.58	42 / 14	6% / 3%	25% / 14%
1 year old	1,266	14.93 / 5.47	44 / 14	6% / 4%	22% / 14%
2 years old	1,304	14.56 / 5.47	42 / 13	6% / 3%	22% / 13%
3 years old	1,311	13.15 / 5.96	38 / 14	6% / 4%	20% / 14%
4 years old	1,293	13.20 / 6.15	37 / 15	6% / 4%	21% / 15%
5 years old	1,291	16.50 / 5.93	50 / 14	7% / 4%	23% / 14%
6 years old	1,294	16.47 / 5.96	51 / 15	7% / 4%	25% / 14%
7 years old	1,282	16.42 / 6.10	47 / 15	7% / 4%	24% / 15%
8 years old	1,287	17.07 / 5.93	48 / 15	7% / 4%	24% / 14%
9 years old	1,320	17.01 / 5.98	49 / 15	7% / 4%	24% / 14%
10 years old	1,328	16.69 / 6.06	49 / 16	7% / 3%	24% / 13%
11 years old	1,309	16.55 / 6.02	46 / 16	7% / 3%	23% / 13%
12 years old	1,306	16.88 / 5.92	50 / 15	7% / 3%	25% / 13%
13 years old	1,310	16.82 / 6.02	48 / 15	7% / 4%	24% / 13%
14 years old	1,325	16.87 / 6.11	49 / 16	7% / 4%	23% / 13%
15 years old	1,350	5.11 / 3.98	13 / 11	3% / 2%	11% / 8%

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session V: Disclosure Avoidance Briefs

New Directions Based on Recommendations

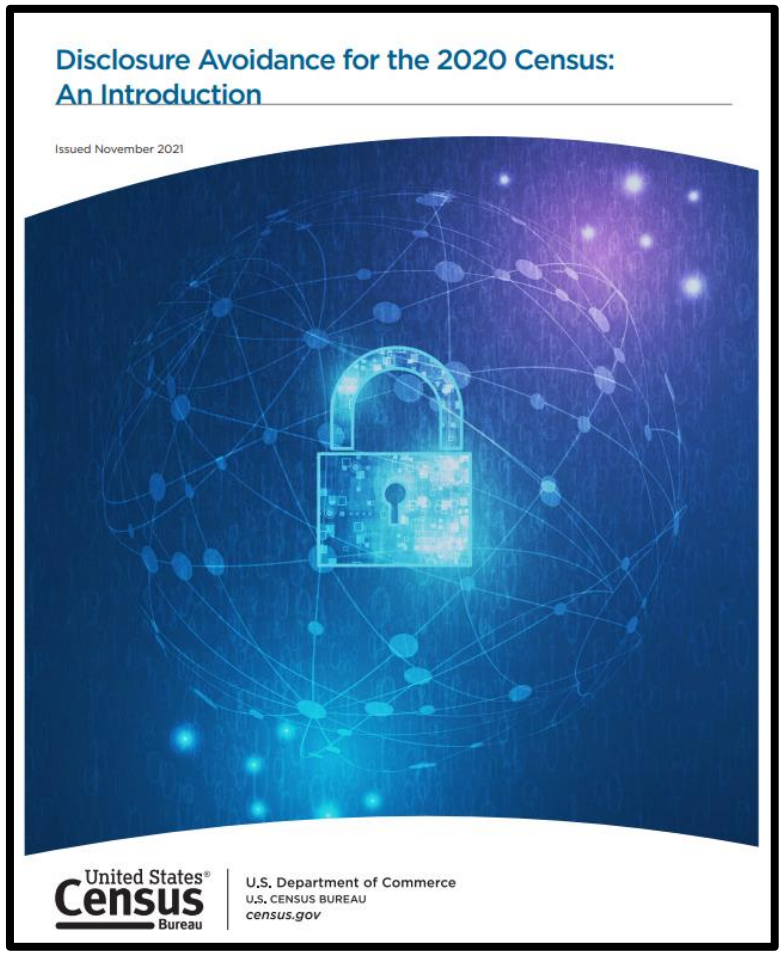
Beth Jarosz
Population Reference Bureau

Cynthia Davis Hollingsworth
Decennial Census Management Division

United States® Census Bureau

Background

- Released “Disclosure Avoidance for the 2020 Census: An Introduction” in November 2021.
  - Described how the disclosure avoidance system was implemented in the 2020 Census Redistricting Data (P.L. 94-171) Summary File.
  - Target Audience: Data users who want a high-level understanding of disclosure avoidance modernization – what they "need to know" to work with the data.
- Received National Advisory Committee on Racial, Ethnic and Other Populations (NAC) and Census Scientific Advisory Committee (CSAC) recommendations for future materials.
- Also gathered information through a survey (April 2022, 12 respondents) and listening sessions at other meetings.



United States® Census Bureau

Common Themes in Feedback

- Overarching themes from feedback received:
  - **Short**, goal no more than 6 pages (but most slightly longer)
  - **Focused**, highlighting key points as opposed to a narrative format, and
  - **Practical**, with information that is easy to understand and found quickly
- Add hyperlinks for those interested in more detailed information.
- Include more graphics and examples.
- Include more equity-focused examples.
- Build in time for external review.

Specific Recommendations: Examples

- Provide concise definition of differential privacy/DAS
- Discuss total error framework
- Address the strengths and limitations of disclosure avoidance methods
- Explain what is published (e.g., location, age, race, sex, relationship to householder), thus at risk of disclosure
- Explain what is not published (e.g., name, birth date, telephone number)
- Explain differential privacy in the context of computer power and technical sophistication to combat the increased threat of reidentification attacks
- Describe real-world impacts

New Direction

- Developing a series of briefs to meet the needs of data users
- Using topics that were suggested by users
- Implementing external review
- Using more graphics

United States® Census Bureau

Briefs on Disclosure Avoidance for the 2020 Census

Briefs 1-3

#1: Disclosure Avoidance and the 2020 Redistricting Data

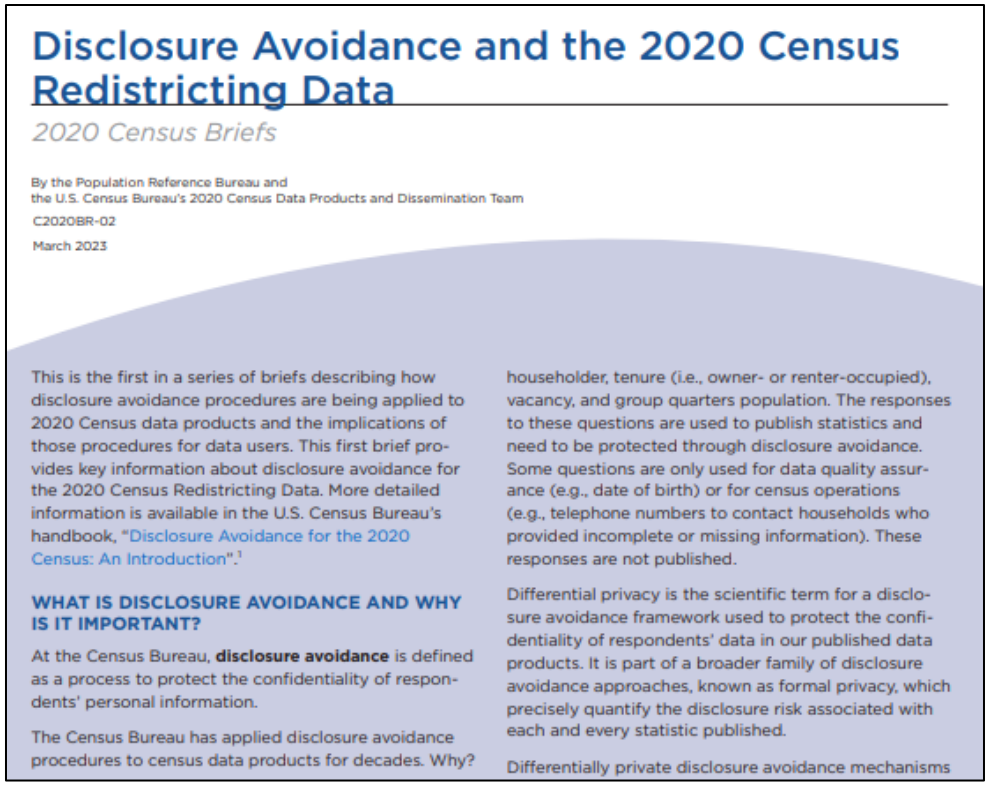
A summary of key points from the previously released Redistricting Data (P.L. 94-171) Summary File handbook.
https://www.census.gov/library/publications/2023/decennial/c2020br-02.html

#2: Why the Census Bureau Chose Differential Privacy

An explanation of how the Census Bureau selected differential privacy over other disclosure avoidance systems.
https://www.census.gov/library/publications/2023/decennial/c2020br-03.html

#3: Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works

A description of the TopDown Algorithm and a concise definition of differential privacy.
https://www.census.gov/library/publications/2023/decennial/c2020br-04.html



Disclosure Avoidance and the 2020 Census Redistricting Data

2020 Census Briefs

By the Population Reference Bureau and
the U.S. Census Bureau's 2020 Census Data Products and Dissemination Team

C2020BR-02

March 2023

This is the first in a series of briefs describing how disclosure avoidance procedures are being applied to 2020 Census data products and the implications of those procedures for data users. This first brief provides key information about disclosure avoidance for the 2020 Census Redistricting Data. More detailed information is available in the U.S. Census Bureau's handbook, "Disclosure Avoidance for the 2020 Census: An Introduction".[1]

WHAT IS DISCLOSURE AVOIDANCE AND WHY IS IT IMPORTANT?

At the Census Bureau, **disclosure avoidance** is defined as a process to protect the confidentiality of respondents' personal information.

The Census Bureau has applied disclosure avoidance procedures to census data products for decades. Why?

householder, tenure (i.e., owner- or renter-occupied), vacancy, and group quarters population. The responses to these questions are used to publish statistics and need to be protected through disclosure avoidance. Some questions are only used for data quality assurance (e.g., date of birth) or for census operations (e.g., telephone numbers to contact households who provided incomplete or missing information). These responses are not published.

Differential privacy is the scientific term for a disclosure avoidance framework used to protect the confidentiality of respondents' data in our published data products. It is part of a broader family of disclosure avoidance approaches, known as formal privacy, which precisely quantify the disclosure risk associated with each and every statistic published.

Differentially private disclosure avoidance mechanisms

United States® Census Bureau

Briefs on Disclosure Avoidance for the 2020 Census

Topics for Briefs 4-6

Brief #4 (Target release - TBD)

DHC – relevant content, including:

- Explain some of the ways stakeholders use DHC data generally.
- Discuss use cases.
- Demonstrate the potential impact of disclosure avoidance methods on various priority use cases/case studies for DHC.

Brief #5 (Target release - Summer 2023)

Detailed DAS methodology (SafeTab-P) for the Detailed DHC-A Product.

Brief #6 (Target release - TBD)

Explain the Total Uncertainty framework including:

- Highlights from the variability paper.
- The degree of error that's introduced with noise infusion.
- Describe the multiple sources of error with examples for each.

United States® Census Bureau

Briefs on Disclosure Avoidance for the 2020 Census

Topics for Briefs 7-8

Brief #7 (Target release - TBD)
Detailed DAS methodology (SafeTab-H) for the Detailed DHC-B Product.

Brief #8 (Target release - TBD)
Detailed DAS methodology (PHSafe) for the Supplemental DHC Product.

Additional guidance and documentation:

https://www.census.gov/programs-surveys/decennial-census/technical-documentation/complete-technical-documents.html

United States® Census Bureau

External Review

- Each brief will have at least 2 external reviewers.
- Reviewers will represent a variety of different data user communities including various geographies and subjects of expertise, including:

  - Children
  - Demography
  - Disability
  - Economics
  - Equity
  - Geography
  - Health
  - Housing
  - Older Adults
  - Privacy
  - Schools/Education
  - Statistics
  - Computer Science

United States® Census Bureau

Thank You!

Beth Jarosz and Cynthia Davis Hollingsworth

bjarosz@prb.org
cynthia.davis.hollingsworth@census.gov

United States® Census Bureau

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session VI: Constructing Uncertainty Estimates for TDA-based Data Products

Robert Ashmead
Research and Methodology

United States® Census Bureau

Acknowledgement

The research presented here is a collaborative effort of a larger team:

Sallie Keller, John Abowd, Michael Hawes, Wendy Martinez, Ryan Cummings, Philip Leclerc, Beth Jarosz, Matthew Spence, Jonathan Spader, Cynthia Hollingsworth, James Whitehorne, Pavel Zhuravlev, Tapan Nayak, Alexandra Krause, Joseph Schafer, Mary Pritts

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product.
*(*CBDRB-FY23-0240 and CBDRB-FY22-DSEP-004*)*

United States® Census Bureau

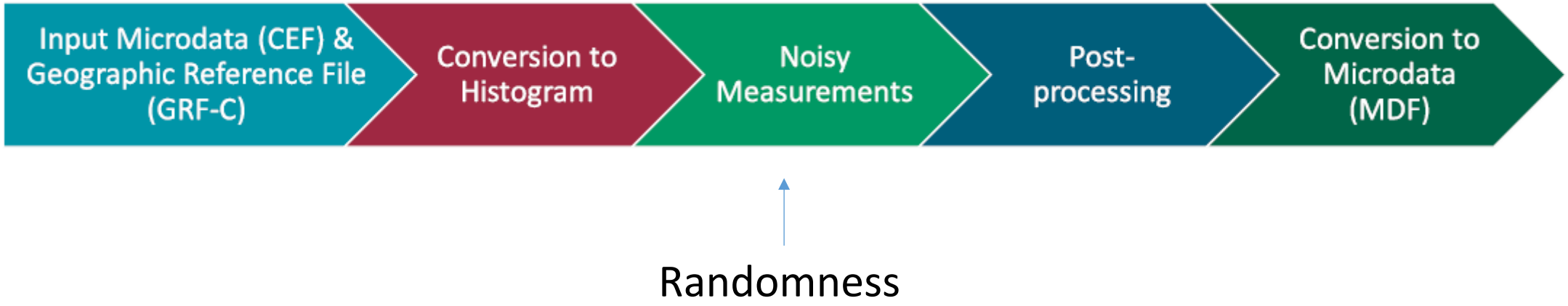
Background

- 2020 P.L.-94 Redistricting data products were created from the Disclosure Avoidance System's (DAS's) TopDown Algorithm (TDA) which used differential privacy (DP) to protect respondent information.
- Small amounts of "noise" (error) were infused into published tabulations
- Ideally, users can take into account the statistical properties of error to adapt their statistical analyses
  - **Goal will be to construct meaningful confidence intervals**

United States® Census Bureau

Outline of DAS Process

- CEF – Census Edited File
- MDF – Microdata Detail File



United States® Census Bureau

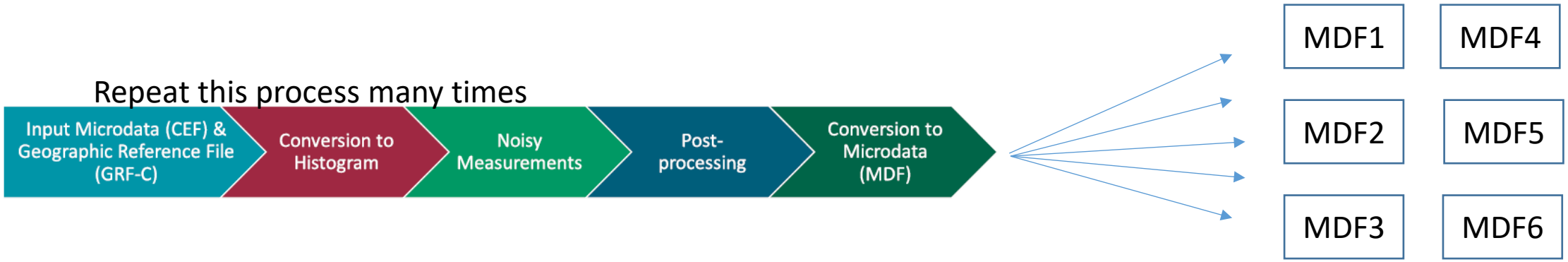
The Challenge

- The TDA enforces non-negativity, maintains integers with controlled rounding, and implements equality and inequality constraints
- Solution found by a complex numerical optimization algorithm
- As a result:
  - No closed-form variance/bias/root mean-squared error (RMSE) formulas for tabulation queries
  - The variance/bias/RMSE do not just depend on the randomness in the noisy measurements, but on the underlying query (e.g. how small or large it is)

United States® Census Bureau

Monte Carlo Simulation Methods

- One way of calculating the statistical properties of a random process is to simulate the process many times and analytically calculate the quantities of interest.



- Then calculate the variance, RMSE, bias of any tabulation empirically based on MDF1, MDF2, MDF3, ….

United States® Census Bureau

Why this doesn’t quite work with formal privacy methods

- Repeating the process counts against the privacy-loss budget for each run since we are generating noisy measurements based on the CEF
- Bias and RMSE calculations require knowing the CEF values
- What can we do instead?

