U.S. Census Bureau Workshop on Using 2020 Census Data

New guidance and resources for assessing the fitness-for-use of differential privacy-adjusted Census data

Population Association of America
April 17, 2024

Workshop on Using 2020 Census Data

Session I: Welcome, Objectives, and Introductions

Tori Velkoff
Associate Director for
Demographic Programs

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 ongoing research and we'll introduce you to some new tools and resources to help you as you use DP-adjusted census data.

United States® Census Bureau

Today's Agenda

1:00pm – 1:30pm	Welcome and Introductions
1:30pm – 2:15pm	Disclosure Avoidance Background and Updates
2:15pm – 2:30pm	Break
2:30pm – 3:30pm	New Guidance and Resources for Data Users
3:30pm – 3:45pm	Break
3:45pm – 4:30pm	Confidence Intervals
4:30pm – 5:00pm	Discussion and Wrap-up

United States® Census Bureau

Who are we?

- Tori Velkoff
- Robert Ashmead
- Cassandra Dorius
- Michael Hawes
- Beth Jarosz
- Alexandra Krause
- 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 Using 2020 Census Data

Session II: Disclosure Avoidance Background and Updates

Michael Hawes
Research and Methodology

United States® Census Bureau

2020 Census Data Collection

- Vacancy status
- Tenure
- Relationship to Householder
- Sex
- Age and date of birth
- Hispanic origin
- Race

3. Is this house, apartment, or mobile home — *Mark ☒ ONE box.*

- ☐ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
- ☐ Owned by you or someone in this household free and clear (without a mortgage or loan)?
- ☐ Rented?
- ☐ Occupied without payment of rent?

3. How is this person related to Person 1? *Mark ☒ ONE box.*

- ☐ Opposite-sex husband/wife/spouse
- ☐ Opposite-sex unmarried partner
- ☐ Same-sex husband/wife/spouse
- ☐ Same-sex unmarried partner
- ☐ Biological son or daughter
- ☐ Adopted son or daughter
- ☐ Stepson or stepdaughter
- ☐ Brother or sister
- ☐ Father or mother
- ☐ Grandchild
- ☐ Parent-in-law
- ☐ Son-in-law or daughter-in-law
- ☐ Other relative
- ☐ Roommate or housemate
- ☐ Foster child
- ☐ Other nonrelative

4. What is this person's sex? *Mark ☒ ONE box.*

☐ Male ☐ Female

5. What is this person's age and what is this person's date of birth? *For babies less than 1 year old, do not write the age in months. Write 0 as the age.*

Print numbers in boxes.

Age on April 1, 2020 — years | Month | Day | Year of birth

➜ NOTE: Please answer BOTH Question 6 about Hispanic origin and Question 7 about race. For this census, Hispanic origins are not races.

6. Is this person of Hispanic, Latino, or Spanish origin?

- ☐ **No**, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin – *Print, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc.*

7. What is this person's race?
*Mark ☒ one or more boxes **AND** print origins.*

- ☐ White – *Print, for example, German, Irish, English, Italian, Lebanese, Egyptian, etc.*
- ☐ Black or African Am. – *Print, for example, African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc.*
- ☐ American Indian or Alaska Native – *Print name of enrolled or principal tribe(s), for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.*
- ☐ Chinese
- ☐ Filipino
- ☐ Asian Indian
- ☐ Other Asian – *Print, for example, Pakistani, Cambodian, Hmong, etc.*
- ☐ Vietnamese
- ☐ Korean
- ☐ Japanese
- ☐ Native Hawaiian
- ☐ Samoan
- ☐ Chamorro
- ☐ Other Pacific Islander – *Print, for example, Tongan, Fijian, Marshallese, etc.*
- ☐ Some other race – *Print race or origin.*

United States® Census Bureau

2020 Census Data Products

Released in 2021

Apportionment
National and state-level total population counts

Redistricting File (Public Law 94-171)
Voting age population by major race and ethnic groups, housing occupancy, and Group Quarters

Released in 2023

Demographic and Housing Characteristics File (DHC)
&
Demographic Profile
Additional demographic and housing characteristics for major race and ethnic groups

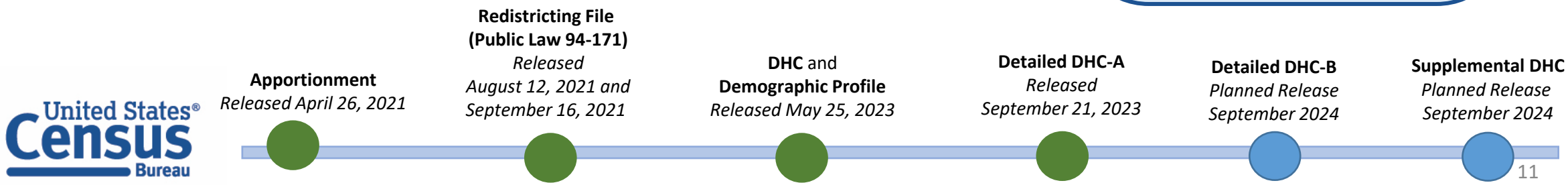
Detailed DHC-A
Total population and sex by age data for detailed race and ethnic groups and AIAN tribes and villages

Planned Release by September 2024

Detailed DHC-B
Household type and tenure data for disaggregated race and ethnic groups and AIAN tribes and villages

Supplemental DHC (S-DHC)
Counts of people in certain household types, including averages by major race and ethnic groups

Privacy-Protected Microdata File (PPMF)



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 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
- Privacy-Protected Microdata File (PPMF)