United States® Census Bureau

Proposed Method: Approximate Monte Carlo

- Previously/alternatively called the “Parametric Bootstrap” method and explored using the 1940 Census data

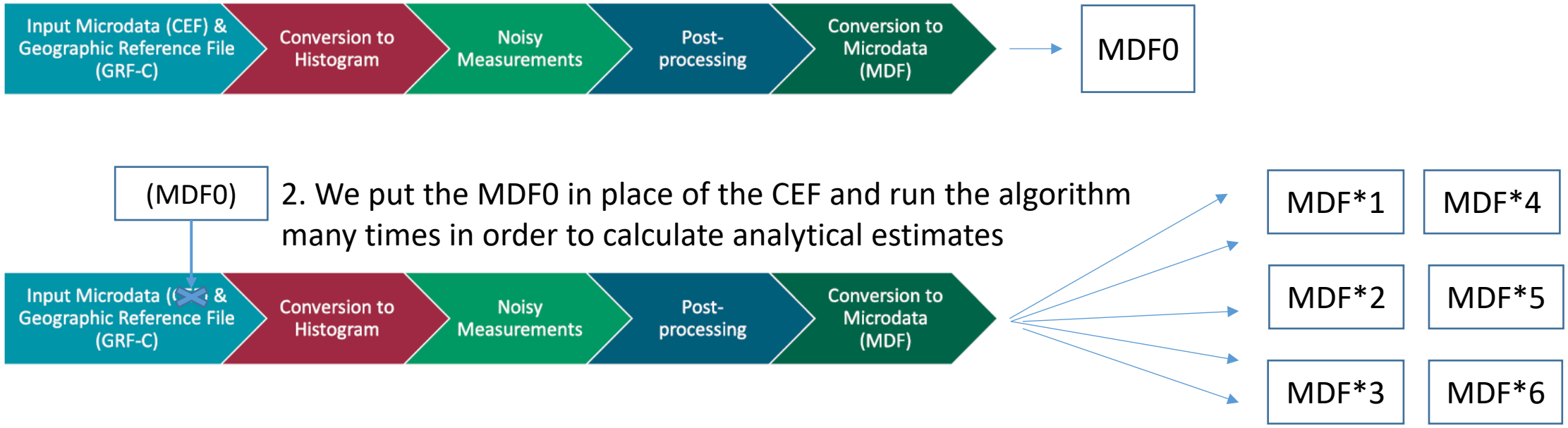
https://www.census.gov/content/dam/Census/newsroom/press-kits/2019/jsm/presentation-estimating-the-variance-of-complex-differentially-private-algorithms.pdf

https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/privacy-methods-2020-census.pdf (pg 95)

United States® Census Bureau

Basic Idea (visually)

1. The TDA was/is run which gives the published tabulations



Does this work for formally private methods?

- Since we are using the MDF0 quantities, we are free to generate noisy measurements based on them without counting any additional privacy-loss (post-processing property)

- For Bias and RMSE calculations, we know the MDF0 values and can freely use them since they are outputs of the TDA.

United States® Census Bureau

Why should this work?

- If the MDF0 is similar to the CEF, then the simulated quantities should be similar to those that would have come from the CEF
- Are the MDF0 and CEF close enough for this to work well?
  - One of our research questions

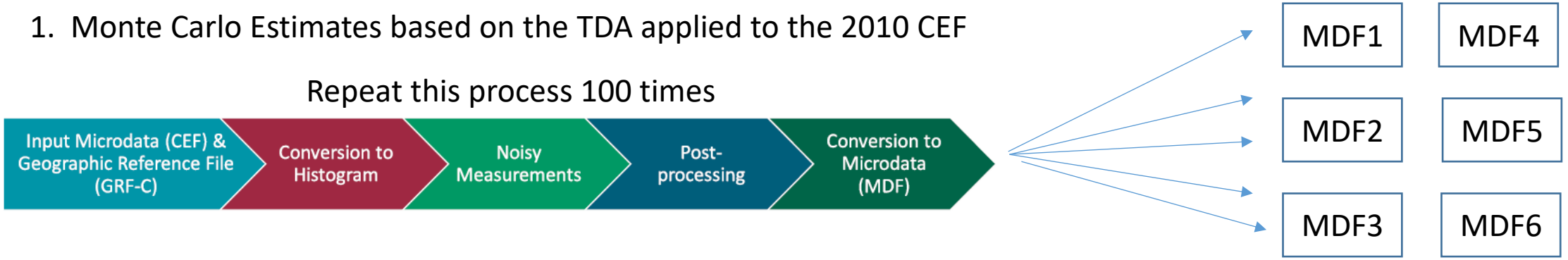
- Other Research Questions
  - How many simulations is enough?
  - Where are the geographic levels/queries where this doesn't work as well? What adjustments can we make to compensate?

United States® Census Bureau

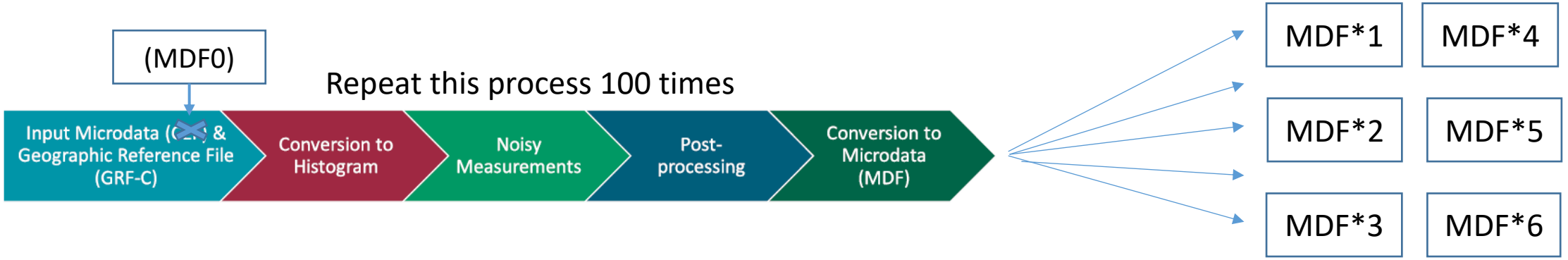
How do we check the proposed method?

1. Monte Carlo Estimates based on the TDA applied to the 2010 CEF

Repeat this process 100 times

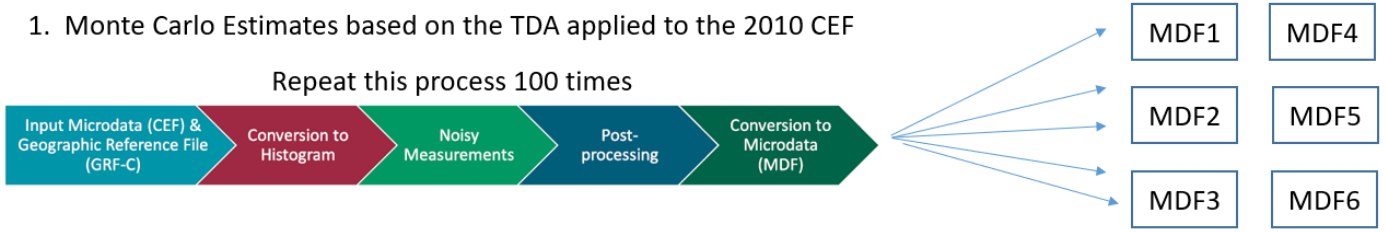


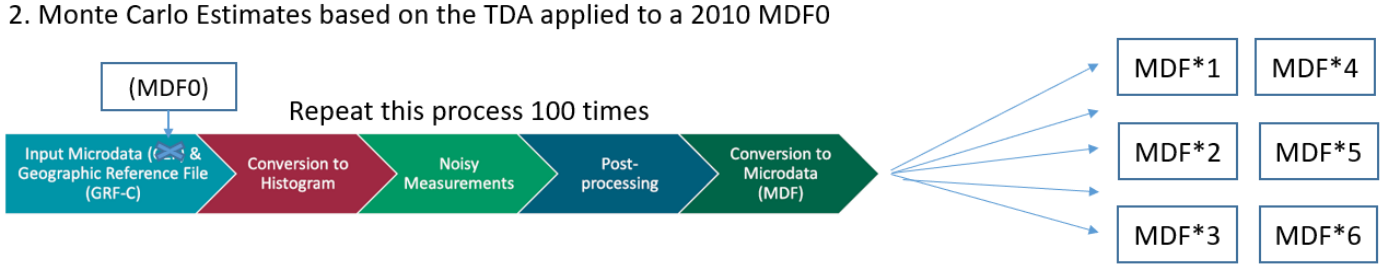
2. Monte Carlo Estimates based on the TDA applied to a 2010 MDF0



United States® Census Bureau

How do we check the proposed method?





3. Compare analytical estimates of bias/standard error/RMSE between methods 1 and 2. Ideally 2 approximates 1, for a given query/tabulation.

4. Construct confidence intervals using quantities from 2 and the MDF0 tabulation as the point estimate. Check the % of time that the CEF tabulation is contained inside the confidence interval.

United States® Census Bureau

Experiment Setup, 2010 Data

- 100 runs of the TDA from the CEF
- 100 runs of the TDA from MDF0
- 301 Queries from the Redistricting Data (P.L. 94-171) Summary File:
  - P1. Race
  - P2. Hispanic or Latino, and not Hispanic or Latino by Race
  - P3. Race for the Population 18 Years and Over
  - P4. Hispanic or Latino, and not Hispanic or Latino by Race for the Population 18 Years and Over
  - P5. Group Quarters Population by Major Group Quarters Type
  - H1. Occupancy Status (Housing)

United States® Census Bureau

Estimated Quantities

For a given query (tabulation) q() we calculate

$$RMSE_{CEF} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF_i) - q(CEF))^2}$$

$$RMSE_{MDF0} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF0_i^*) - q(MDF0))^2}$$

$$Bias_{CEF} = \frac{1}{n}\sum_{i=1}^{100} q(MDF_i) - q(CEF)$$

$$Bias_{MDF0} = \frac{1}{n}\sum_{i=1}^{100} q(MDF0_i^*) - q(MDF0)$$

United States® Census Bureau

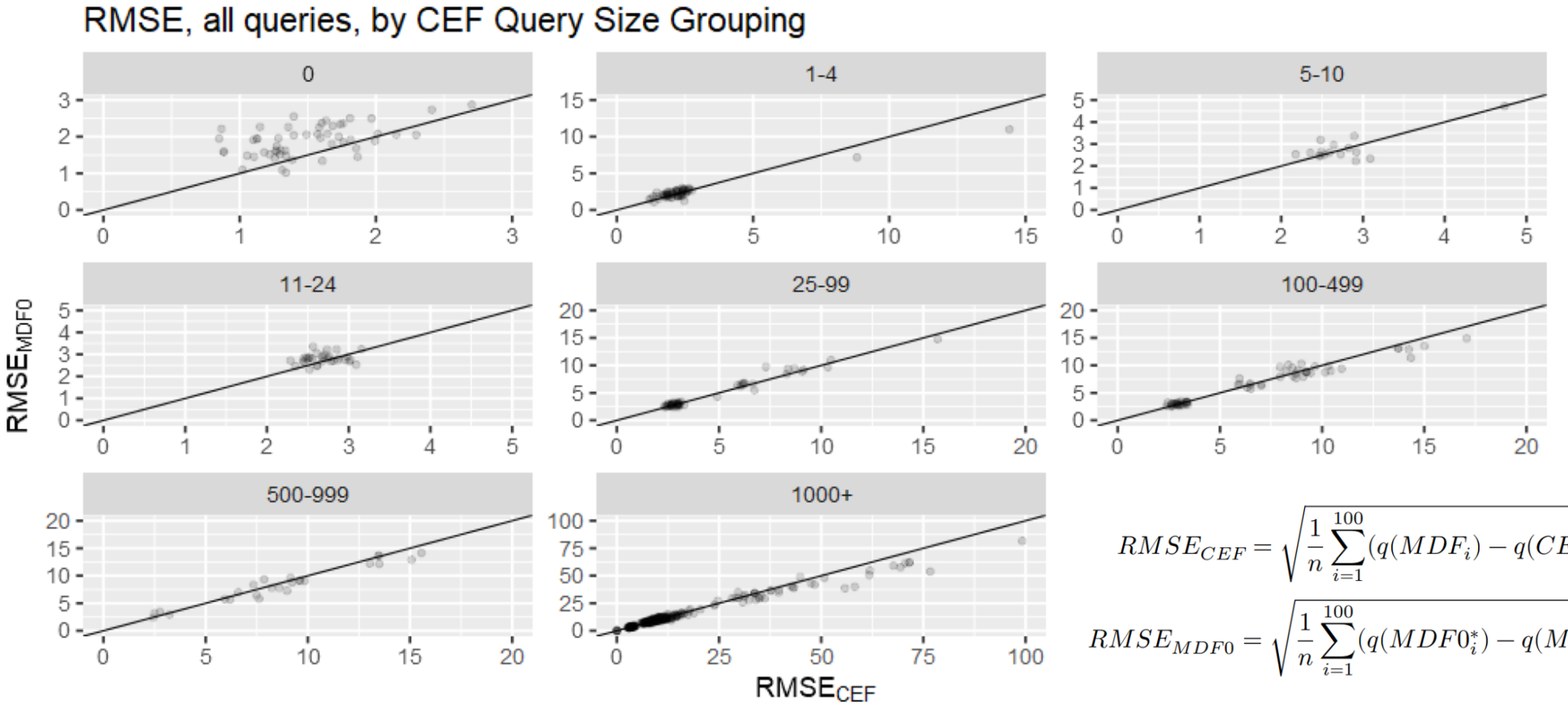
Estimated Quantities Continued

$$SE_{CEF} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF_i) - \overline{q(MDF)})^2}$$

$$SE_{MDF0} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF0_i^*) - \overline{q(MDF0^*)})^2}$$

United States® Census Bureau

National Level (US and PR) RMSE Estimation



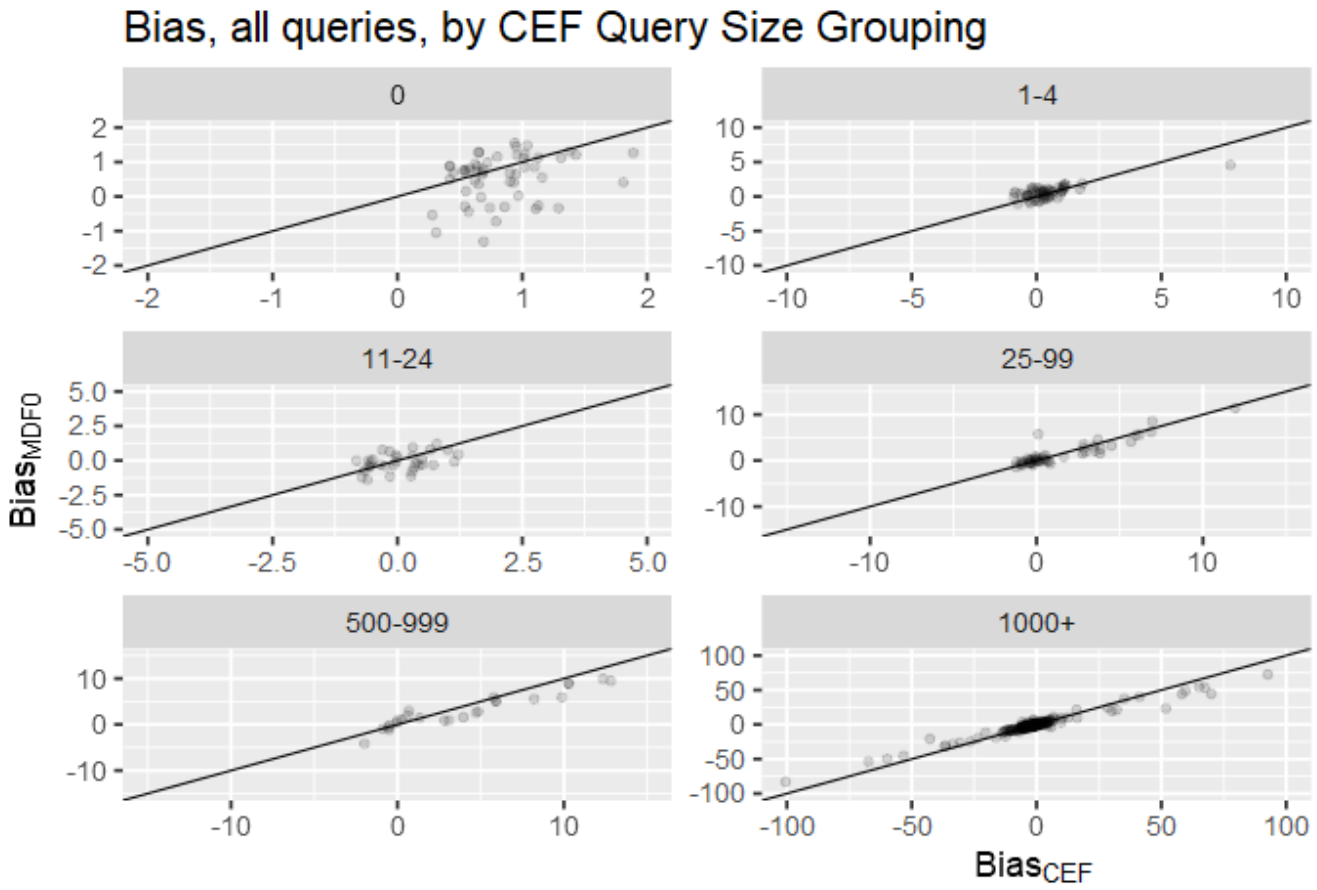
$$RMSE_{CEF} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF_i) - q(CEF))^2}$$

$$RMSE_{MDF0} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF0_i^*) - q(MDF0))^2}$$

United States® Census Bureau

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

National Level (US and PR) Bias Estimation

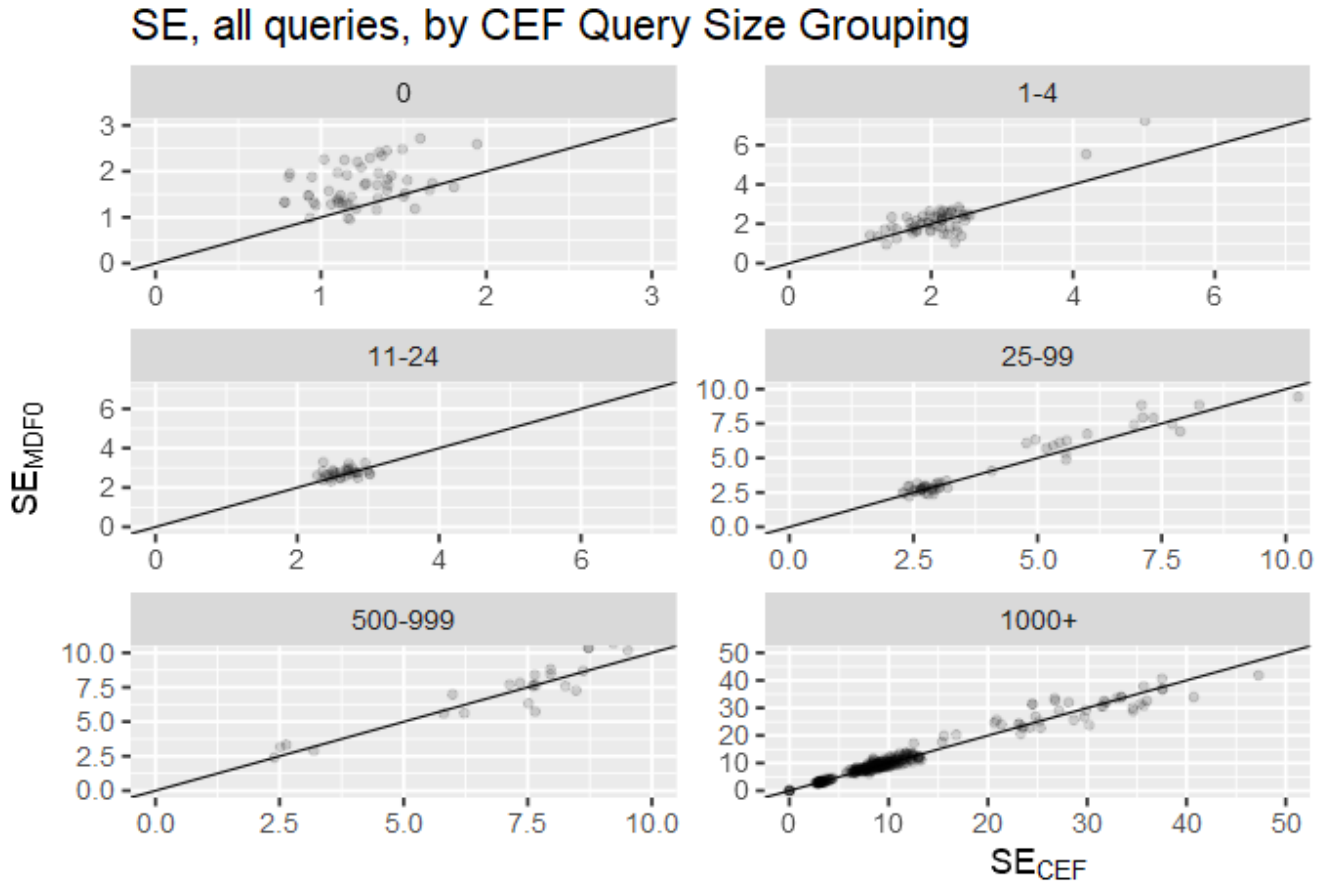


$$Bias_{CEF} = \frac{1}{n}\sum_{i=1}^{100} q(MDF_i) - q(CEF)$$

$$Bias_{MDF0} = \frac{1}{n}\sum_{i=1}^{100} q(MDF0_i^*) - q(MDF0)$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

National Level (US and PR) SE Estimation

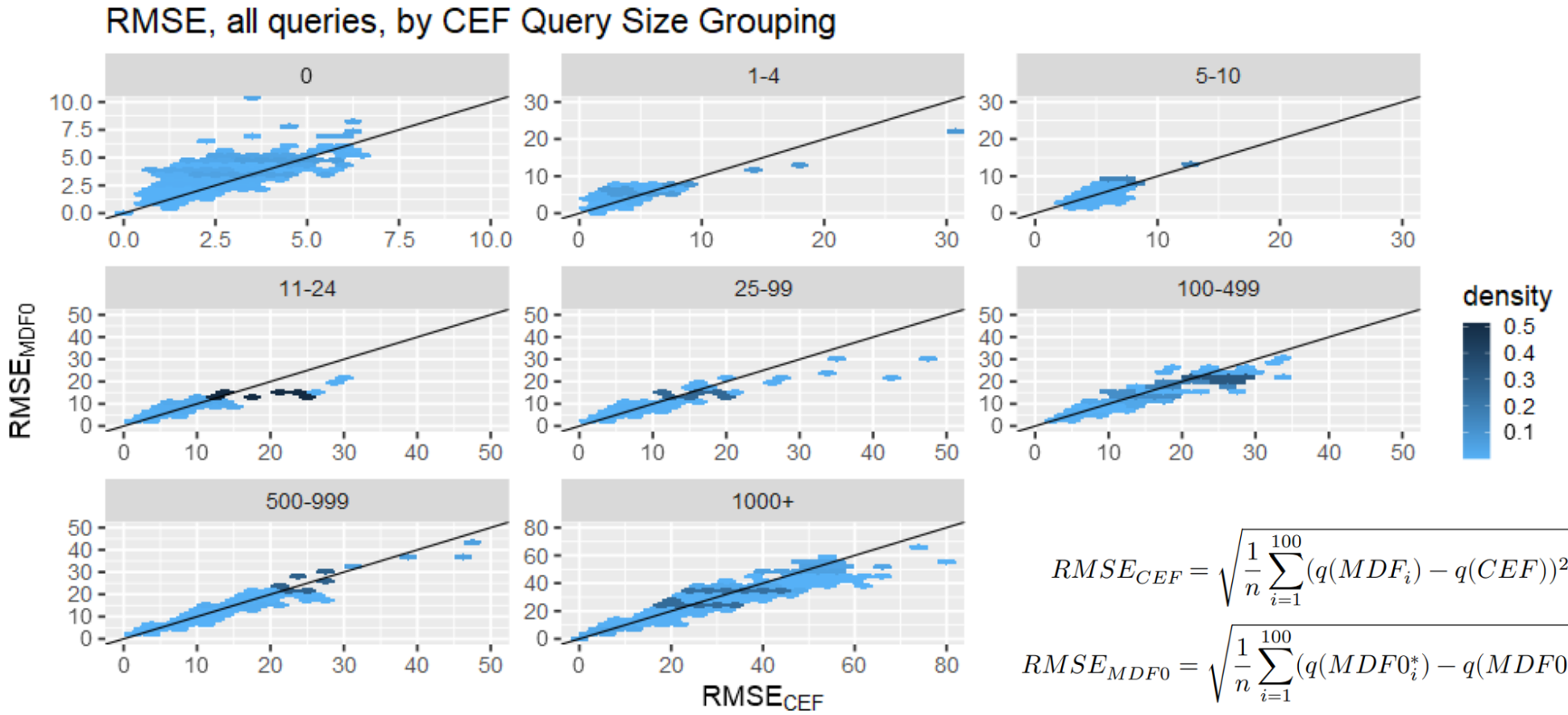


$$SE_{CEF} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF_i) - \overline{q(MDF)})^2}$$

$$SE_{MDF0} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF0_i^*) - \overline{q(MDF0^*)})^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

State Level RMSE Estimation

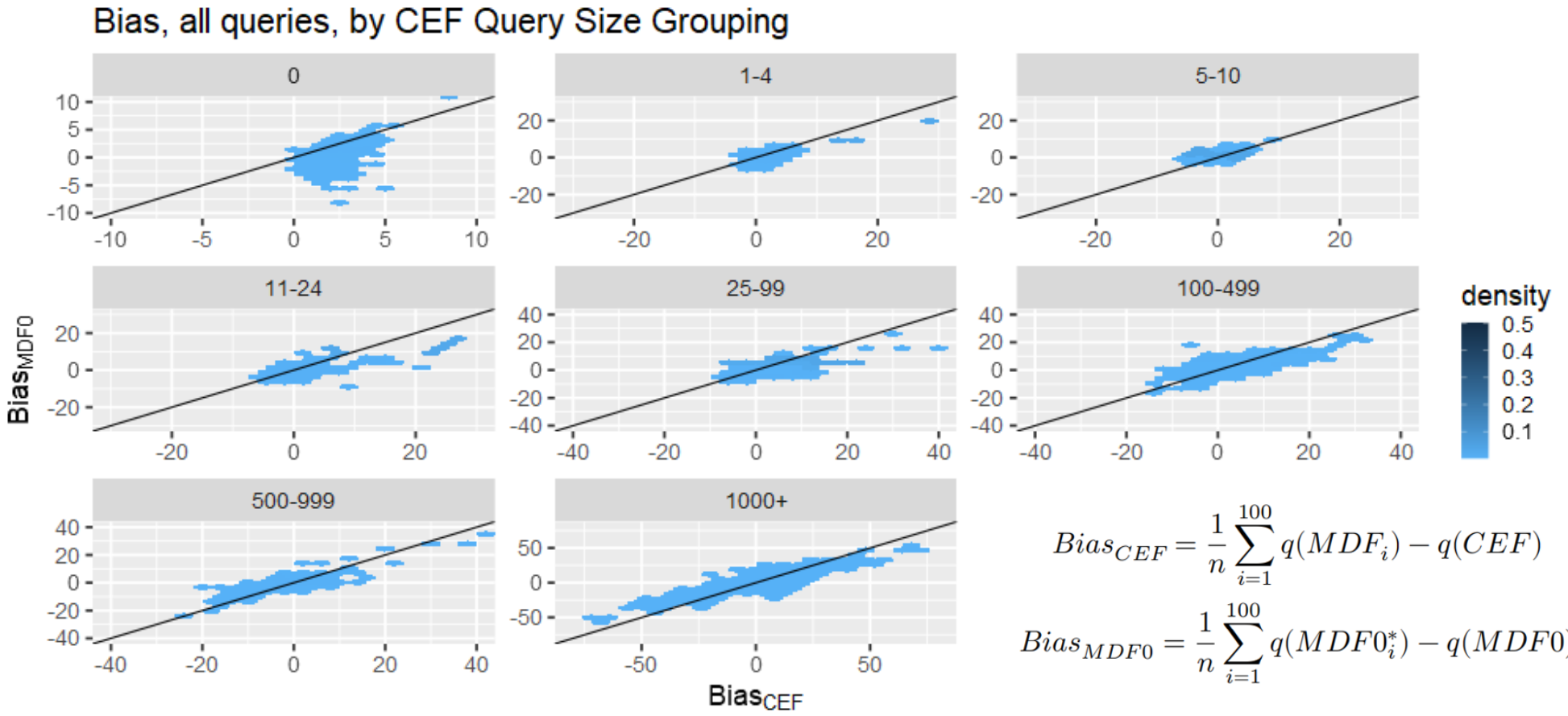


$$RMSE_{CEF} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF_i) - q(CEF))^2}$$

$$RMSE_{MDF0} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF0_i^*) - q(MDF0))^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

State Level Bias Estimation

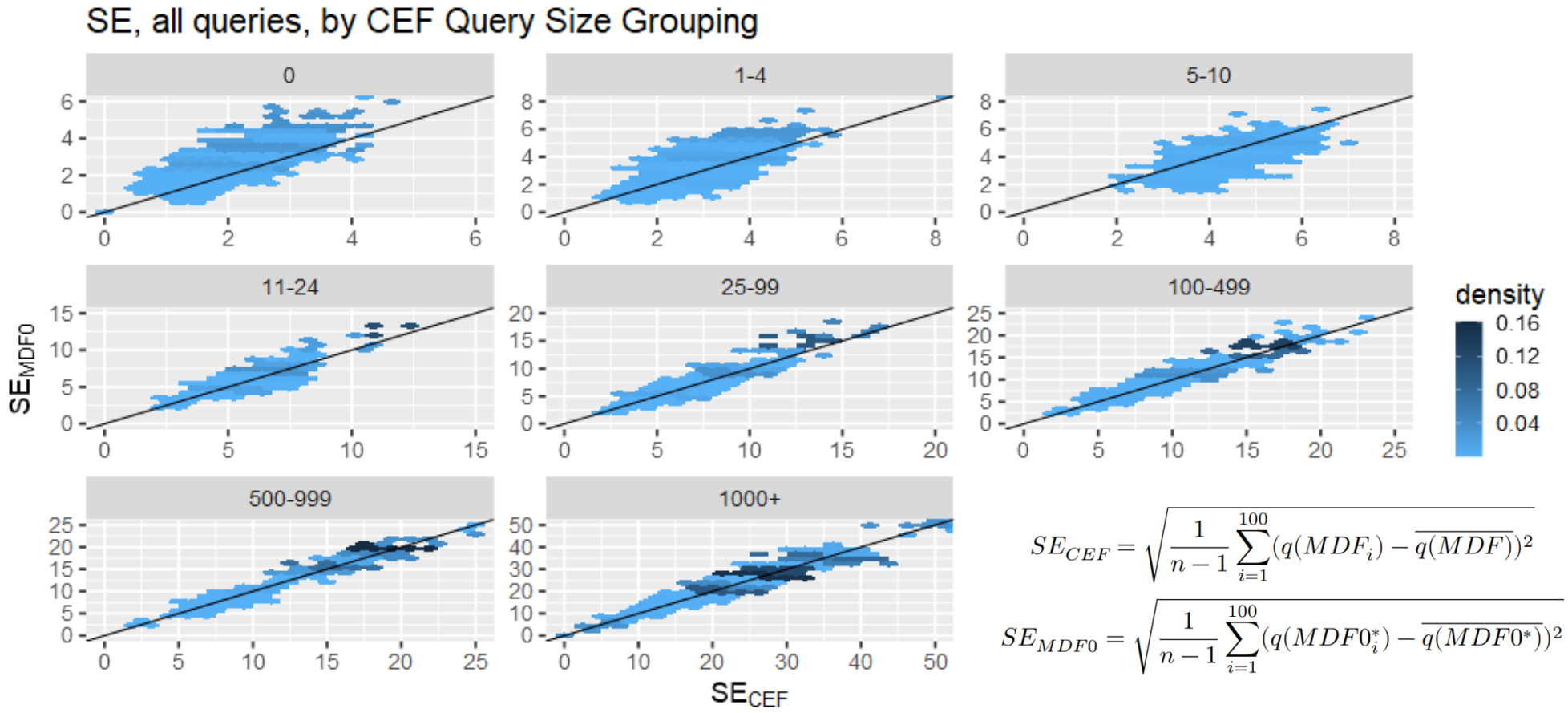


$$Bias_{CEF} = \frac{1}{n}\sum_{i=1}^{100} q(MDF_i) - q(CEF)$$

$$Bias_{MDF0} = \frac{1}{n}\sum_{i=1}^{100} q(MDF0_i^*) - q(MDF0)$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

State Level SE Estimation

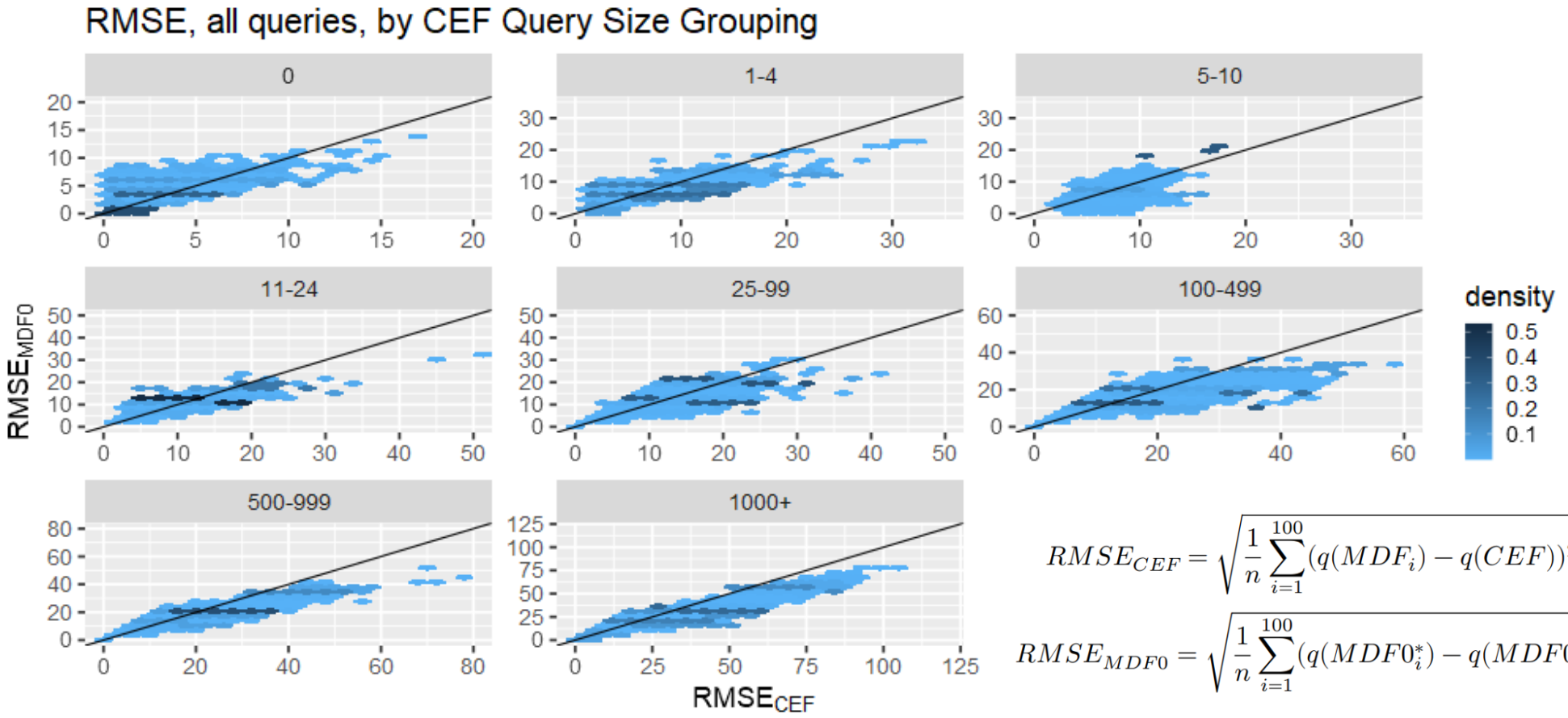


$$SE_{CEF} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF_i) - \overline{q(MDF)})^2}$$

$$SE_{MDF0} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF0_i^*) - \overline{q(MDF0^*)})^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