SafeTab PHSafe

Produce privacy-protected tabulations

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

United States® Census Bureau

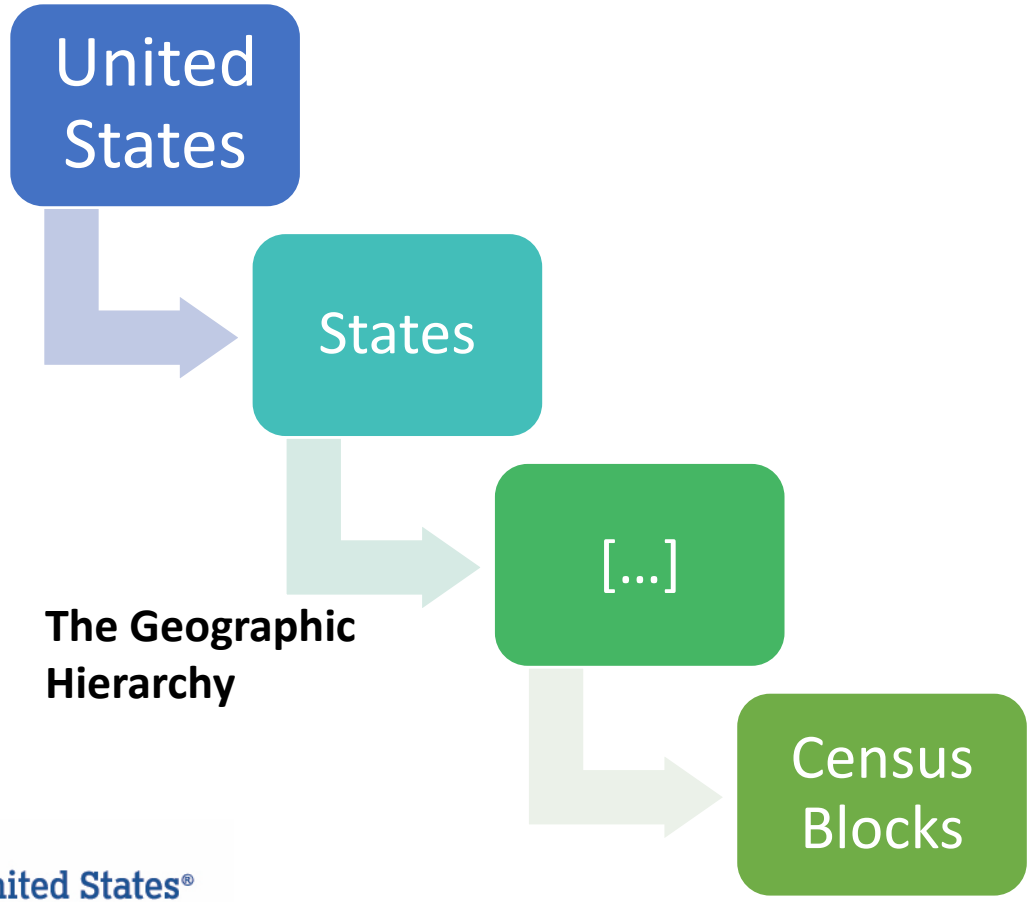
The TopDown Algorithm

*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

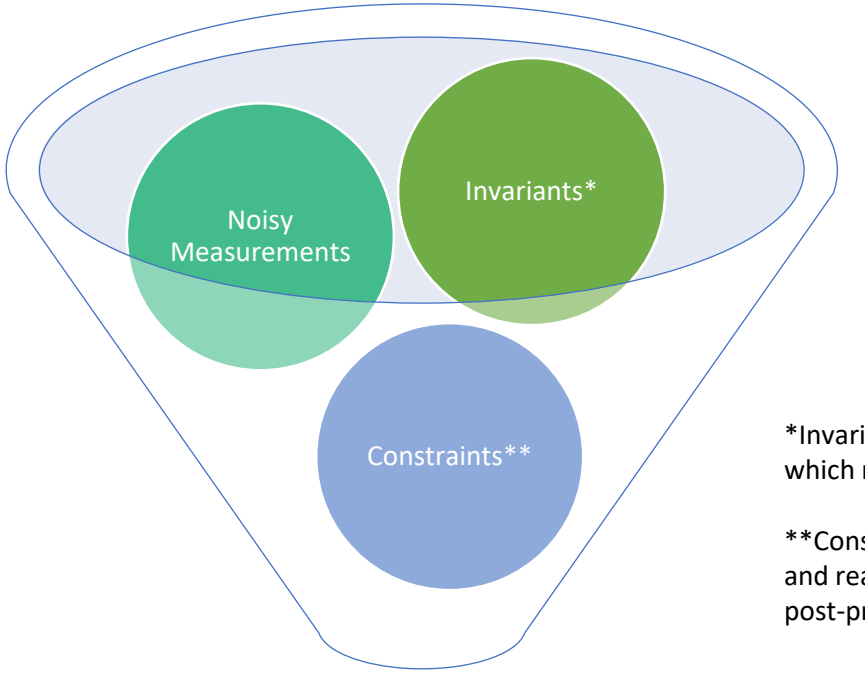
United States® Census Bureau

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

Queries and Privacy-loss Budget Allocation

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

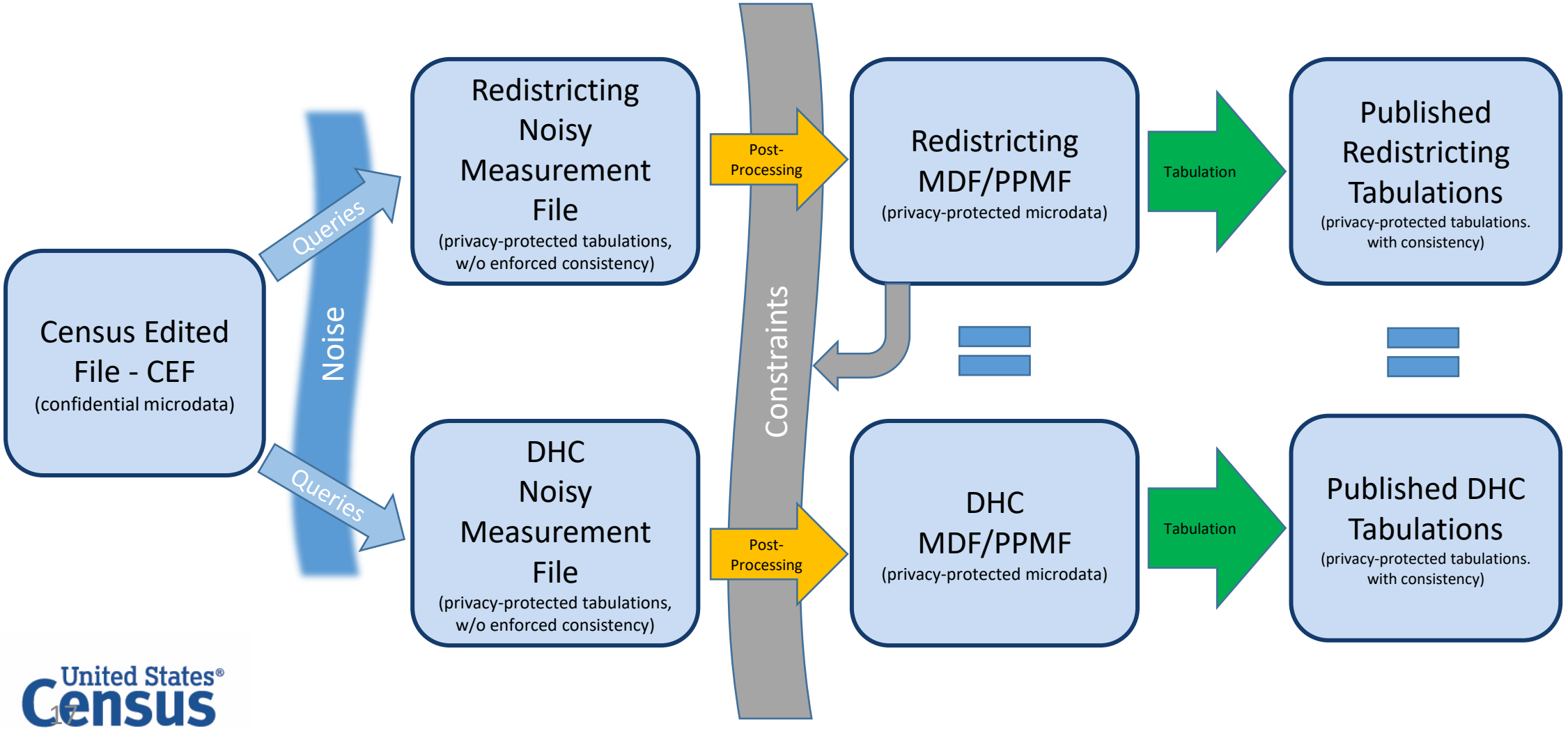
	***rho* Allocation by Geographic Level**
US	2.54%
State	35.13%
County	10.91%
Tract	16.76%
Optimized Block Group*	30.64%
Block	4.03%

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)		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%

Complete PLB allocations for the 2020 Census Redistricting Data (P.L. 94-171) Summary File and Demographic and Housing Characteristics File are available at: DHC File - Privacy-loss Budget Allocation for Person Tables Production Settings 2023-04-03 (census.gov)

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



Assessing Fitness for Use

Guidance and Resources for Data Users

United States® Census Bureau

2020 Accuracy Measures

Direct comparisons of the 2020 Census DHC to the 2020 Census Edited File (CEF)

Available at: https://www2.census.gov/programs-surveys/decennial/2020/data/demographic-and-housing-characteristics-file/2020-Census-Disclosure-Avoidance-System-Detailed-Summary-Metrics.xlsx

United States® Census Bureau

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)

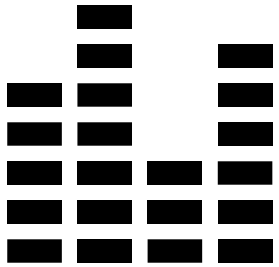
United States® Census Bureau

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

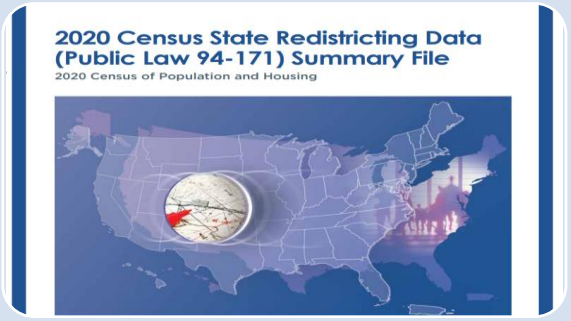


Post-processing

- 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

Using the NMFs to Generate Unbiased Confidence Intervals

- Recorded webinar on how to use the NMFs available here: Noisy Measurement Files for the Redistricting and DHC Data Products (census.gov)

- Sample code, examples, and useful tips: Computing Confidence Intervals Using the 2010 Census Redistricting (P.L. 94-171) Demonstration Data Noisy Measurement File (2023-04-03) by Ryan Cumings-Menon, Michael Hawes, and Matthew Spence (April 2023)

United States® Census Bureau

Generating Confidence Intervals

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: 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

How do these estimates compare?