County Level RMSE Estimation

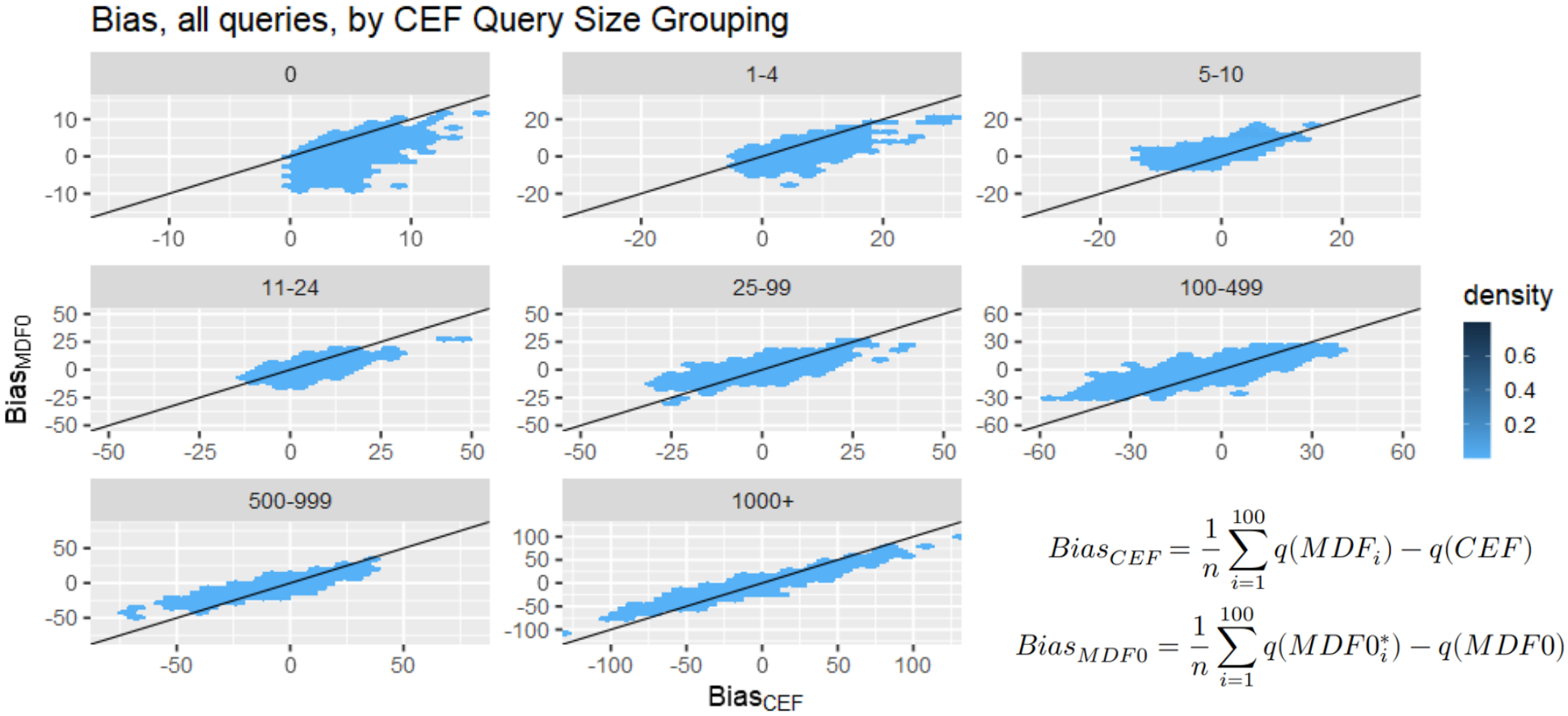


$$RMSE_{CEF} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF_i) - q(CEF))^2}$$

$$RMSE_{MDF0} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF0_i^*) - q(MDF0))^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

County Level Bias Estimation

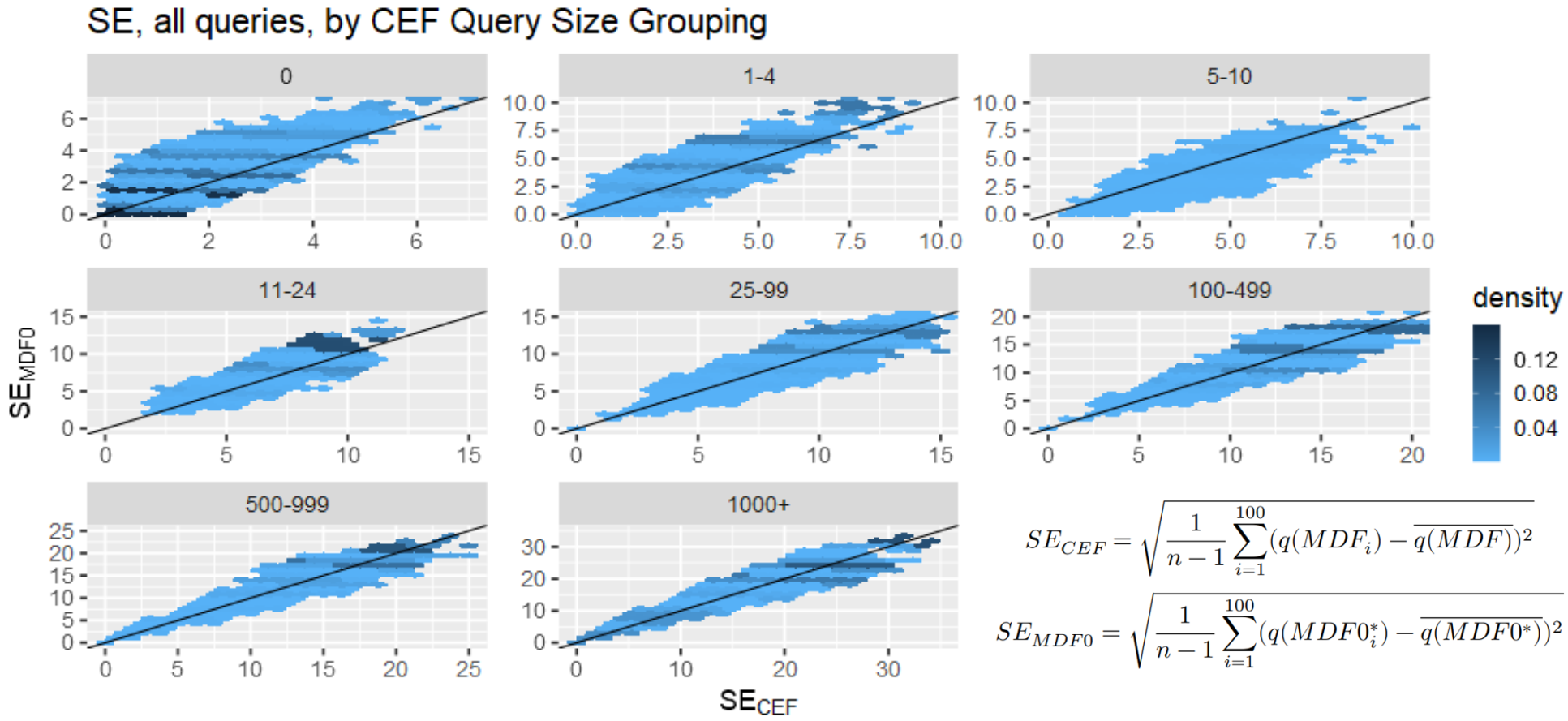


$$Bias_{CEF} = \frac{1}{n}\sum_{i=1}^{100} q(MDF_i) - q(CEF)$$

$$Bias_{MDF0} = \frac{1}{n}\sum_{i=1}^{100} q(MDF0_i^*) - q(MDF0)$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

County Level SE Estimation

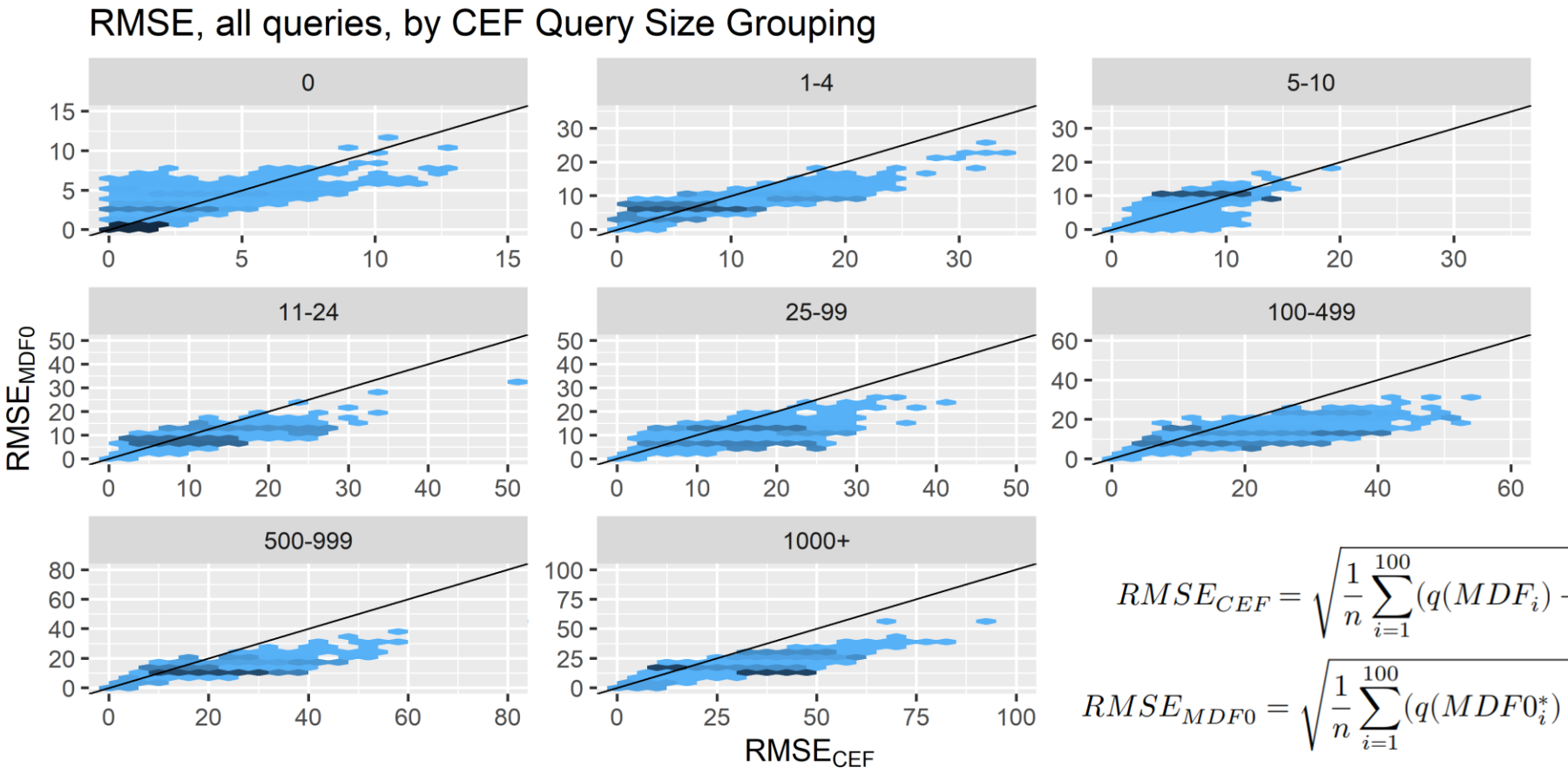


$$SE_{CEF} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF_i) - \overline{q(MDF)})^2}$$

$$SE_{MDF0} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF0_i^*) - \overline{q(MDF0^*)})^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

Tract Level RMSE Estimation

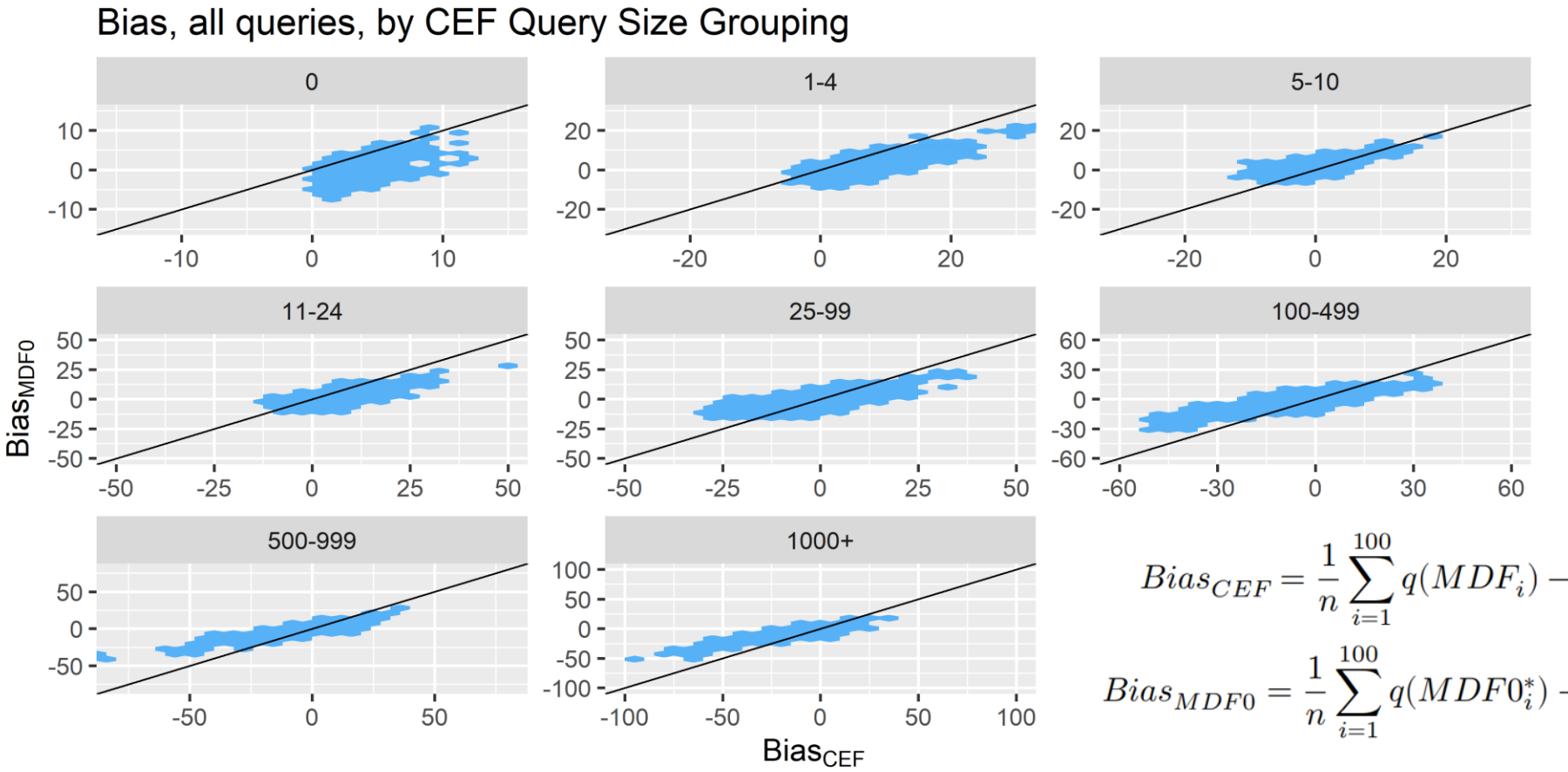


$$RMSE_{CEF} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF_i) - q(CEF))^2}$$

$$RMSE_{MDF0} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF0_i^*) - q(MDF0))^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

Tract Level Bias Estimation

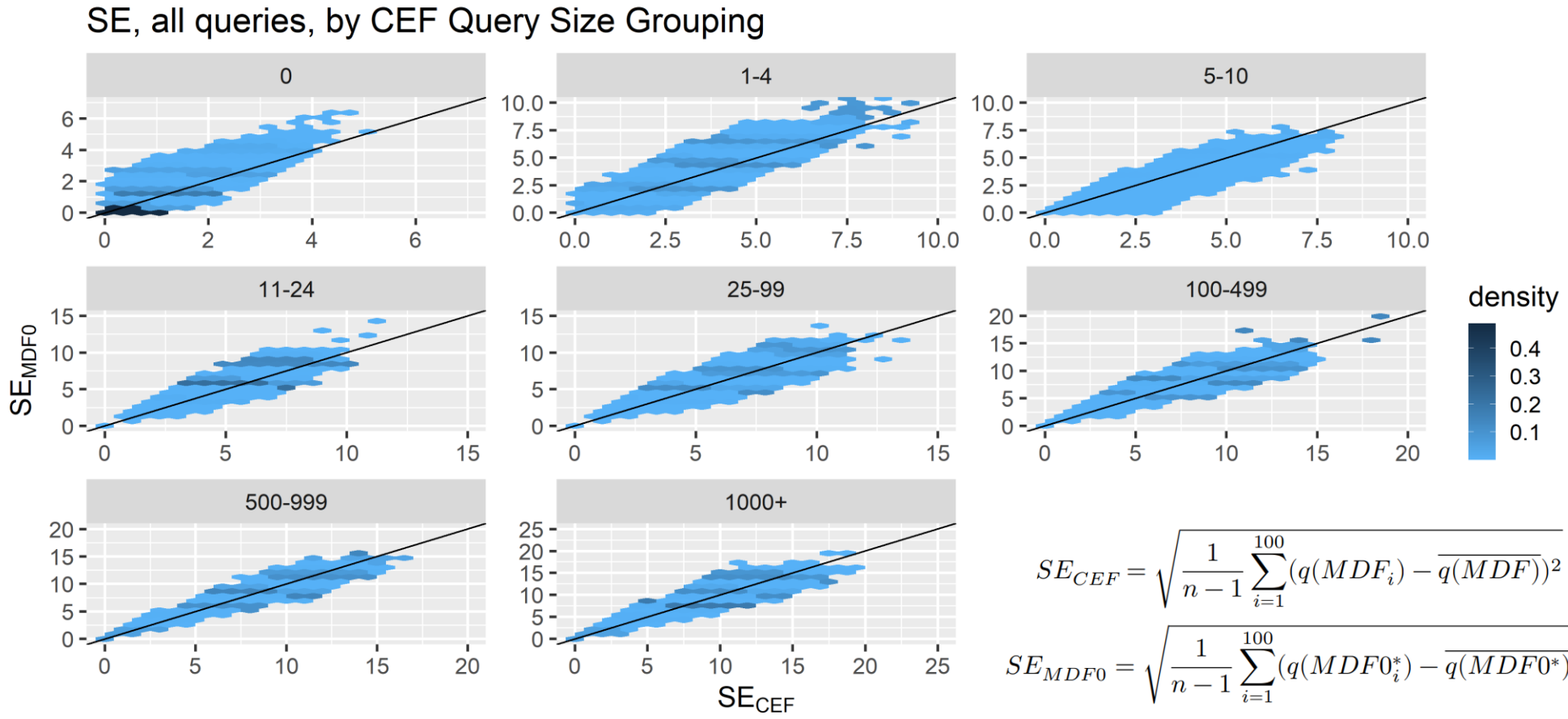


$$Bias_{CEF} = \frac{1}{n}\sum_{i=1}^{100} q(MDF_i) - q(CEF)$$

$$Bias_{MDF0} = \frac{1}{n}\sum_{i=1}^{100} q(MDF0_i^*) - q(MDF0)$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

Tract Level SE Estimation



$$SE_{CEF} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF_i) - \overline{q(MDF)})^2}$$

$$SE_{MDF0} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF0_i^*) - \overline{q(MDF0^*)})^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

CI Formulas

1. Normal Approximation w/ RMSE:

$$q(MDF0) \pm Z_{1-\frac{\alpha}{2}} RMSE_{MDF0}$$

2. Normal Approximation w/ adjusted RMSE

$$q(MDF0) \pm Z_{1-\frac{\alpha}{2}} adjRMSE_{MDF0}$$

$$adjRMSE_{MDF0} = \sqrt{(m * Bias_{MDF0})^2 + Variance_{MDF0}}$$

m is a multiplier (e.g. we will use m = 2 below)).

- Round the lower/upper end points to integers with a floor/ceiling function and use a lower limit of 0.

United States® Census Bureau

CI Examples National Level

QUERY	Description	US/PR	MDF0	90% CI	CI Width
P0030015	White; Some other race	PR	19798	19792-19804	12
P0040053	White; Black or African American; Asian; Native Hawaiian and Other Pacific Islander	PR	0	0-2	2
P0020053	White; Black or African American; Asian; Native Hawaiian and Other Pacific Islander	US	3509	3494-3524	30
P0020044	Black or African American; Native Hawaiian and Other Pacific Islander; Some Other Race	PR	5	0-10	10
P0040047	American Indian and Alaska Native; Native Hawaiian and Other Pacific Islander; Some other race	US	211	198-224	26
P0020060	Black or African American; American Indian and Alaska Native; Asian; Native Hawaiian and Other Pacific Islander	US	672	659-685	26
P0010040	Black or African American; Asian; Native Hawaiian and Other Pacific Islander	US	7324	7302-7346	44

United States® Census Bureau

CI Examples County Level

QUERY	Description	Geoid	MDF0	90% CI	CI Width
P0030026	Population of three races	22077	28	15-41	26
P0020026	Asian; Some Other Race	01093	3	0-10	10
P0020042	Black or African American; Asian; Native Hawaiian and Other Pacific Islander	02180	1	0-4	4
P0040018	Black or African American; American Indian and Alaska Native	40019	187	179-195	16
P0040043	Black or African American; Asian; Some other race	34033	3	0-9	9
H0010002	Occupied	26121	65614	65601-65627	26
P0040003	Not Hispanic or Latino	13177	20008	19985-20031	46

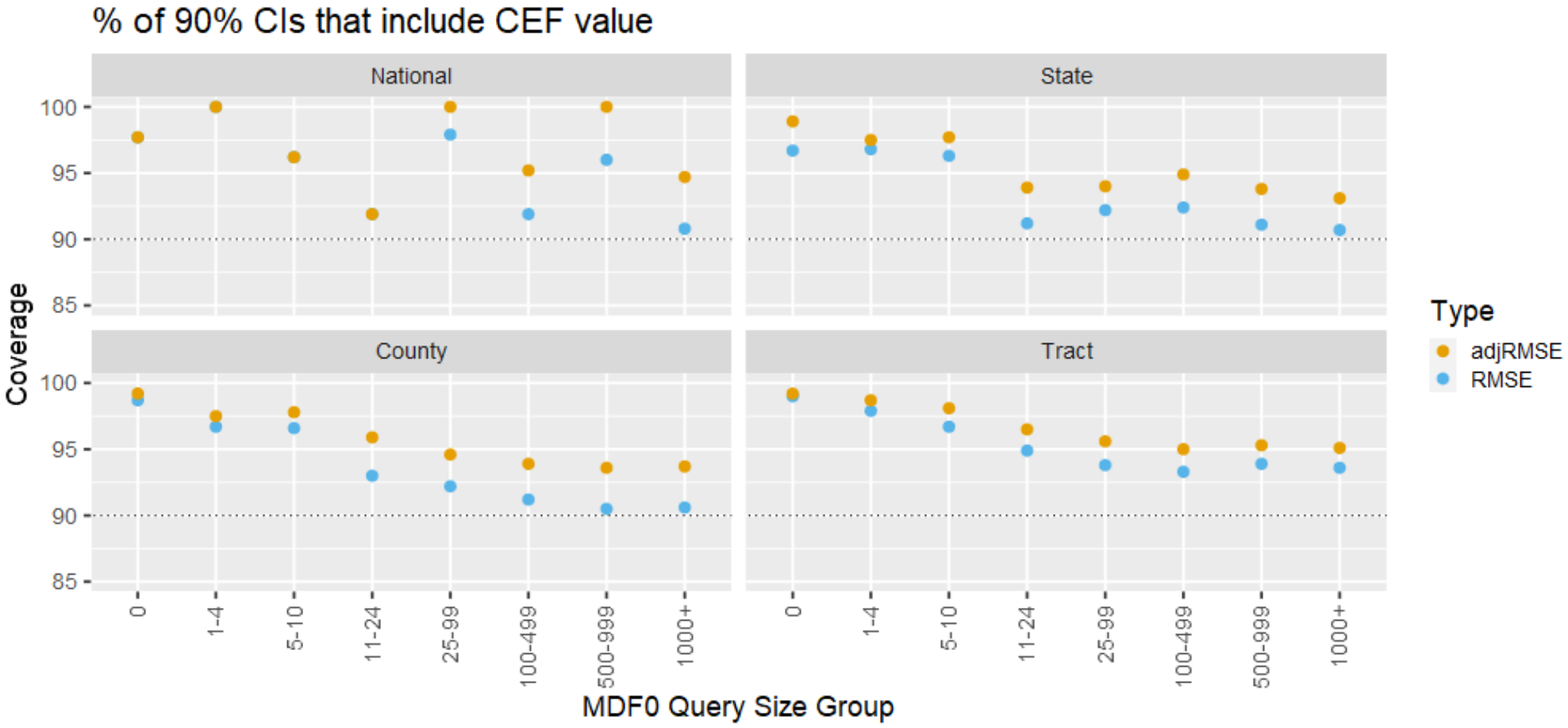
United States® Census Bureau

CI Examples Tract Level

QUERY	Description	Geoid	MDF0	90% CI	CI Width
P0050005	Nursing facilities/Skilled-nursing facilities	06081610000	10	0-20	20
P0010003	White alone	31173940100	2023	2015-2031	16
P0010014	White; Native Hawaiian and Other Pacific Islander	48029171922	1	0-4	4
P0010009	Population of two or more races:	45019002002	96	89-103	14
P0040012	Population of two races	50007003000	45	35-55	20
P0030005	American Indian and Alaska Native alone	17043846603	21	16-26	10
P0010001	Total:	06085503001	4061	4056-4066	10

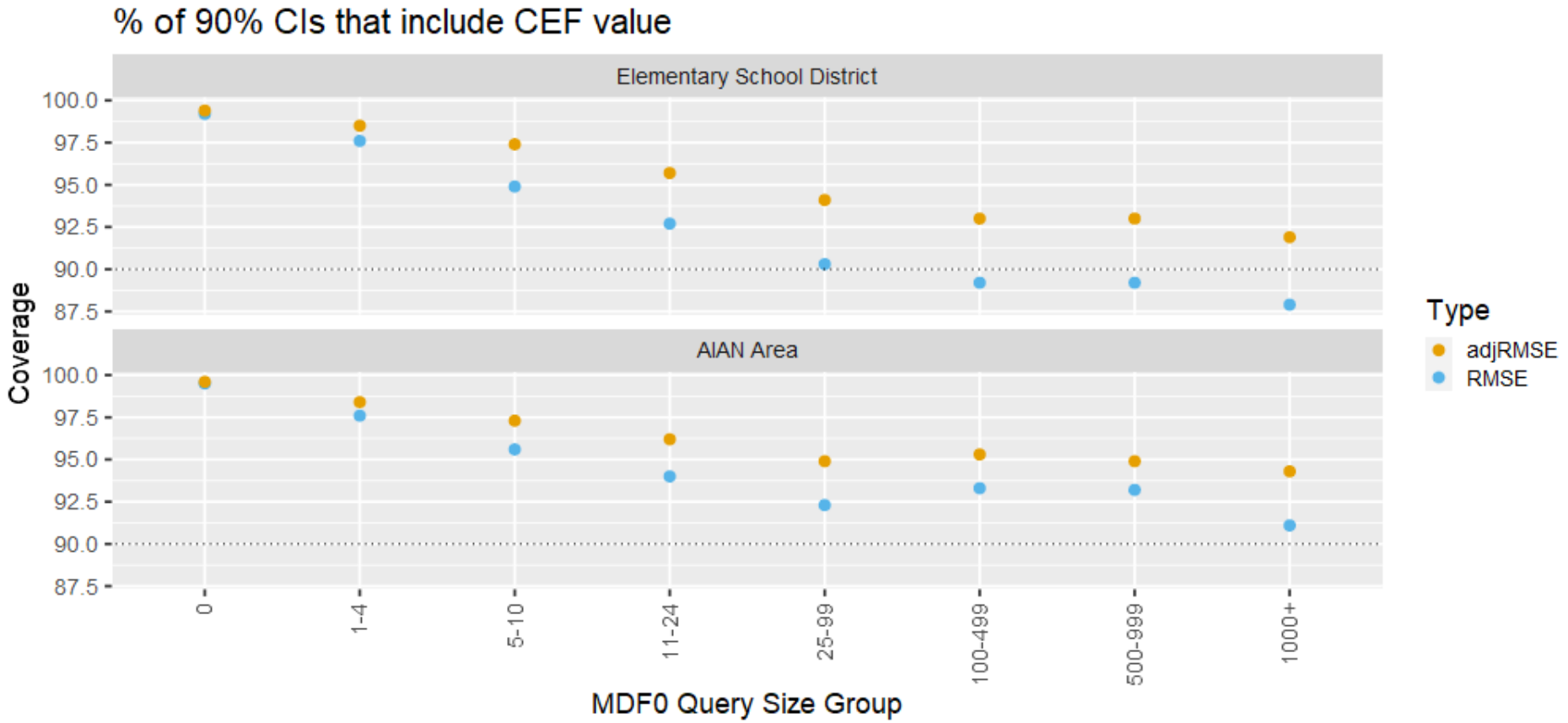
United States® Census Bureau

90% CI Results



United States® Census Bureau

Off Spine – Elementary School Districts and AIAN Areas



90% Confidence Interval Width Summary

State Level

MDF0 Query Size Group	% of Queries	CI Width: Median	CI Width: 1st Perc.	CI Width: 99th Perc.
0	9.0	6	2	17
1-4	9.8	9	4	20
5-10	7.5	15	10	24
11-24	9.7	20	10	39
25-99	14.5	22	12	58
100-499	13.7	26	13	106
500-999	4.7	30	18	129
1000+	31.1	40	0	194

United States® Census Bureau

90% Confidence Interval Width Summary

County Level

MDF0 Query Size Group	% of Queries	CI Width: Median	CI Width: 1st Perc.	CI Width: 99th Perc.
0	57.9	1	0	9
1-4	11.6	7	4	17
5-10	4.7	15	10	28
11-24	4.7	20	12	45
25-99	5.8	22	12	70
100-499	5.3	28	14	98
500-999	1.6	32	12	126
1000+	8.5	36	0	176

United States® Census Bureau

90% Confidence Interval Width Summary

Tract Level

MDF0 Query Size Group	% of Queries	CI Width: Median	CI Width: 1st Perc.	CI Width: 99th Perc.
0	71.2	1	0	4
1-4	9.3	6	3	13
5-10	3.3	10	6	21
11-24	3.7	10	6	32
25-99	4.1	14	6	42
100-499	2.4	16	6	54
500-999	0.7	14	0	52
1000+	5.3	16	0	54

United States® Census Bureau

Strengths/Weaknesses of the Approach

- Strengths
  - Possible for any query at any geography
  - Accurate estimates of variance/SE
  - Ability to compute meaningful confidence intervals using a normal approximation and the RMSE under most circumstances
- Weaknesses
  - Struggle to accurately estimate bias (relative to the size of the tabulation) in some situations
    - Very small tabulations (query answer <10)
    - Larger tabulations in lower geographies (underestimates the bias)

United States® Census Bureau

Proposed Next Steps

- Continue reviewing results and decide on a rule for the adjusted RMSE CI method
- Explore using < 100 simulations
- Extend the method to the 2020 Redistricting Data MDF and prepare for creating a research data product
- Test the approach on the DHC with 2010 data

United States® Census Bureau

Questions?

United States® Census Bureau

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session VII: Using the Noisy Measurement File

Michael Hawes
Research and Methodology

Matthew Spence
Population Division

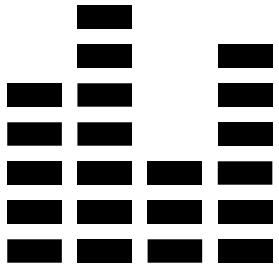
United States® Census Bureau

Should I Use the NMF, the PPMF, or the Tabulations?

- There are two sources of error in the published statistics (PPMF and Tabulations):