2010 Total Population	SF1	2010 PPMF	2010 NMF
Bullock County, AL	10,914	10,912	10,911.61 +/- 3.19
Menominee, WI	4,232	4,230	4,232 +/- 5
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

Reader-Friendly Disclosure Avoidance Briefs

- Disclosure Avoidance and the 2020 Redistricting Data
- Why the Census Bureau Chose Differential Privacy
- Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works

More resources are in development, as well as additional specific guidance and training for using the 2020 Census data.

United States® Census Bureau

Updated Research on the Census Bureau's Evaluation of the 2010 Disclosure Avoidance System

United States® Census Bureau

Technical Paper:
https://www.census.gov/library/working-papers/2023/adrm/CES-WP-23-63.html

Github Replication Package:
https://github.com/uscensusbureau/recon_replication

The 2010 Census Confidentiality Protections Failed, Here's How And Why

December 2023

Written by: JOHN M. ABOWD, TAMARA ADAMS, ROBERT ASHMEAD, DAVID DARAIS, SOURYA DEY, SIMSON L. GARFINKEL, NATHAN GOLDSCHLAG, DANIEL KIFER, PHILIP LECLERC, ETHAN LEW, SCOTT MOORE, ROLANDO A. RODRÍGUEZ, RAMY N. TADROS, AND LARS VILHUBER

Working Paper Number CES-23-63

Abstract

Using only 34 published tables, we reconstruct five variables (census block, sex, age, race, and ethnicity) in the confidential 2010 Census person records. Using the 38-bin age variable tabulated at the census block level, at most 20.1% of reconstructed records can differ from their confidential source on even a single value for these five variables. Using only published data, an attacker can verify that all records in 70% of all census blocks (97 million people) are perfectly reconstructed. The tabular publications in Summary File 1 thus have prohibited disclosure risk similar to the unreleased confidential microdata. Reidentification studies confirm that an attacker can, within blocks with perfect reconstruction accuracy, correctly infer the actual census response on race and ethnicity for 3.4 million vulnerable population uniques (persons with nonmodal characteristics) with 95% accuracy, the same precision as the confidential data achieve and far greater than statistical baselines. The flaw in the 2010 Census framework was the assumption that

Download The 2010 Census Confidentiality Protections Failed, Here's How And Why [PDF - > 1MB]

Related Information

Center for Economic Studies (CES) Working Paper Series

PUBLICATION
Discussion Paper Series

United States® Census Bureau

Planning for 2030

United States® Census Bureau

We are establishing a robust research, development, and evaluation program for disclosure avoidance for the 2030 Census.

Having learned from the 2020 Census experience, our goal is to have all major decision-making and development completed by the 2028 Dress Rehearsal.

Our research program has two major paths

Our research program has two major paths

Communication

We are also researching better ways to conceptualize and discuss disclosure risk, and more effective mechanisms for communicating about uncertainty in published statistics.

Questions?

2020DAS@census.gov

BREAK

Workshop on Using 2020 Census Data

Session III: New Guidance and Resources for Data Users

United States® Census Bureau

2020 Census Disclosure Avoidance Briefs

Status Update

Beth Jarosz
Population Reference Bureau

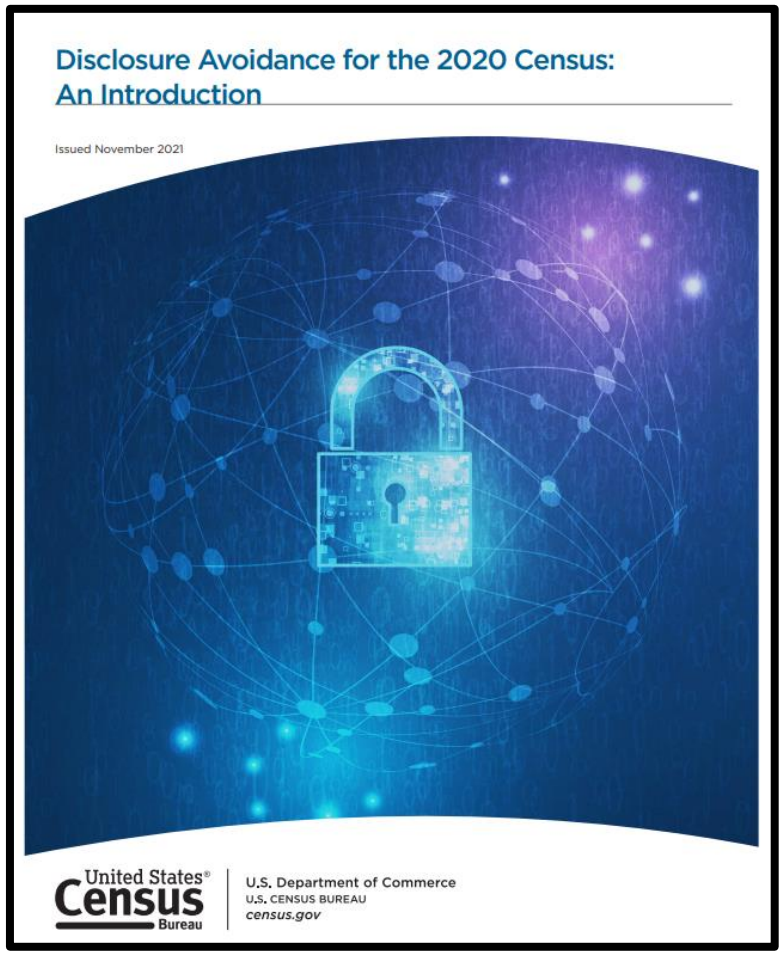
Cynthia Davis Hollingsworth
Decennial Census Programs

Population Association of America Pre-Conference Workshop
April 17, 2024

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.

United States® Census Bureau

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

United States® Census Bureau

Progress to Date

- 4 briefs completed + 1 in production
- More graphics
- Focus on concise language, but not always “brief”
- External review on all briefs
  - Each brief has had at least two (usually three) external reviewers.
  - Reviewers represent a variety of different data user communities including various geographies and subjects of expertise.
  - Suggestions are incorporated where possible—even sometimes when it means going back to the drawing board.

United States® Census Bureau

Briefs on Disclosure Avoidance for the 2020 Census

Briefs 1-3

#1: Disclosure Avoidance and the 2020 Redistricting Data

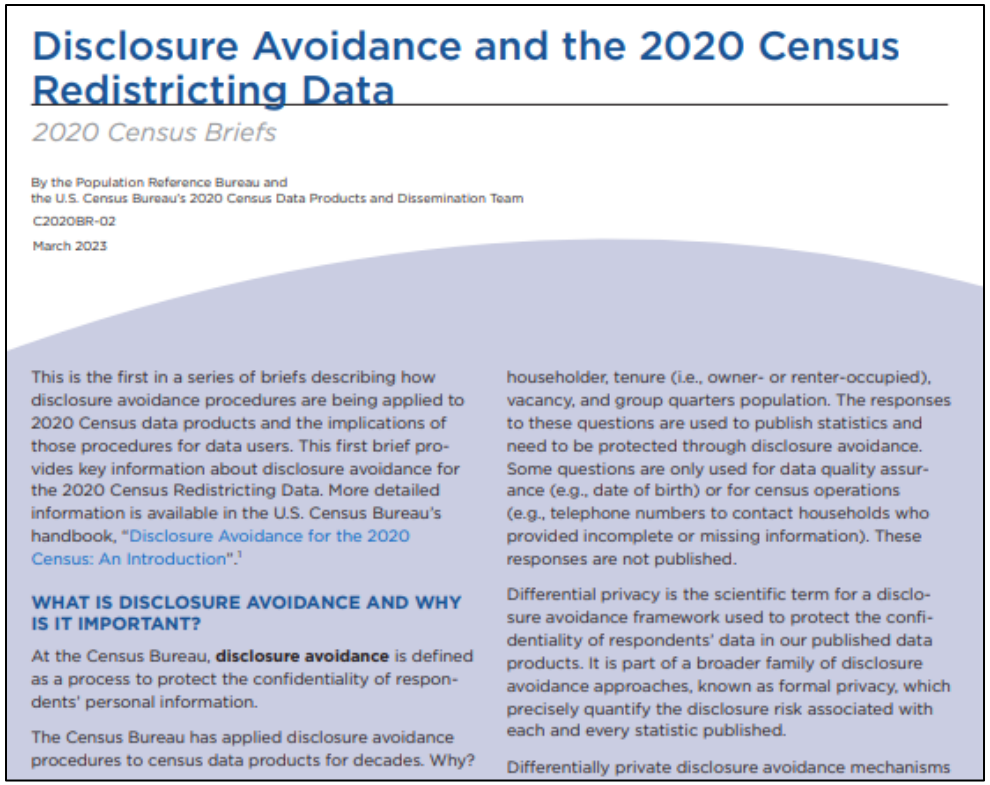
A summary of key points from the previously released Redistricting Data (P.L. 94-171) Summary File handbook.
https://www.census.gov/library/publications/2023/decennial/c2020br-02.html