Differentially private noise

- Unbiased
- Known distribution
- Reflected in the noisy measurements



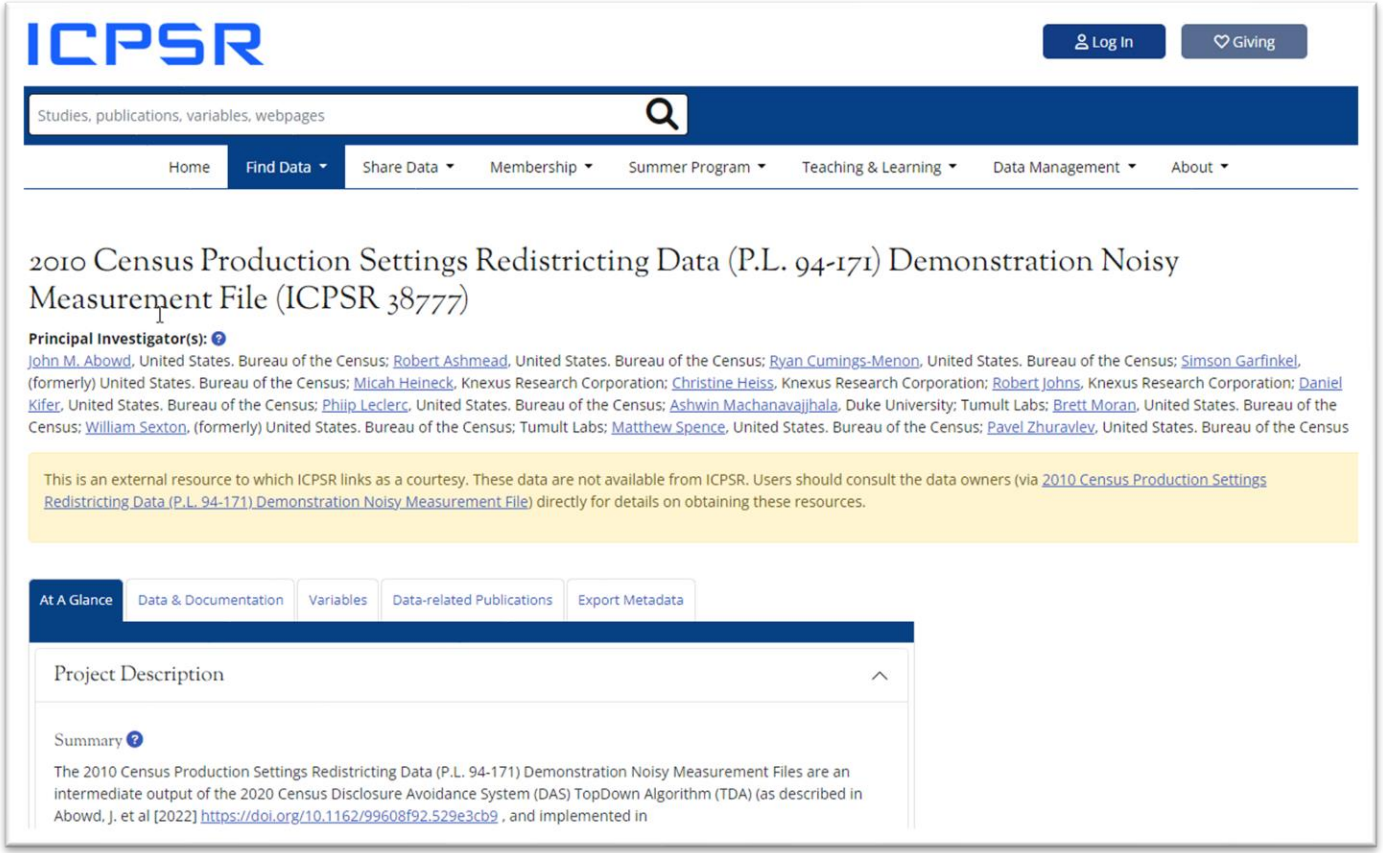
Post-processing

- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

United States® Census Bureau

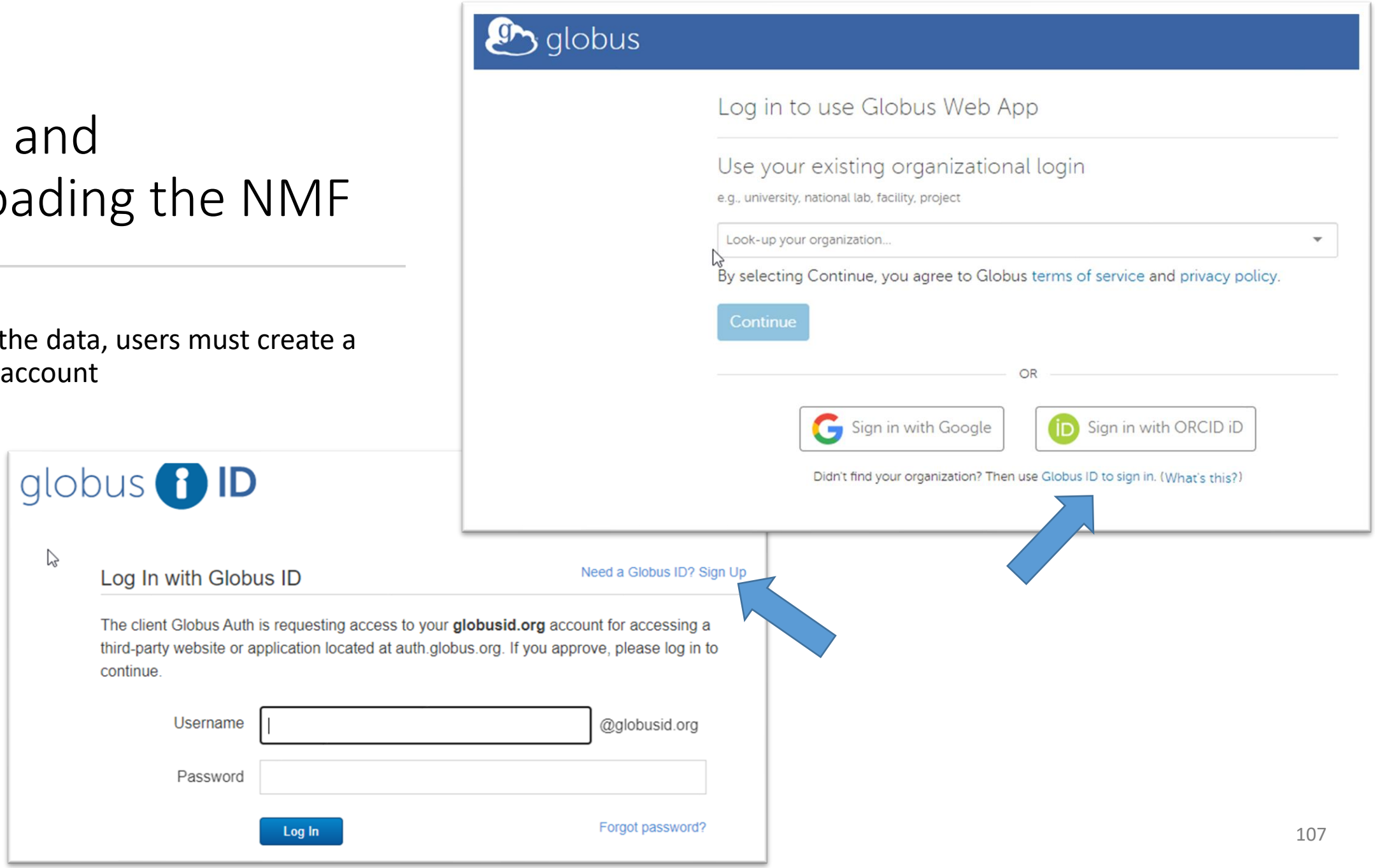
Finding and Downloading the NMF

The 2010 Redistricting NMF and supporting metadata and documentation is available via ICPSR at: https://www.icpsr.umich.edu/web/ICPSR/studies/38777



Finding and Downloading the NMF

To download the data, users must create a (free) Globus account

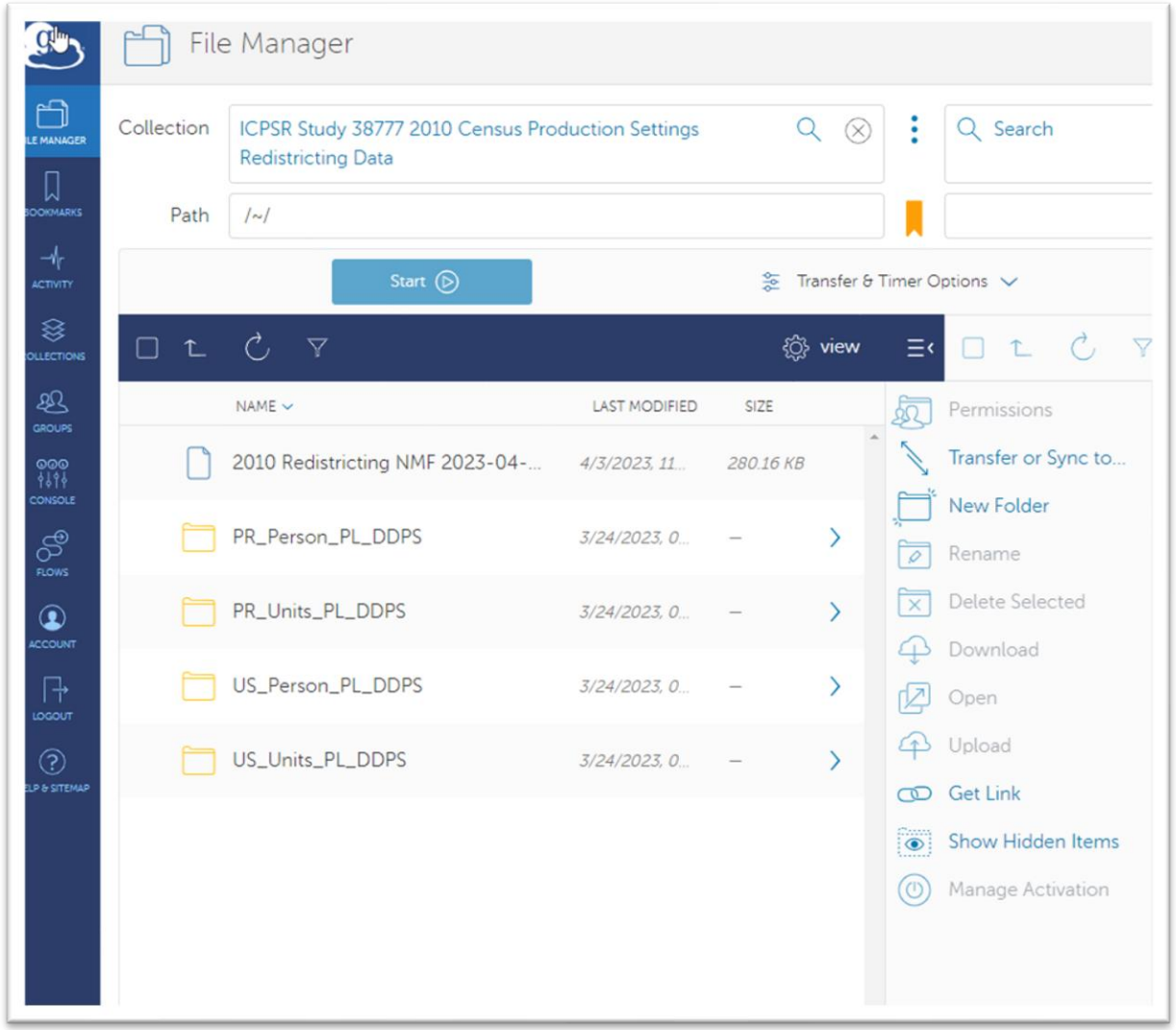


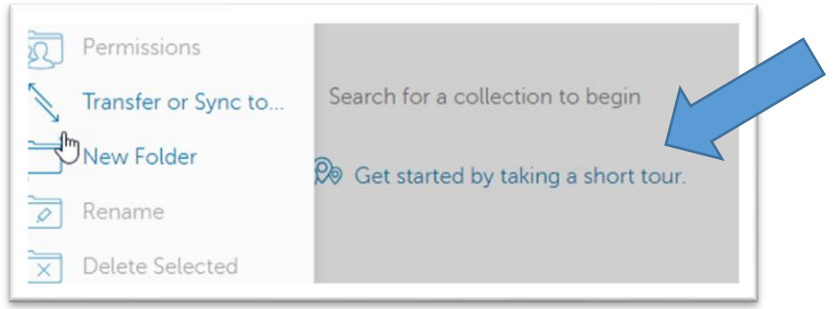
Finding and Downloading the NMF

Once you are logged into Globus, the NMFs are stored in the “ICPSR Study 38777 2010 Census Production Settings Redistricting Data” collection.

You can select which component file(s) to transfer to your local machine.

Globus has a tutorial for how to create a workspace (“collection”) on your local machine into which you can transfer the files.





Understanding 2020 DAS Geocodes

- Because of spine optimization, DAS geocodes differ from traditional tabulation geocodes.

Bullock County, AL

DAS Geocode: ***0 0 1 1 0 0 1 1***

United States® Census Bureau

Understanding 2020 DAS Geocodes

- Because of spine optimization, DAS geocodes differ from traditional tabulation geocodes.

Bullock County, AL

DAS Geocode: ***0 0 1 1 0 0 1 1***

Indicates non-AIANNH spine

United States® Census Bureau

Understanding 2020 DAS Geocodes

- Because of spine optimization, DAS geocodes differ from traditional tabulation geocodes.

Bullock County, AL

DAS Geocode: ***0 0 1 1 0 0 1 1***

State code 01

Understanding 2020 DAS Geocodes

- Because of spine optimization, DAS geocodes differ from traditional tabulation geocodes.

Bullock County, AL

DAS Geocode: ***0 0 1 1 0 0 1 1***

Ignore for all DAS geocodes

United States® Census Bureau

Understanding 2020 DAS Geocodes

- Because of spine optimization, DAS geocodes differ from traditional tabulation geocodes.

Bullock County, AL

DAS Geocode: ***0 0 1 1 0 0 1 1***

County

United States® Census Bureau

Understanding 2020 DAS Geocodes

- In most cases, spine optimization has transformed geocodes for tracts and block groups, so the DAS geocodes for these geographies are not directly comparable.
- These geographies can be constructed from their component blocks.
- All DAS block-level geocodes contain the full 16-digit tabulation block geocode as the final 16 digits of the DAS geocode.

Bourbon St, New Orleans, LA:

Tabulation Block Geocode

0 2 2 1 0 0 7 1 0 1 6 4 1 0 2 2 2 0 7 1 0 1 3 5 0 0 3 3 0 0 6

DAS Block Geocode

United States® Census Bureau

Structure of the NMF Component Files

The component files of the NMF are first divided into the four separate Persons and Units runs for the U.S. and Puerto Rico

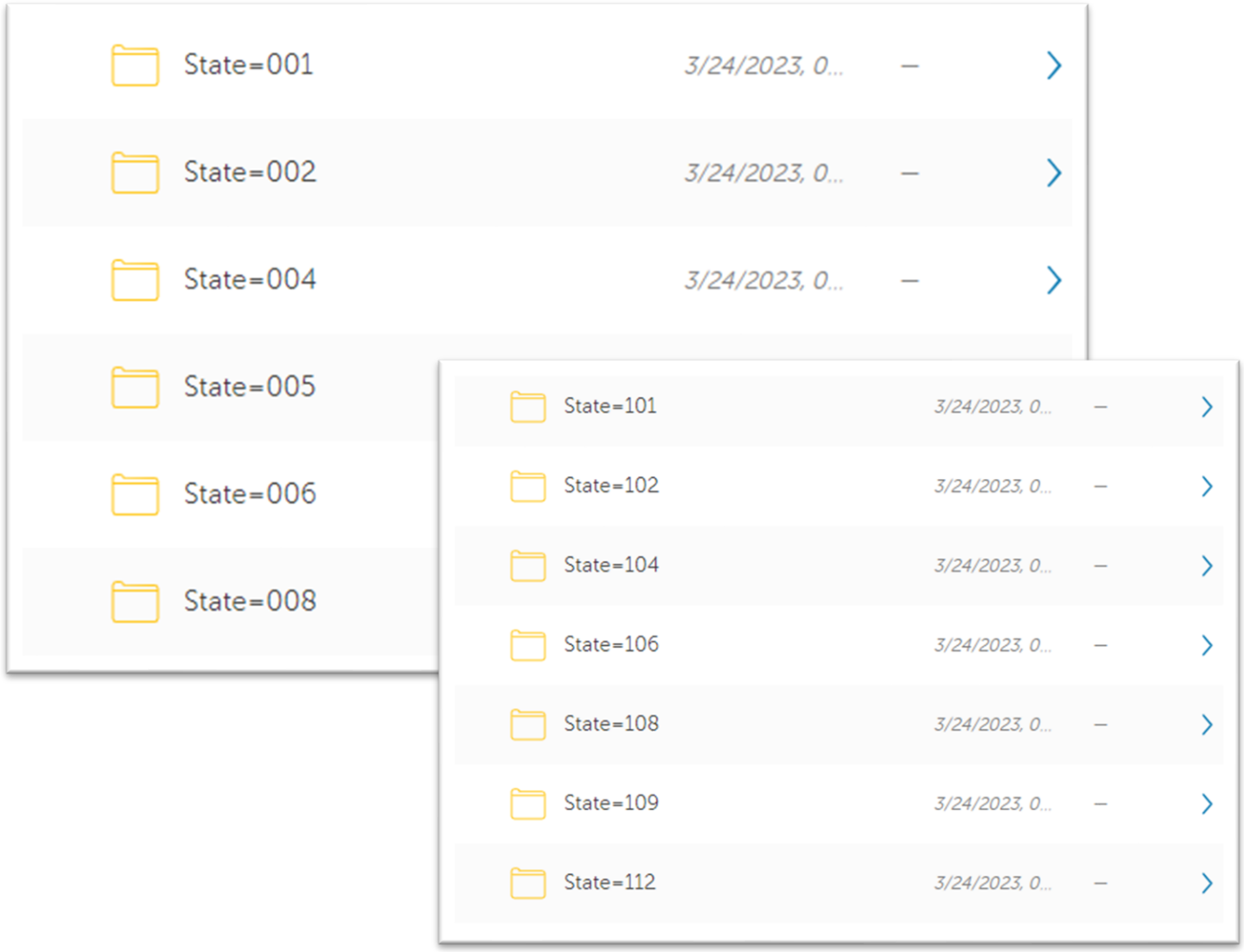
Structure of the NMF Component Files

The component files are then subdivided by geographic level.

Structure of the NMF Component Files

Within each geographic level, the files are divided into DPQueries (the noisy measurements) and Constraints (the equalities and inequalities that the TDA uses during post-processing).

Structure of the NMF Component Files

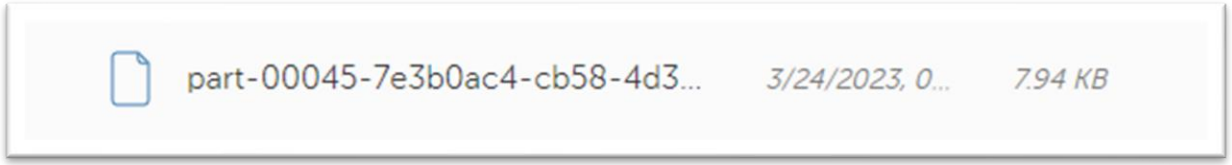


At the state level and below, files are further separated by their state DAS geocode.

Note, the first digit (0 or 1) of the DAS state geocode indicates the AIANNH (1) or non-AIANNH (0) branch of the spine.

States that have AIANNH areas will have two folders.

Structure of the NMF Component Files



Within the selected geography, you will find the .parquet file containing the noisy measurements for all geounits of that particular level within the selected state geocode.