#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

Briefs 4-5

#4: Disclosure Avoidance and the 2020 Census: How SafeTab-P Works

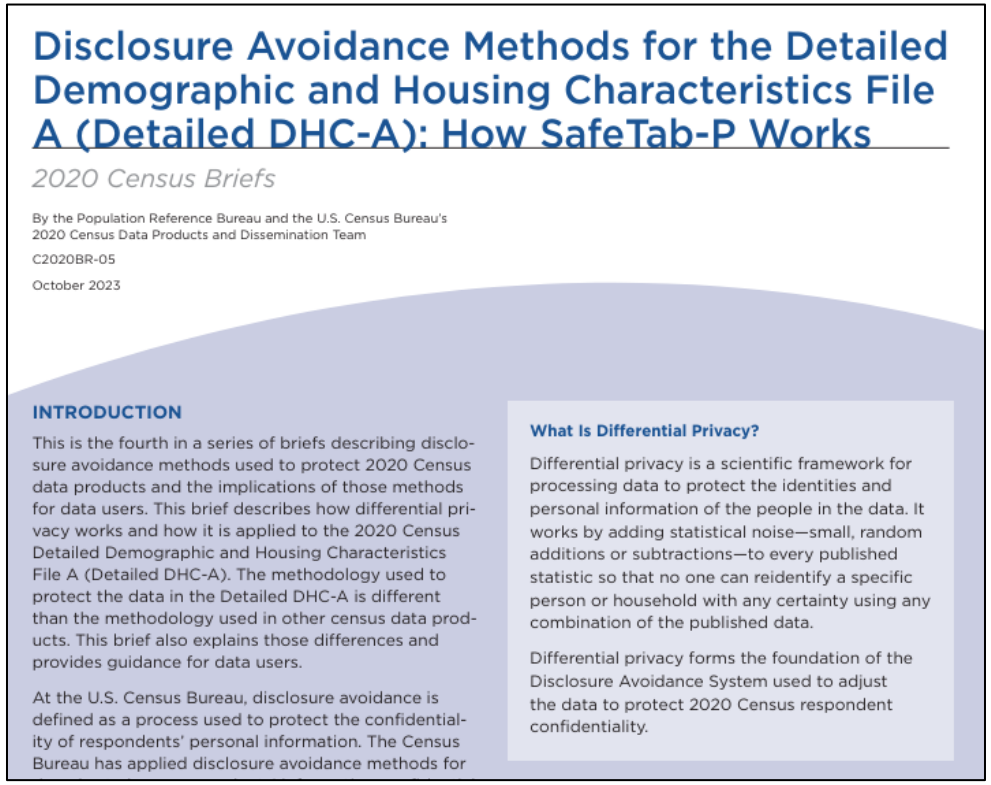
Describes how differential privacy works and is applied to the 2020 Census Detailed Demographic and Housing Characteristics File A (Detailed DHC-A).

https://www.census.gov/library/publications/2023/decennial/c2020br-05.html

#5: Disclosure Avoidance for the Demographic and Housing Characteristics File (DHC) and Guidance for Data Users

Will provide examples of how to interpret disclosure avoidance-related noise for specific use cases.

Forthcoming



Disclosure Avoidance Methods for the Detailed Demographic and Housing Characteristics File A (Detailed DHC-A): How SafeTab-P Works

2020 Census Briefs

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

C2020BR-05

October 2023

INTRODUCTION

This is the fourth in a series of briefs describing disclosure avoidance methods used to protect 2020 Census data products and the implications of those methods for data users. This brief describes how differential privacy works and how it is applied to the 2020 Census Detailed Demographic and Housing Characteristics File A (Detailed DHC-A). The methodology used to protect the data in the Detailed DHC-A is different than the methodology used in other census data products. This brief also explains those differences and provides guidance for data users.

At the U.S. Census Bureau, disclosure avoidance is defined as a process used to protect the confidentiality of respondents' personal information. The Census Bureau has applied disclosure avoidance methods for

What Is Differential Privacy?

Differential privacy is a scientific framework for processing data to protect the identities and personal information of the people in the data. It works by adding statistical noise—small, random additions or subtractions—to every published statistic so that no one can reidentify a specific person or household with any certainty using any combination of the published data.

Differential privacy forms the foundation of the Disclosure Avoidance System used to adjust the data to protect 2020 Census respondent confidentiality.

United States® Census Bureau

What's Next?

Topics for Briefs 6-7

#6: Disclosure Avoidance and the 2020 Census: How SafeTab-H Works

Will describe how differential privacy works and is applied to the 2020 Census Detailed Demographic and Housing Characteristics File B (Detailed DHC-B) and provide guidance for working with the data.

#7: Disclosure Avoidance and the 2020 Census: How PHSafe Works

Will describe how differential privacy works and is applied to the 2020 Census Supplemental Demographic and Housing Characteristics File (S-DHC) and provide guidance for working with the data.

United States® Census Bureau

Thank You!

Beth Jarosz and Cynthia Davis Hollingsworth

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

United States® Census Bureau

2020 Census Demographic and Housing Characteristics File (DHC) Averages

Population Association of America (PAA)

April 17, 2024

Alexandra Krause
Population Division

Background for History of DHC and S-DHC

United States® Census Bureau

DHC

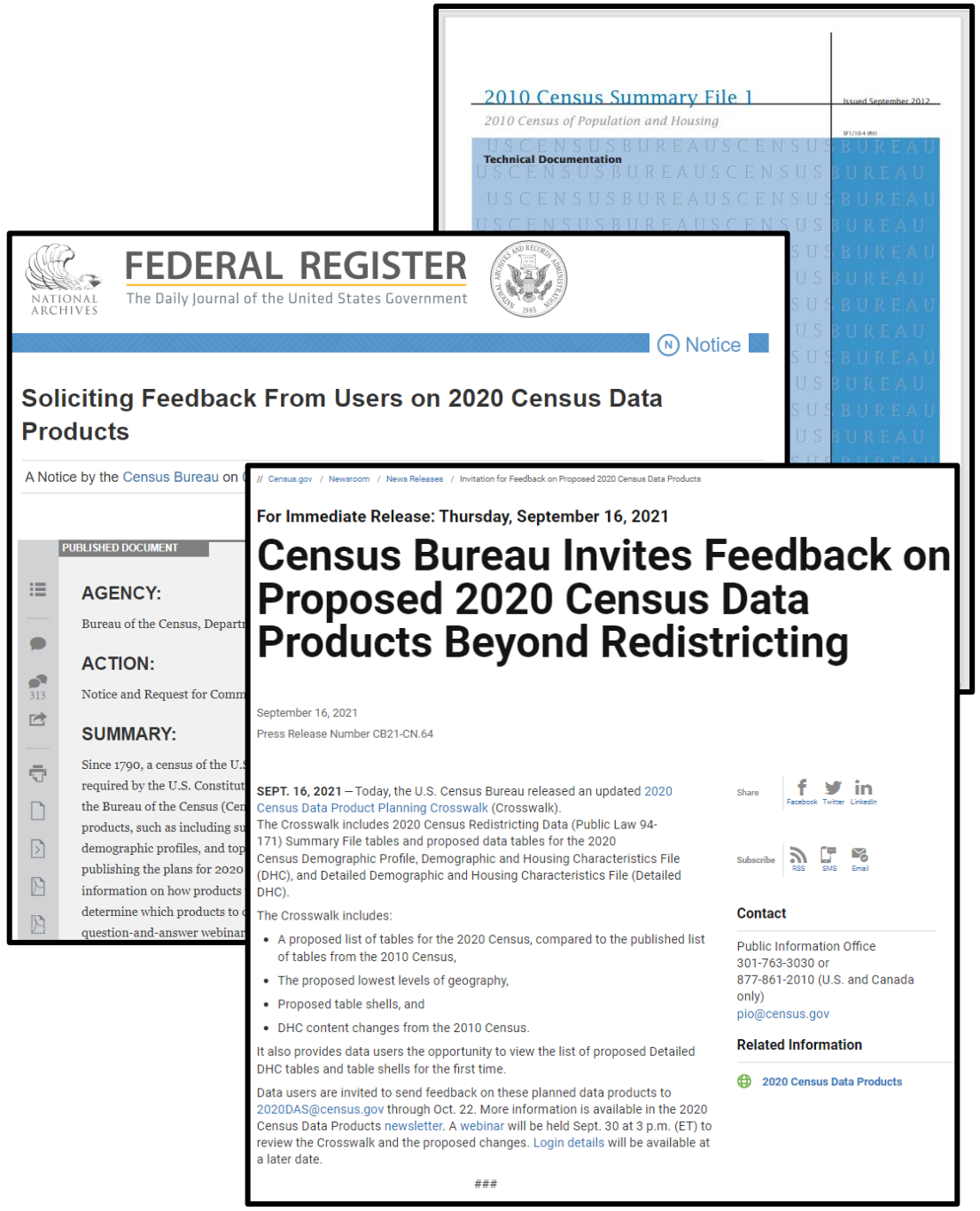
Includes many of the demographic and housing tables previously included in 2010 Summary File 1. Some tables are repeated by race and ethnicity.

- **Released**: May 25, 2023
- **Subjects include**:
  - Sex by single year-of-age
  - Hispanic or Latino origin of householder by race of householder
  - Group Quarters 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 Top-Down Algorithm (TDA)

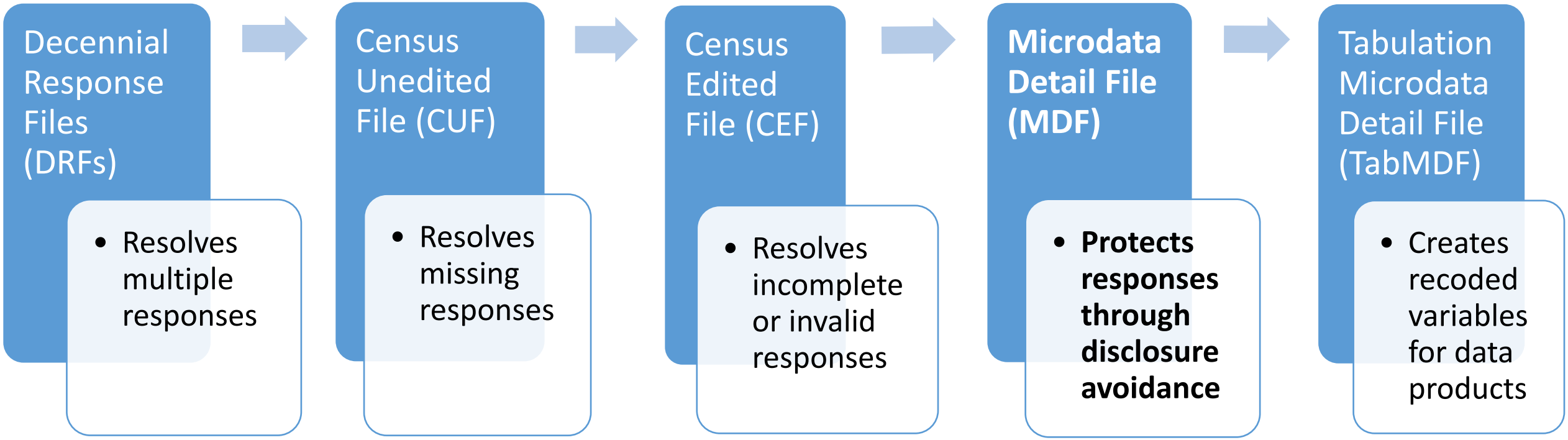
United States® Census Bureau

DHC Table Proposal Development

- Updated 2010 Summary File 1 based on 2018 Federal Register Notice (FRN) and Census Bureau subject matter expertise
- **Added and removed content based on decision to not maintain the person and household join**
- Added content and geographies based on data user feedback
  - Kept block-level tables
  - For tables available at state and county level, lowered geography to tract level (when offered at tract or block for 2010 Census)
  - Added sex by single year of age iterations at tract level



DHC Processing: Original Plan



United States® Census Bureau

DHC Processing: Person and Unit Files

- Throughout processing, there are person and unit files
- Can merge them on DRF through CEF files to learn about individual's household
- Early Decision: cannot merge them on the MDF or TabMDF
  - Without link, can only tabulate information about individuals or households

MDF Person Variables
Record type
Group Quarters type
Relationship to householder
Sex
Age
Hispanic origin
Race

MDF Housing Unit Variables
Record type
Group Quarters type
Tenure
Vacancy status
Household size

United States® Census Bureau

Examples of Tables Impacted

P16 | HOUSEHOLD TYPE

Decennial Census | Universe: Households | 2020: DEC 118th Congressional District Summary File

Label
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

PCT15 | COUPLED HOUSEHOLDS, BY TYPE

Decennial Census | Universe: Households | 2020: DEC Demographic and Housing Characteristics

Label
Total:
Married couple household:
Opposite-sex married couple household
Same-sex married couple household:
Male-male married couple households
Female-female married couple households
Unmarried-partner household:
Opposite-sex unmarried partner household
Same-sex unmarried partner households:
Male-male unmarried partner households
Female-female unmarried partner household
All other households

PCT14 | PRESENCE OF MULTIGENERATIONAL HOUSEHOLDS

Decennial Census | Universe: Households | 2020: DEC 118th Congressional District Summary File

Label
Total:
Household has three or more generations
Household does not have three or more generations

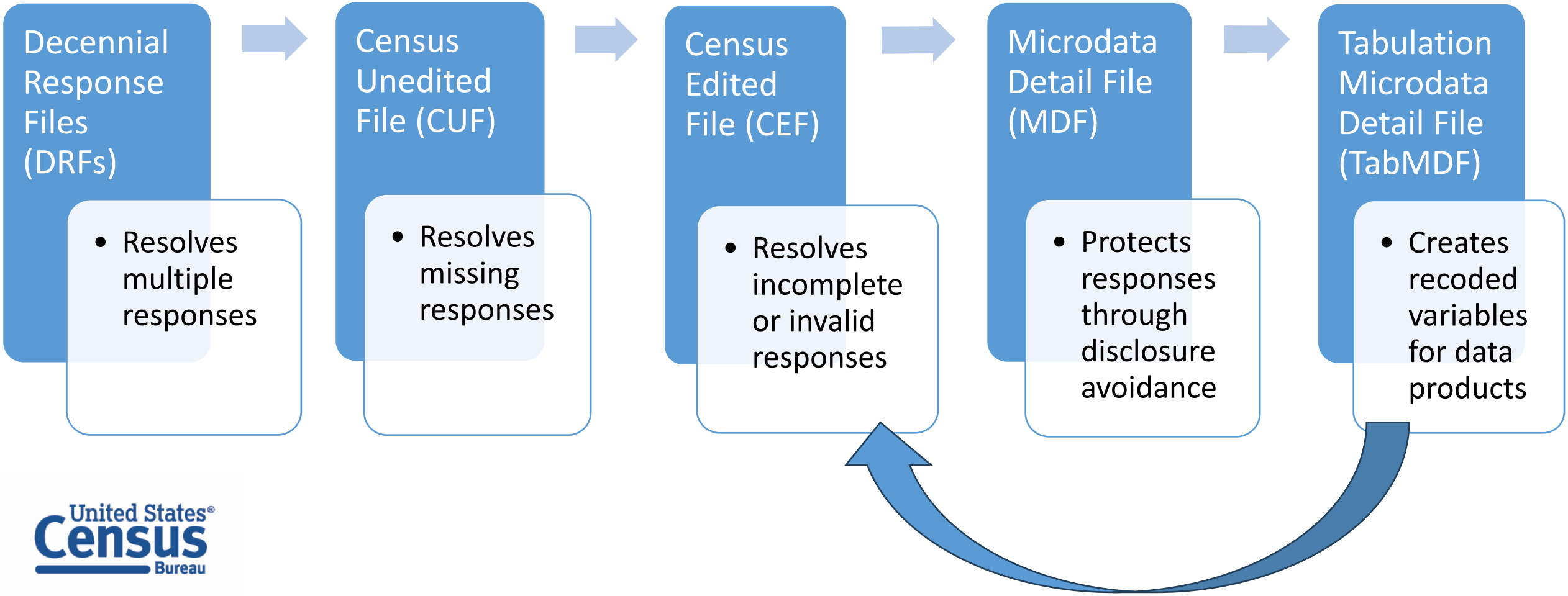
H15 | TENURE BY PRESENCE OF PEOPLE UNDER 18 YEARS (EXCLUDING HOUSEHOLDERS, SPOUSES, AND UNMARRIED PARTNERS)

Decennial Census | Universe: Occupied housing units | 2020: DEC 118th Congressional District Summary ...