Converting from Parquet Format

- The NMF is stored in a compressed Parquet file format.
- In order to extract the noisy measurements for use in most statistical packages, it is necessary to convert the .parquet files into .csv (or another format).
- Python and R have modules to do this conversion.

```
In [1]: import pandas as pd
        df = pd.read_parquet('filename.parquet')
        df.to_csv('filename.csv')
```

Insert R code here to convert to .csv

United States® Census Bureau

Content of the NMF

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
0	1100890045	total_dpq	*	*	*	*	[1 1 1 1]	[2980]	6.110105	[52707577526 322048826043]
1	1100890045	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[1975 837 22 6 3 51 9 -9 26 7 -20 -40 72 47 -55 -9 44 -52 -44 44 55 -49 11 -39 -41 -1 -70 -20 12 20 22 -17 51 45 34 76 23 20 0 -27 5 48 6 22 -66 -24 -41 20 -6 -4 -15 -12 -7 -72 -42 1 -26 -14 16 -30 -14 26 -48]	1196.8168	[269087824 322048826043]
2	1100890045	hispanic_dpq	*	*	hispanic	*	[1 1 2 1]	[2890 123]	1914.9069	[168179890 322048826043]
3	1100890045	votingage_dpq	*	votingage	*	*	[1 2 1 1]	[379 2645]	1914.9069	[168179890 322048826043]
4	1100890045	hhinstlevels_dpq	hhinstlevels	*	*	*	[3 1 1 1]	[1781 1127 29]	1914.9069	[168179890 322048826043]
5	1100890045	hhgq_dpq	hhgq	*	*	*	[8 1 1 1]	[1754 1017 59 120 6 -68 -84 -10]	1914.9069	[168179890 322048826043]
6	1100890045	hispanic * cenrace_dpq	*	*	hispanic	cenrace	[1 1 2 63]	[1933 856 24 18 2 2 11 15 3 1 1 7 1 -2 -4 -1 2 -1 5 3 2 -1 -3 -3 1 -2 2 0 2 0 -1 0 -1 0 1 2 -6 -3 -1 3 1 0 -1 -1 1 -2 0 0 -3 -5 1 2 1 1 -2 -1 0 7 -3 -4 0 1 2 67 8 10 -2 -1 13 0 4 0 0 3 -1 0 -2 3 1 1 7 6 2 0 2 0 4 -2 -1 -1 3 0 2 -2 3 3 -2 4 3 4 -1 1 -3 -5 2 -2 -2 1 2 0 -1 4 0 1 -1 -1 4 3 2 -2 4 2 0 0 3 -1]	4.9531994	[65018345474 322048826043]

United States® Census Bureau

Generating Confidence Intervals – Example 1

Generating a simple county-level estimate and CI

Bullock County, AL (00110011):

total_dpq: 10911

variance: 4.70162740

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
385	110011	total_dpq	*	*	*	*	[1 1 1 1]	[10911]	4.7016273	[45609211112 214437516707]
386	110011	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[2463 7606 83 -12 13 627 0 72 -66 35 -98 42 -11 42 -11 13 -17 -25 40 30 5 -49 34 12 -9 0 31 -9 48 21 -73 54 14 83 20 32 4 -1 -26 35 63 -1 -11 11 -30 21 -50 4 -113 -3 59 7 -45 -37 -10 20 23 1 -86 1 29 21 -57]	1469.7288	[437708360 643312550121]
387	110011	hispanic_dpq	*	*	hispanic	*	[1 1 2 1]	[10166 800]	1469.7288	[437708360 643312550121]
388	110011	votingage_dpq	*	votingage	*	*	[1 2 1 1]	[2481 8435]	1469.7288	[437708360 643312550121]
389	110011	hhinstlevels_dpq	hhinstlevels	*	*	*	[3 1 1 1]	[9179 1664 33]	1469.7288	[437708360 643312550121]
390	110011	hhgq_dpq	hhgq	*	*	*	[8 1 1 1]	[9243 1540 57 157 -48 26 4 -39]	1469.7288	[437708360 643312550121]
391	110011	hispanic * cenrace_dpq	*	*	hispanic	cenrace	[1 1 2 63]	[2356 7605 51 -7 -37 9 17 33 -14 -19 5 54 13 6 -6 -1 4 33 11 -37 22 -4 5 14 19 -26 36 -2	524.90314	[1225583408 643312550121]

United States® Census Bureau

Generating Confidence Intervals – Example 1

Generating a simple county-level estimate and CI

Bullock County, AL (00110011):

total_dpq: 10911

variance: 4.70162740

Approximation of 90% CI via normal distribution:

10911 +/- 1.64 * Sqrt(4.7016) = 10911 +/- 4.2499

Calculation of 90% CI via discrete gaussian distribution:

```
In [5]: # Constants in this example:
        sigma_sqrd_county = 4.70162740
        dp_answer = np.array([10911])

        for ci_half_width in range(1000):
            prob = discrete_gaussian_distribution(ci_half_width, sigma_sqrd=sigma_sqrd_county)
            if prob >= 1 - alpha:
                print(f"The confidence interval is: {dp_answer} +/- {ci_half_width}")
                break

        The confidence interval is: [10911] +/- 4
```

United States® Census Bureau

Tip No. 1: Though the TDA uses discrete gaussian noise, a close approximation of the CI can be easily calculated for any particular noisy measurement using a normal distribution and the share of privacy-loss budget allocated to that measurement's query. Calculating the CI using the discrete gaussian distribution (and a sufficient number of simulation iterations) will yield a more exact CI.

Generating Confidence Intervals – Example 2

Generating a more precise county-level estimate and CI by leveraging multiple noisy measurements

Bullock County, AL (00110011):
total_dpq: 10911
variance: 4.70162740

Tract 1 (001100110001):
total_dpq: 1433
variance: 6.11010487

Tract 2 (001100110002):
total_dpq: 7105
variance: 6.11010487

Tract 3 (001100110003):
total_dpq: 2376
variance: 6.11010487

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
385	110011	total_dpq	*	*	*	*	[1 1 1 1]	[10911]	4.7016273	[45609211112 214437516707]

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
922	1100110001	total_dpq	*	*	*	*	[1 1 1 1]	[1433]	6.110105	[52707577526 322048826043]
436	1100110002	total_dpq	*	*	*	*	[1 1 1 1]	[7105]	6.110105	[52707577526 322048826043]
436	1100110002	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[1334 5047 7 3 57 541 29 -72 34 55 18 38 -76 12	1196.8168	[269087824 322048826043]
143	1100110003	total_dpq	*	*	*	*	[1 1 1 1]	[2376]	6.110105	[52707577526 322048826043]
143	1100110003	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[1025 1291 13 1 -50 42 11 63 0 63 6 -63 -14 -27	1196.8168	[269087824 322048826043]

United States® Census Bureau

Generating Confidence Intervals – Example 2

Generating a more precise county-level estimate and CI by leveraging multiple noisy measurements

Bullock County, AL (00110011):	Tract 1 (001100110001):	Tract 2 (001100110002):	Tract 3 (001100110003):
total_dpq: 10911 variance: 4.70162740	total_dpq: 1433 variance: 6.11010487	total_dpq: 7105 variance: 6.11010487	total_dpq: 2376 variance: 6.11010487

```
In [6]: # Constants in this example:
        coef_mat = np.vstack((np.ones((1, 3)), np.eye(3)))
        var_county = 4.70162740
        var_tracts = 6.11010487
        sigma_sqrds = np.array([var_county] + [var_tracts] * 3)
        dp_answers = np.array([10911, 1433, 7105, 2376])

        weights = np.diag(1 / sigma_sqrds)
        point_estimate, ci_half_width = simulation_based_ci(dp_answers, coef_mat, weights, sigma_sqrds, prng,
        num_sims, alpha)
        print(f"The confidence interval is: {point_estimate} +/- {ci_half_width}")

        The confidence interval is: 10911.612405249798 +/- 3.193797375100286
```

Tip No. 2: Even when the statistic of interest is measured directly in the NMFs, it can be advantageous to incorporate additional information (e.g., noisy measurements for the child geounits) in order to obtain more precise estimates and smaller CIs.

United States® Census Bureau

Generating Confidence Intervals – Example 3

Combining noisy measurements from the AIANNH and non-AIANNH branches of the geographic hierarchy

Menominee County (non-AIAN Portion) (05510078):	Menominee County (AIAN Portion) (15510078):
total_dpq: 1339	total_dpq: 2893
variance: 4.70162740	variance: 4.70162740

In [7]:
```
# Constants in this example:
coef_mat = np.eye(2)
var_county = 4.70162740
sigma_sqrds = np.array([var_county] * 2)
dp_answers = np.array([1339, 2893])

weights = np.diag(1 / sigma_sqrds)
point_estimate, ci_half_width = simulation_based_ci(dp_answers, coef_mat, weights, sigma_sqrds, prng,
num_sims, alpha)
print(f"The confidence interval is: {point_estimate} +/- {ci_half_width}")
```

```
The confidence interval is: 4232.0 +/- 5.0
```

Tip No. 3: Remember that DAS geocodes differ from traditional tabulation geocodes. Always verify if the geounit of interest has been split into AIAN/non-AIAN portions

United States® Census Bureau

Generating Confidence Intervals – Example 4

Calculating an estimate and CI for an off-spine geography

Redfield, IA – Block Group 1 (019100490006103):
- total_dpq: 696
- variance: 3.06932331

Redfield, IA – Block Group 2 (019100490006104):
- total_dpq: 139
- variance: 3.06932331

In [8]:
```
# Constants in this example:
coef_mat = np.eye(2)
var_optimized_bg = 3.06932331
sigma_sqrds = np.array([var_optimized_bg] * 2)
dp_answers = np.array([696, 139])

weights = np.diag(1 / sigma_sqrds)
point_estimate, ci_half_width = simulation_based_ci(dp_answers, coef_mat, weights, sigma_sqrds, prng,
num_sims, alpha)
print(f"The confidence interval is: {point_estimate} +/- {ci_half_width}")
```

```
The confidence interval is: 835.0 +/- 4.0
```

Tip No. 4: When generating estimates and CIs for custom geographies (those for which there are no direct noisy measurements), it is generally better to minimize the number of on-spine geounit queries that have to be combined to form the geography of interest. For example, if a school district spanned three complete tracts and one additional census block, combining the noisy measurements of the three tracts and the lone census block will typically yield a more precise estimate and narrower CI than adding up all of the individual census blocks' measurements would.

United States® Census Bureau

Generating Confidence Intervals – Example 5

Incorporating multiple queries in a calculation

Bullock County, AL (00110011):
total_dpq: 10911
variance: 4.70162740

Bullock County, AL (00110011):
votingage_dpq: 2481 8435
variance: 1469.72873

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
385	110011	total_dpq	*	*	*	*	[1 1 1 1]	[10911]	4.7016273	[45609211112 214437516707]
386	110011	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[2463 7606 83 -12 13 627 0 72 -66 35 -98 42 -11 42 -11 13 -17 -25 40 30 5 -49 34 12 -9 0 31 -9 48 21 -73 54 14 83 20 32 4 -1 -26 35 63 -1 -11 11 -30 21 -50 4 -113 -3 59 7 -45 -37 -10 20 23 1 -86 1 29 21 -57]	1469.7288	[437708360 643312550121]
387	110011	hispanic_dpq	*	*	hispanic	*	[1 1 2 1]	[10166 800]	1469.7288	[437708360 643312550121]
388	110011	votingage_dpq	*	votingage	*	*	[1 2 1 1]	[2481 8435]	1469.7288	[437708360 643312550121]
389	110011	hhinstlevels_dpq	hhinstlevels	*	*	*	[3 1 1 1]	[9179 1664 33]	1469.7288	[437708360 643312550121]
390	110011	hhgq_dpq	hhgq	*	*	*	[8 1 1 1]	[9243 1540 57 157 -48 26 4 -39]	1469.7288	[437708360 643312550121]
391	110011	hispanic * cenrace_dpq	*	*	hispanic	cenrace	[1 1 2 63]	[2356 7605 51 -7 -37 9 17 33 -14 -19 5 54 13 6 -6 -1 4 33 11 -37 22 -4 5 14 19 -26 36 -2	524.90314	[1225583408 643312550121]

United States® Census Bureau

Generating Confidence Intervals – Example 5

Incorporating multiple queries in a calculation

Bullock County, AL (00110011):	Bullock County, AL (00110011):
total_dpq: 10911	votingage_dpq: 2481 8435
variance: 4.70162740	variance: 1469.72873

In [9]:
```
# Constants in this example:
coef_mat = np.vstack((np.ones((1, 2)), np.eye(2)))
var_total_pop = 4.70162740
var_votingage = 1469.72873
sigma_sqrds = np.array([var_total_pop] + [var_votingage] * 2)
dp_answers = np.array([10911, 2481, 8435])

weights = np.diag(1 / sigma_sqrds)
point_estimate, ci_half_width = simulation_based_ci(dp_answers, coef_mat, weights, sigma_sqrds, prng,
num_sims, alpha)
print(f"The confidence interval is: {point_estimate} +/- {ci_half_width}")
```

```
The confidence interval is: 10911.007984669362 +/- 3.86214468265589
```

Tip No. 5: Noisy measurements with larger PLB allocations will generally produce more precise estimates with narrower CIs. When selecting between alternate combinations of noisy measurements to construct an estimate, choose those that have larger PLB allocations.

United States® Census Bureau

Generating Confidence Intervals – Example 6

Leveraging additional noisy measurements to increase precision

Bullock County, AL: Total Population and Voting Age Queries for the county and the 3 child tracts (12 noisy measurements)

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
385	110011	total_dpq	*	*	*	*	[1 1 1 1]	[10911]	4.7016273	[45609211112 214437516707]
386	110011	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[2463 7606 83 -12 13 627 0 72 -66 35 -98 42 -11 42 -11 13 -17 -25 40 30 5 -49 34 12 -9 0 31 -9 48 21 -73 54 14 83 20 32 4 -1 -26 35 63 -1 -11 11 -30 21 -50 4 -113 -3 59 7 -45 -37 -10 20 23 1 -86 1 29 21 -57]	1469.7288	[437708360 643312550121]
387	110011	hispanic_dpq	*	*	hispanic	*	[1 1 2 1]	[10166 800]	1469.7288	[437708360 643312550121]
388	110011	votingage_dpq	*	votingage	*	*	[1 2 1 1]	[2481 8435]	1469.7288	[437708360 643312550121]
389	110011	hhinstlevels_dpq	hhinstlevels	*	*	*	[3 1 1 1]	[9179 1664 33]	1469.7288	[437708360 643312550121]
390	110011	hhgq_dpq	hhgq	*	*	*	[8 1 1 1]	[9243 1540 57 157 -48 26 4 -39]	1469.7288	[437708360 643312550121]
391	110011	hispanic * cenrace_dpq	*	*	hispanic	cenrace	[1 1 2 63]	[2356 7605 51 -7 -37 9 17 33 -14 -19 5 54 13 6 -6 -1 4 33 11 -37 22 -4 5 14 19 -26 36 -2	524.90314	[1225583408 643312550121]

United States® Census Bureau

Generating Confidence Intervals – Example 6

Leveraging additional noisy measurements to increase precision

Bullock County, AL: Total Population and Voting Age Queries for the county and the 3 child tracts (12 noisy measurements)

In [10]:
```
# Constants in this example:
per_geounit_query_mat = np.vstack((np.ones((1, 2)), np.eye(2)))
spine_mat = np.vstack((np.ones((1, 3)), np.eye(3)))
coef_mat = np.kron(spine_mat, per_geounit_query_mat)
var_total_pop_county = 4.70162740
var_votingage_county = 1469.72873
var_total_pop_tract = 6.11010487
var_votingage_tract = 1914.90687
sigma_sqrds_county = np.array([var_total_pop_county] + [var_votingage_county] * 2)
sigma_sqrds_tract = np.array([var_total_pop_tract] + [var_votingage_tract] * 2)
sigma_sqrds = np.concatenate([sigma_sqrds_county] + [sigma_sqrds_tract] * 3)
dp_answers = np.array([10911, 2481, 8435, 1433, 298, 1048, 7105, 1550, 5609, 2376, 490, 1914])

weights = np.diag(1 / sigma_sqrds)
point_estimate, ci_half_width = simulation_based_ci(dp_answers, coef_mat, weights, sigma_sqrds, prng,
num_sims, alpha)
print(f"The confidence interval is: {point_estimate} +/- {ci_half_width}")
```

```
The confidence interval is: 10911.617132194644 +/- 3.2198631260586152
```

United States® Census Bureau

Tip No. 6: Incorporating additional noisy measurements with small PLBs into the weighted calculation of an estimate and CI can improve precision, but those improvements may be minor.

Generating Confidence Intervals – Example 7

Be strategic in selecting which noisy measurements to use, and don't ignore negative values

Bullock County, AL (00110011):

hhgq_dpq: 9243 1540 57 157 -48 26 4 -39

variance: 1914.9069

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
385	110011	total_dpq	*	*	*	*	[1 1 1 1]	[10911]	4.7016273	[45609211112 214437516707]
386	110011	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[2463 7606 83 -12 13 627 0 72 -66 35 -98 42 -11 42 -11 13 -17 -25 40 30 5 -49 34 12 -9 0 31 -9 48 21 -73 54 14 83 20 32 4 -1 -26 35 63 -1 -11 11 -30 21 -50 4 -113 -3 59 7 -45 -37 -10 20 23 1 -86 1 29 21 -57]	1469.7288	[437708360 643312550121]
387	110011	hispanic_dpq	*	*	hispanic	*	[1 1 2 1]	[10166 800]	1469.7288	[437708360 643312550121]
388	110011	votingage_dpq	*	votingage	*	*	[1 2 1 1]	[2481 8435]	1469.7288	[437708360 643312550121]
389	110011	hhinstlevels_dpq	hhinstlevels	*	*	*	[3 1 1 1]	[9179 1664 33]	1469.7288	[437708360 643312550121]
390	110011	hhgq_dpq	hhgq	*	*	*	[8 1 1 1]	[9243 1540 57 157 -48 26 4 -39]	1469.7288	[437708360 643312550121]
391	110011	hispanic * cenrace_dpq	*	*	hispanic	cenrace	[1 1 2 63]	[2356 7605 51 -7 -37 9 17 33 -14 -19 5 54 13 6 -6 -1 4 33 11 -37 22 -4 5 14 19 -26 36 -2	524.90314	[1225583408 643312550121]

United States® Census Bureau

Generating Confidence Intervals – Example 7

Be strategic in selecting which noisy measurements to use, and don't ignore negative values

Bullock County, AL (00110011):

hhgq_dpq: 9243 1540 57 157 -48 26 4 -39

variance: 1914.9069

```
In [11]: # Constants in this example:
         coef_mat = np.eye(8)
         var_county = 1914.9069
         sigma_sqrds = np.array([var_county] * 8)
         dp_answers = np.array([9243, 1540, 57, 157, -48, 26, 4, -39])

         weights = np.diag(1 / sigma_sqrds)
         point_estimate, ci_half_width = simulation_based_ci(dp_answers, coef_mat, weights, sigma_sqrds, prng,
         num_sims, alpha)
         print(f"The confidence interval is: {point_estimate} +/- {ci_half_width}")

         The confidence interval is: 10940.0 +/- 205.0
```

Tip No. 7: Don't ignore negative noisy measurements. The inclusion of negative measurements can increase precision of estimates produced using the noisy measurements.

Tip No. 8: Be mindful of how you construct estimates from the noisy measurements. While the detailed query may contain noisy measurements for every cross-tabulation of attributes included in the NMFs, the low share of PLB allocated to each of those queries means that most estimates that could be calculated from the detailed query would be substantially less precise that comparable estimates that included more focused noisy measurements that received higher shares of PLB.

United States® Census Bureau

How do these estimates compare?

2010 Total Population	SF1	2010 PPMF	2010 NMF	Parametric Bootstrap
Bullock County, AL	10,914	10,912	10,911.61 +/- 3.19	10,909-10,917 (est. bias 0.2)
Menominee, WI	4,232	4,230	4,232 +/- 5	4,225-4,237 (est. bias 0.64)
Redfield, IA	835	835	835 +/- 4	

Simulated Error in Population in Households for Counties (Excluding Puerto Rico) From Nonsampling Variability

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**117.27**	**-248**	**+230**
Counties with housing unit population between 0-999	37	10.03	-10	+27
Counties with housing unit population between 1,000-9,999	691	28.23	-38	+71
Counties with housing unit population between 10,000-99,999	1,849	74.45	-131	+177
Counties with housing unit population between 100,000-999,999	527	292.21	-784	+545
Counties with housing unit population at or above 1,000,000	39	1,463.12	-3,659	+1,351

Source: U.S. Census Bureau, Research and Methodology Directorate, simulations based on 2010 Census Coverage Measurement research and 2010 Census housing unit population.

Simulated Error in Population in Households for Counties (Excluding Puerto Rico) From Census Coverage Error

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**964.00**	**-1,841**	**+2,048**
Counties with housing unit population between 0-999	37	23.00	-22	+54
Counties with housing unit population between 1,000-9,999	691	121.00	-146	+284
Counties with housing unit population between 10,000-99,999	1,849	446.00	-832	+1,053
Counties with housing unit population between 100,000-999,999	527	2,930.00	-7,222	+6,278
Counties with housing unit population at or above 1,000,000	39	14,848.00	-44,833	+20,007

Source: U.S. Census Bureau, Research and Methodology Directorate, simulations based on 2010 Census Coverage Measurement research and 2010 Census housing unit population.

Source: Understanding Disclosure Avoidance-Related Variability 2020 Census

United States® Census Bureau

Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session VIII: Concluding Discussion

Sallie Ann Keller
Associate Director for Research and
Methodology and Chief Scientist

United States® Census Bureau

What additional resources would you like to see?

United States® Census Bureau

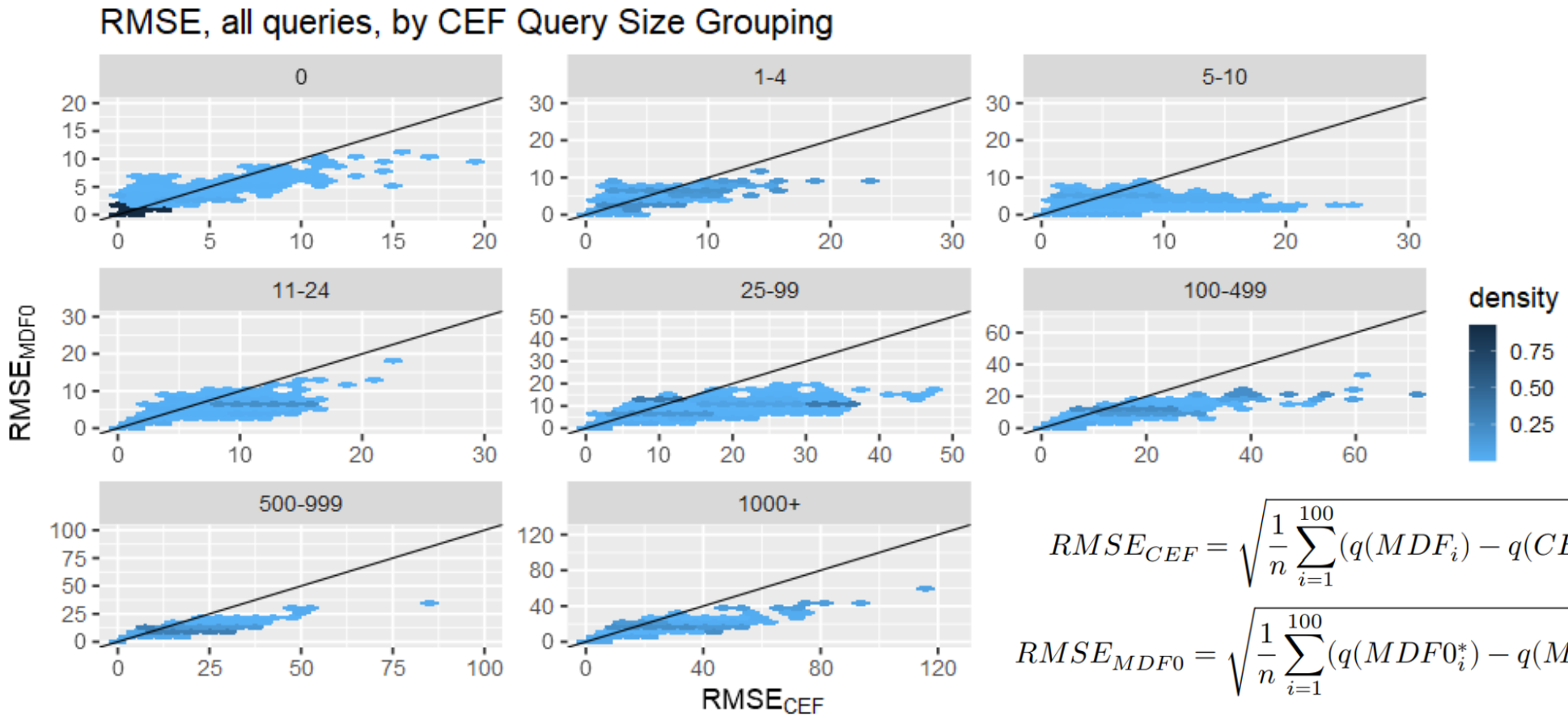
Workshop on Assessing Fitness-for-Use of Differential Privacy-Adjusted Census Data

Session VI: Appendix Slides

Robert Ashmead
Research and Methodology

United States® Census Bureau

Block Level RMSE Estimation (Sample)

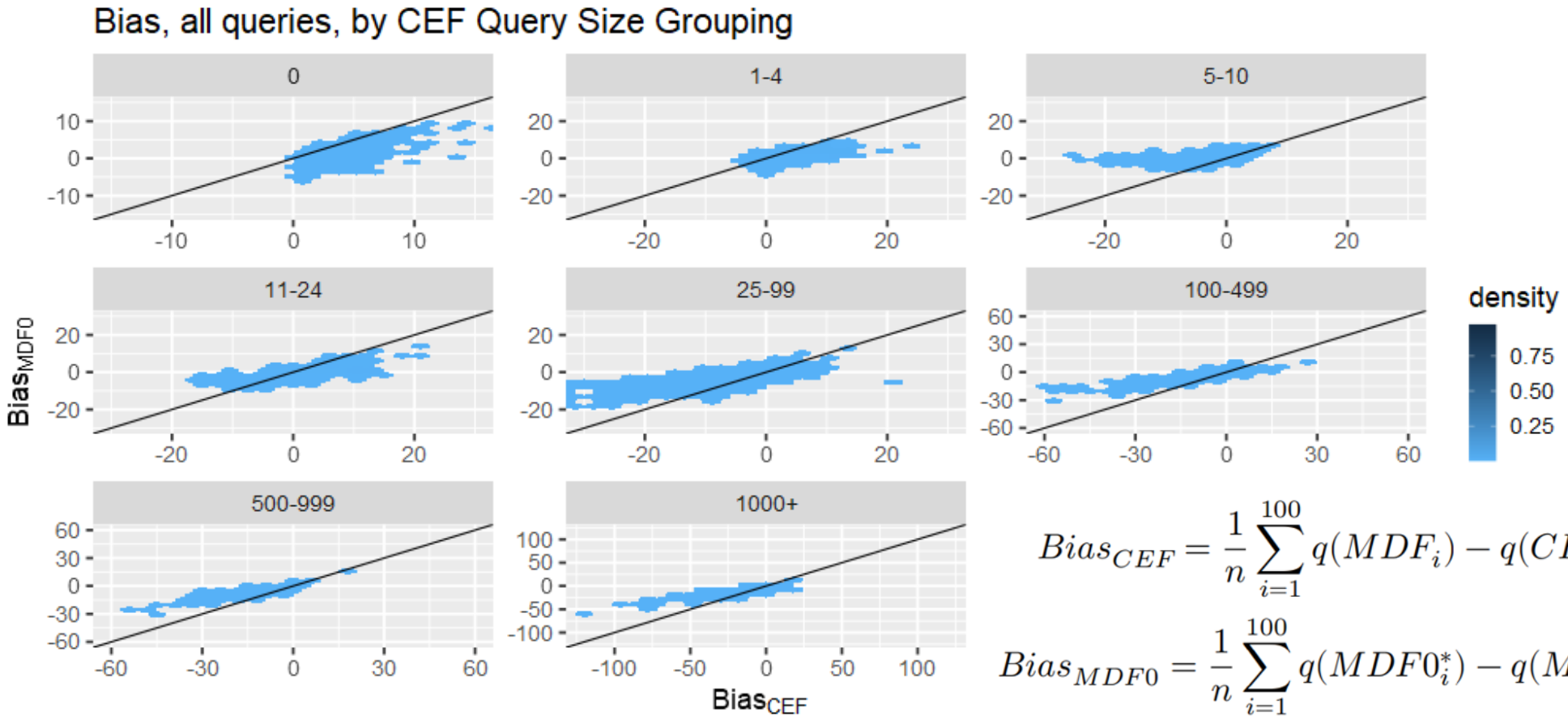


$$RMSE_{CEF} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF_i) - q(CEF))^2}$$

$$RMSE_{MDF0} = \sqrt{\frac{1}{n}\sum_{i=1}^{100}(q(MDF0_i^*) - q(MDF0))^2}$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

Block Level Bias Estimation (Sample)

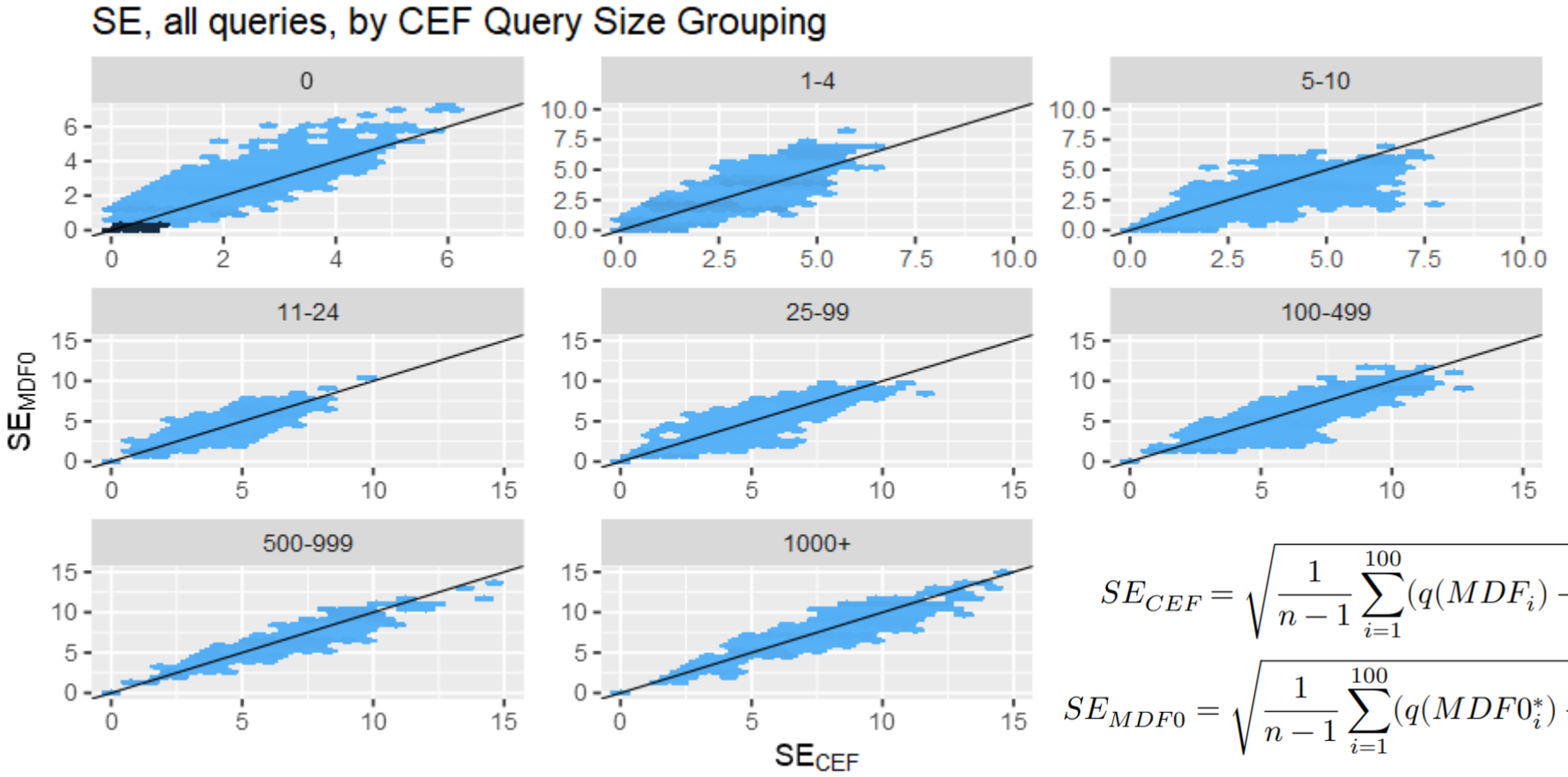


$$Bias_{CEF} = \frac{1}{n}\sum_{i=1}^{100} q(MDF_i) - q(CEF)$$

$$Bias_{MDF0} = \frac{1}{n}\sum_{i=1}^{100} q(MDF0_i^*) - q(MDF0)$$

Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

Block Level SE Estimation (Sample)



$$SE_{CEF} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF_i) - \overline{q(MDF)})^2}$$

$$SE_{MDF0} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(q(MDF0_i^*) - \overline{q(MDF0^*)})^2}$$

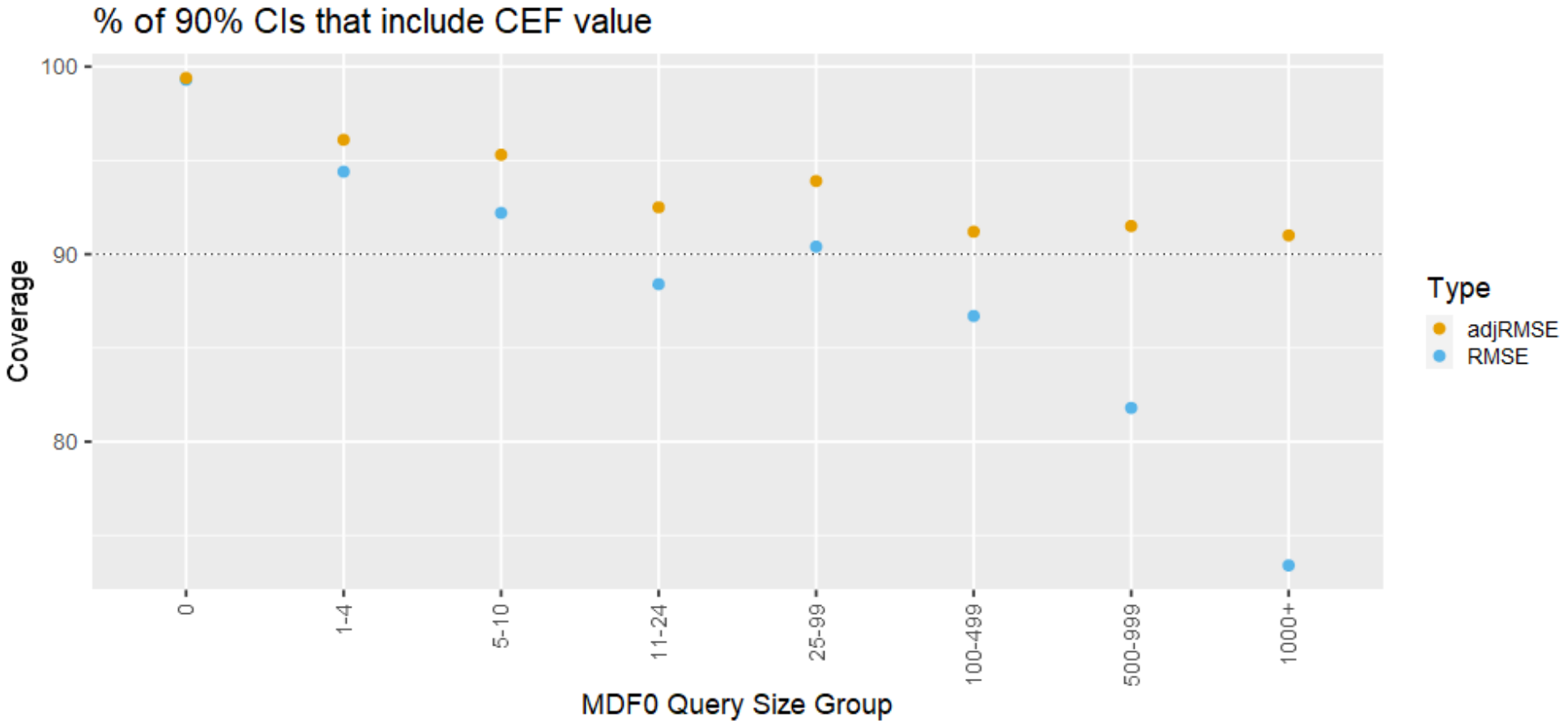
Points near the 45 degree line indicate the MDF0-based approach accurately estimates the CEF-based approach

CI Examples Block Level

QUERY	Description	Geoid	MDF0	90% CI	CI Width
P0010002	Population of one race:	440070144003020	4	0-14	14
P0030009	Population of two or more races	170190060001041	11	5-17	12
P0010024	Asian; Some Other Race	440090514001017	1	0-4	4
P0020005	White alone	440070131022010	7	0-15	15
P0020003	Not Hispanic or Latino:	470370136014008	40	31-49	18
P0010026	Population of three races:	080310068144007	4	0-13	13
P0040011	Population of two or more races	410050215002031	1	0-4	4

United States® Census Bureau

90% CI Results – Block Level



United States® Census Bureau

CI Examples, AIAN Areas

QUERY	Description	geoid	MDF0	90% CI	CI Width
P0020001	Total:	4740	96	93-99	6
P0040012	Population of two races	5111	81	67-95	28
P0010047	Population of four races:	2375	5	0-11	11
P0030009	Population of two or more races	0525	55	41-69	28
P0020002	Hispanic or Latino	6495	1	0-6	6
P0010010	Population of two races:	1800	7	0-19	19
P0030011	White; Black or African American	4710	1	0-4	4

United States® Census Bureau