Label
Total:
Owner occupied:
With children under 18 years
No children under 18 years
Renter occupied:
With children under 18 years
No children under 18 years

United States® Census Bureau

DHC Processing: Updated Plan



United States® Census Bureau

DHC Processing: Added Recoded Variables to CEF

- Added recoded variables to the CEF as input into the disclosure avoidance system to retain tables

Added Person Recoded Variable
Person Living Alone

Added Housing Unit Recoded Variables
Household/Family Type
Household/Family Type (Includes Cohabiting)
Couple Type
Presence and Type of Unmarried Partner Household
Multigenerational Household
Sex of Householder
Presence and Age of Children Under 18
Presence and Age of Own Children Under 18
Presence of People Under 18 Years in Household
Presence of People 60 Years and Over in Household
Presence of People 65 Years and Over in Household
Presence of People 75 Years and Over in Household

Housing Unit Recoded Variables that Already Existed
Age of Householder
Hispanic Householder
Race of Householder

United States® Census Bureau

S-DHC

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

- **Planned release date**: September 2024
- **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*
- **Truncation**: 10 (total household population) and 6 (children under 18 years)
- **Proposed levels of geography**: Nation and state
- **Disclosure avoidance**: Differentially private PHSafe algorithm

United States® Census Bureau

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

S-DHC Example Table

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

Major Race and Ethnic Groups

A. White alone
B. Black or African American alone
C. American Indian or Alaska Native (AIAN) alone
D. Asian alone
E. Native Hawaiian and Other Pacific Islander (NHPI) alone
F. Some Other Race (SOR) alone
G. Two or More Races
H. Hispanic or Latino
I. White alone, not Hispanic or Latino

United States® Census Bureau

S-DHC Table Proposal Development

- Started with tables requiring person and unit file join that could not be produced in the DHC
- Removed tables based on anticipated privacy-loss budget and Census Bureau subject matter expertise
- Developed initial proposal based on preliminary analysis of disclosure risk and accuracy
  - Considered additional geographies based on public feedback
- Removed geographies after completing extensive testing with disclosure avoidance settings and modeling
  - Accuracy did not meet Census standards given the privacy-loss budget allocated for the data product

United States® Census Bureau

DHC Average Guidance

United States® Census Bureau

DHC Average Methods

- Resources
  - Calculating and Interpreting Average Household Size Ratios in the DHC
  - Brief 5: Disclosure Avoidance Methods for the DHC and Implications for Data Users
- Analysis based on 2010 Demonstration Data Products Suite (Production Settings 2023-04-03) Privacy-Protected Microdata File (PPMF), compared to 2010 Census published data
- Examine average household size for total population and by race and Hispanic origin, voting age, and tenure
- Creating averages can result in implausible and impossible scenarios
- S-DHC provides official ratios for nation and state

Calculations: Average Household Size

Person File Only: $$\frac{\text{Count of people in households}}{\text{Count of householders}}$$

Unit File Only: $$\frac{\text{Sum of household size}}{\text{Count of occupied housing units}}$$

Combination: $$\frac{\text{Count of people in households}}{\text{Count of occupied housing units}}$$

United States® Census Bureau

Household Size is top-coded at 7+ people

Calculations: Average Household Size by Race and Hispanic Origin

Person File Only:

$$\frac{\text{Count of people in households by race and Hispanic origin}}{\text{Count of householders by race and Hispanic origin}}$$

Unit File Only:

$$\frac{\text{Sum of household size by race and Hispanic origin of householder}}{\text{Count of occupied housing units by race and Hispanic origin of householder}}$$

Combination:

$$\frac{\text{Count of people in households by race and Hispanic origin}}{\text{Count of occupied housing units by race and Hispanic origin of householder}}$$

United States® Census Bureau

Household Size is top-coded at 7+ people

Tract Results: Average Household Size

- At tract level, **Person File Only** calculation performs best for total population
  - For tracts with more households, the **Combination** calculation performs similarly
- Errors are larger for geographies with small counts
- Remaining analyses are limited to 100+ households per cell

Average Household Size for Total Population at Tract			
Number of Households per Tract	Absolute Mean Difference		
	Person Only	Unit Only	Combination
1-5	**1.40**	1.84	4.41
5-20	**0.62**	0.87	1.42
20-50	**0.24**	0.52	0.64
50-100	**0.13**	0.37	0.22
100-200	**0.07**	0.21	0.11
200-400	**0.03**	0.12	0.04
400-600	**0.02**	0.10	**0.02**
600-1000	**0.01**	0.08	**0.01**
1000-5000	**0.01**	0.05	**0.01**
5000+	**0.00**	0.03	**0.00**

Treated ratios of 0/0 as 0 instead of undefined

United States® Census Bureau

Tract Results: Race and Hispanic Origin

- For total population at tract, Person File only calculation performs best
- For race and ethnicity at tract, the best calculation varies
- Errors are larger for race and ethnicity, compared to Total Population
- Results also vary by geography (not shown on slide)

Average Household Size at Tract						
Total Population and by Race and Ethnicity Group	Person File Only		Unit File Only		Combination	
	MAE	90th Percentile	MAE	90th Percentile	MAE	90th Percentile
Total Population	**0.01**	**0.02**	0.05	0.12	0.01	0.02
A. White alone	0.05	0.10	0.05	0.11	**0.04**	**0.09**
B. Black or African American alone	0.10	0.25	0.11	0.23	**0.07**	**0.16**
C. AIAN alone	0.21	0.46	**0.18**	**0.35**	0.18	0.43
D. Asian alone	0.15	0.34	0.14	0.31	**0.11**	**0.24**
E. NHPI alone	0.28	0.59	0.37	0.65	**0.19**	**0.43**
F. SOR alone	0.17	0.37	0.21	0.44	**0.11**	**0.24**
G. Two or More Races	0.85	1.75	**0.19**	**0.40**	0.72	1.54
H. Hispanic or Latino	0.25	0.60	**0.19**	**0.40**	0.20	0.47
I. White alone, not Hispanic	0.07	0.14	**0.04**	**0.09**	0.07	0.14

Mean absolute error (MAE)

Treated ratios of 0/0 as 0 instead of undefined

Analysis is limited to geographies with 100+ households for that race or ethnic group

United States® Census Bureau

Calculations: Average Household Size by Voting Age

Combination: $\frac{\text{Count of people in households by voting age}}{\text{Count of occupied housing units}}$

United States® Census Bureau

Calculations: Average Household Size by Tenure

Unit File Only: $\dfrac{\text{Sum of household size by tenure}}{\text{Count of occupied housing units by tenure}}$

United States® Census Bureau

Household Size is top-coded at 7+ people

Tract Results: Voting Age – Under 18 Years

- Errors are larger, compared to Average Household Size by Race and Ethnicity
- Average errors and 90th percentiles are under 0.5
  - Excluding under 18 years for Two or More Races
- 3 cells have max differences greater than 1

Average Household Size for People Under 18 Years at Tract			
Total Population and by Race and Hispanic Origin	Under 18 Years		
	MAE	90th Percentile	Max
Total Population	0.0060	0.0126	0.2596
A. White alone	0.0281	0.0575	0.6090
B. Black or African American alone	0.0357	0.0870	0.5258
C. AIAN alone	0.1214	0.3005	0.8265
D. Asian alone	0.0453	0.1033	0.5742
E. NHPI alone	0.1013	0.2188	0.3233
F. SOR alone	0.0666	0.1412	0.7045
G. Two or More Races	0.6258	1.3144	3.0570
H. Hispanic or Latino	0.1306	0.3089	1.9134
I. White alone, not Hispanic	0.0438	0.0915	1.9790

United States® Census Bureau

Treated ratios of 0/0 as 0 instead of undefined
Analysis is limited to geographies with 100+ households for a given cell

Tract Results: Tenure - Owned

- Errors are larger, compared to Average Household Size by Voting Age
- Average errors and 90th percentile are below .5
  - Excluding NHPI and SOR which perform worse, compared to other race and ethnic groups
- 9 cells have max differences greater than 1

Average Household Size by Tenure at Tract			
Total Population and by Race and Hispanic Origin	Owned		
	MAE	90th Percentile	Max
Total Population	0.0576	0.1325	1.8220
A. White alone	0.0492	0.1081	2.1550
B. Black or African American alone	0.1096	0.2340	0.9071
C. AIAN alone	0.1711	0.3488	1.0872
D. Asian alone	0.1468	0.3156	1.6411
E. NHPI alone	0.4644	0.9542	1.5503
F. SOR alone	0.2696	0.5761	1.8507
G. Two or More Races	0.2032	0.3968	1.1011
H. Hispanic or Latino	0.2054	0.4602	2.7298
I. White alone, not Hispanic	0.0399	0.0862	2.1458

United States® Census Bureau

Treated ratios of 0/0 as 0 instead of undefined
Analysis is limited to geographies with 100+ households for a given cell

Guidance

- Use S-DHC for nation and state because they are official statistics
- Limit ratio calculations to larger areas to ensure adequate accuracy
  - Our analysis is limited to at least 100 households per cell
- Which Average Household Size calculation works best?
  - For total population, **Unit File Only** does not perform as well as other calculations
  - For race and ethnicity, calculation varies by group
- Accuracy varies by geography, major race and Hispanic origin groups, voting age, and tenure categories
  - Combine smaller race and ethnicity groups to get larger base populations
- Reference the Fact Sheet to understand nuanced differences rather than relying on overgeneralizations

United States® Census Bureau

Questions?

United States® Census Bureau

Back-Up Slides

United States® Census Bureau

DHC Tables Removed due to Reasons Other than Differential Privacy

2010 Table ID	Table Title	Iterated A-I
P21	HOUSEHOLDS BY AGE OF HOUSEHOLDER BY HOUSEHOLD TYPE BY PRESENCE OF RELATED CHILDREN	
P27	HOUSEHOLDS BY PRESENCE OF NONRELATIVES	
P35	FAMILIES	X
P39	FAMILY TYPE BY PRESENCE AND AGE OF RELATED CHILDREN	X
P42	GROUP QUARTERS POPULATION BY MAJOR GROUP QUARTERS TYPE	
PCT19	NONRELATIVES BY HOUSEHOLD TYPE	X
PCT22	GROUP QUARTERS POPULATION BY SEX BY MAJOR GROUP QUARTERS TYPE FOR THE POPULATION 18 YEARS AND OVER	
PCT24	HOUSEHOLD TYPE BY RELATIONSHIP FOR THE POPULATION 65 YEARS AND OVER	
HCT3	TENURE BY PRESENCE AND AGE OF RELATED CHILDREN	

United States® Census Bureau

DHC Tables Removed after Implementing Differential Privacy

2010 Table ID	Table Title
P16A-I	POPULATION IN HOUSEHOLDS BY AGE
P33	HOUSEHOLD TYPE FOR THE POPULATION UNDER 18 YEARS IN HOUSEHOLDS
P44	POPULATION SUBSTITUTED
P45-P51	ALLOCATION OF POPULATION ITEMS, RACE, HISPANIC OR LATINO ORIGIN, SEX, AGE, RELATIONSHIP, and POPULATION ITEMS FOR THE POPULATION IN GROUP QUARTERS
PCT16	HOUSEHOLD TYPE BY NUMBER OF CHILDREN UNDER 18
PCT17	PRESENCE OF UNMARRIED PARTNER OF HOUSEHOLDER BY HOUSEHOLD TYPE FOR THE POPULATION UNDER 18 YEARS IN HOUSEHOLDS
PCT20A-I	GROUP QUARTERS POPULATION BY GROUP QUARTERS TYPE
H8	TOTAL RACES TALLIED FOR HOUSEHOLDERS
H9	HISPANIC OR LATINO ORIGIN OF HOUSEHOLDERS BY TOTAL RACES TALLIED
H20	OCCUPIED HOUSING UNITS SUBSTITUTED
H21, H22	ALLOCATION OF VACANCY STATUS and TENURE

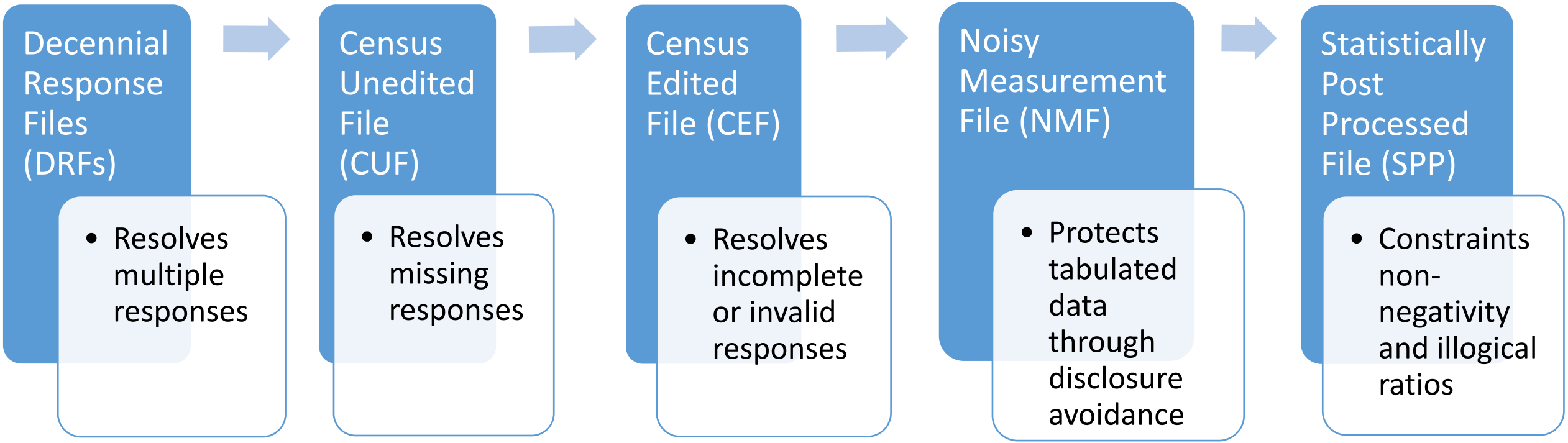
United States® Census Bureau

Additional Standardization before MDF

Housing Unit Variables, except Group Quarters Type		
Variable	**Update prior to MDF**	**Impact on Data Availability**
Group Quarters Type	Collapsed a few group quarter types	None
Household Size	Top coded at 7 or more people	None
Householder Age	Collapsed single year of age into 9 categories	None
Presence of People Under 18 Years in Household	Changed **number** to **presence** of people	None
Presence of People Over 60 Years in Household	Changed **number** to **presence** of people	None
Presence of People Over 65 Years in Household	Changed **number** to **presence** of people	None
Presence of People Over 75 Years in Household	Changed **number** to **presence** of people	None
Race of Householder	Collapsed 63 categories into 7 categories	Eliminated 2 tables and some iterations for 2 tables

United States® Census Bureau

S-DHC Processing



United States® Census Bureau

Assessing Disclosure Avoidance Uncertainty in the 2020 Census:

Determining Reliability Thresholds for Demographic and Housing Characteristics Data

Matthew Spence
Population Division, U.S. Census Bureau

United States® Census Bureau

Need to Aggregate Geographies or Groups

Even from the first brief, data users were encouraged to aggregate counts across geographies or demographic groups

3. RECOMMENDATIONS AND CONSIDERATIONS WHEN USING THE REDISTRICTING DATA

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

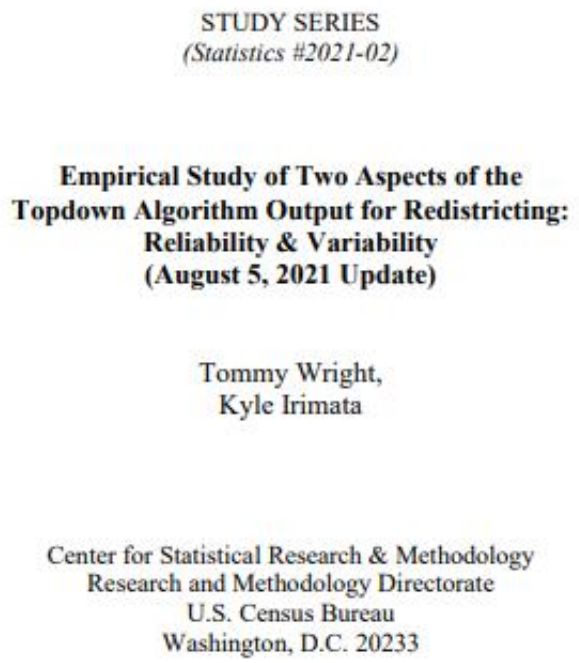
Block-level data should be aggregated before use. The amount of noise added to statistics does not depend on population or geographic size, so block-level data are most affected by disclosure avoidance procedures. For example, it is equally likely that five people could be added to an area with a population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

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

But this prompts the question: "**How much do I need to aggregate before I have reliable data?**"

United States® Census Bureau

Need to Aggregate Geographies or Groups



STUDY SERIES
(Statistics #2021-02)

Empirical Study of Two Aspects of the Topdown Algorithm Output for Redistricting: Reliability & Variability (August 5, 2021 Update)

Tommy Wright,
Kyle Irimata

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

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

– Disclosure Avoidance for the 2020 Census: An Introduction

United States® Census Bureau

Going Beyond

- Can we look at other characteristics?
- Can we quantify "reliability" relative to something other than *total population*?
- Idea: look to a table's *universe* as the basis

Table	Numerator	Denominator
H4C	Housing units owned free and clear with an American Indian and Alaska Native householder	American Indian and Alaska Native households
H5	Housing units for seasonal, recreational, or occasional use	Vacant housing units
HCT2	Owner-occupied households with own children under 6 years only	Occupied households
P12	Males under 5 years	Male population
P16	Male householder, no spouse present	Occupied households
P16	Married couple households	Occupied households
P19	Same-sex unmarried partners	Persons in households
PCO11 / custom universe denominator	Grandchildren under 3 years	Children under 3 years
PCT15 / P16	Female-female married couple households	Occupied households
PCO8 / custom universe denominator	Own children of householder under 3 years of age	Persons in households
PCT19	Males aged 18 to 64 years in emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness	Males aged 18 to 64 years in any group quarters
PCT19	Males aged 18 to 64 years in group homes intended for adults	Males aged 18 to 64 years in any group quarters

An Example: First, Calculate Differences

	Published Data			Demonstration Data			Result
Tract	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Absolute Percentage Point Difference
1	3	100	3.00%	2	99	2.02%	0.98
2	0	10	0.00%	1	11	9.09%	9.09
3	2	250	0.80%	2	249	0.80%	0.00
4	1	1,000	0.10%	2	1,003	0.20%	0.10
5	7	200	3.50%	8	199	4.02%	0.52

Hypothetical Data

United States® Census Bureau

An Example: Next, Group by Size

	Published Data			Demonstration Data			Result
Tract	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Absolute Percentage Point Difference
2	0	10	0.00%	1	11	9.09%	9.09
1	3	100	3.00%	2	99	2.02%	0.98
5	7	200	3.50%	8	199	4.02%	0.52
3	2	250	0.80%	2	249	0.80%	0.00
4	1	1,000	0.10%	2	1,003	0.20%	0.10

Hypothetical Data

Here, we're grouping by 100: 0 – 99, 100 – 199, 200 – 299, …, 1,000 – 1,100, etc.

United States® Census Bureau

An Example: Next, Compare to Baseline

	Published Data			Demonstration Data			Result
Tract	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Absolute Percentage Point Difference
2	0	10	0.00%	1	11	9.09%	9.09
1	3	100	3.00%	2	99	2.02%	0.98
5	7	200	3.50%	8	199	4.02%	0.52
3	2	250	0.80%	2	249	0.80%	0.00
4	1	1,000	0.10%	2	1,003	0.20%	0.10

Hypothetical Data

Here, we're calling anything above a five-percentage point change in share "unreliable"

United States® Census Bureau

An Example: Finally, Establish the Size Threshold

	Published Data			Demonstration Data			Result
Tract	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Female-female Married Couple Households	Households	Female-female Married Couple Households (Share)	Absolute Percentage Point Difference
2	0	10	0.00%	1	11	9.09%	9.09
1	3	100	3.00%	2	99	2.02%	0.98
5	7	200	3.50%	8	199	4.02%	0.52
3	2	250	0.80%	2	249	0.80%	0.00
4	1	1,000	0.10%	2	1,003	0.20%	0.10

Hypothetical Data

Which size categories have 90% of their geographies meeting the “reliability” standard?
What’s the smallest category where this standard is consistently met?

United States® Census Bureau

Results: Five-Percentage Point Change or Less

Part 1 of 2

Numerator	Denominator	Block Groups	Places	Tracts
Owned free and clear, AIAN householder	AIAN households	225 – 249	575 – 599	200 – 224
Housing units for seasonal, recreational, or occasional use	Vacant housing units	150 – 174	125 – 149	75 – 99
Owner-occupied with own children under 6 years only	Households	50 – 74	50 – 74	50 – 74
Males under 5	Males	75 – 99	75 – 99	125 – 149
Male householder, no spouse present households	Households	150 – 174	125 – 149	175 – 199
Married couple households	Households	125 – 149	225 – 249	75 – 99
Same-sex unmarried partners	Persons in households	0 – 24	0 – 24	0 – 24

United States® Census Bureau

Results: Five-Percentage Point Change or Less

Part 2 of 2

Numerator	Denominator	Block Groups	Places	Tracts
Grandchildren under 3	Children under 3	175 – 199	175 – 199	175 – 199
Female-female married couple households	Households	0 – 24	0 – 24	0 – 24
Own children of householder under 3 years of age	Persons in households	75 – 99	50 – 74	100 – 124
Males aged 18 – 64 in emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness	Males aged 18 – 64 in group quarters	0 – 24	0 – 24	75 – 99
Males aged 18 – 64 in group homes intended for adults	Males aged 18 – 64 in group quarters	50 – 74	150 – 174	100 – 124

United States® Census Bureau

Results: Three-Percentage Point Change or Less

Part 1 of 2

Numerator	Denominator	Block Groups	Places	Tracts
Owned free and clear, AIAN householder	AIAN households	400 – 424	975 – 999	575 – 599
Housing units for seasonal, recreational, or occasional use	Vacant housing units	450 – 474	225 – 249	125 – 149
Owner-occupied with own children under 6 years only	Households	125 – 149	125 – 149	75 – 99
Males under 5	Males	200 – 224	200 – 224	150 – 174
Male householder, no spouse present households	Households	325 – 349	250 – 274	350 – 374
Married couple households	Households	225 – 249	375 – 399	175 – 199
Same-sex unmarried partners	Persons in households	0 – 24	0 – 24	50 – 74

United States® Census Bureau

Results: Three-Percentage Point Change or Less

Part 2 of 2

Numerator	Denominator	Block Groups	Places	Tracts
Grandchildren under 3	Children under 3	475 – 499	475 – 499	375 – 399
Female-female married couple households	Households	0 – 24	0 – 24	50 – 74
Own children of householder under 3 years of age	Persons in households	125 – 149	100 – 124	125 – 149
Males aged 18 – 64 in emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness	Males aged 18 – 64 in group quarters	100 – 124	250 – 274	250 – 274
Males aged 18 – 64 in group homes intended for adults	Males aged 18 – 64 in group quarters	75 – 99	275 – 299	150 – 174

United States® Census Bureau

Takeaways

- The choice of reliability threshold matters (larger universe required to hit ±3 pp)
- How "broad" a distribution is also matters – compare those based on "all households" versus those based on a subset:

Numerator	Denominator	Block Groups	Places	Tracts
Owner-occupied with own children under 6 years only	Households	50 – 74	50 – 74	50 – 74
Married couple households	Households	125 – 149	225 – 249	75 – 99
Owned free and clear, AIAN householder	AIAN households	225 – 249	575 – 599	200 – 224
Housing units for seasonal, recreational, or occasional use	Vacant housing units	150 – 174	125 – 149	75 – 99

- Proportions with small numerators and large denominators – like same-sex unmarried partners among all persons in households – may be accurate even for small geographies
- Places (and presumably other off-spine geographies) require larger universes to hit reliability
- More research needed to actually aggregate smaller geographies and confirm findings

United States® Census Bureau

Questions?

matthew.spence@census.gov

United States® Census Bureau

BREAK

Workshop on Using 2020 Census Data

Session IV: An Approximate Monte Carlo Simulation Method For Estimating Uncertainty and Constructing Confidence Intervals for 2020 Census Statistics

Robert Ashmead
Research and Methodology

United States® Census Bureau

Disclaimer

Any views expressed are those of the authors and not those of the U.S. Census Bureau. 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.

Statistics reported in this presentation have been cleared for public release by the Census Bureau's Disclosure Review Board (DRB clearance number: CBDRB-FY24-DSEP-0002).

Background

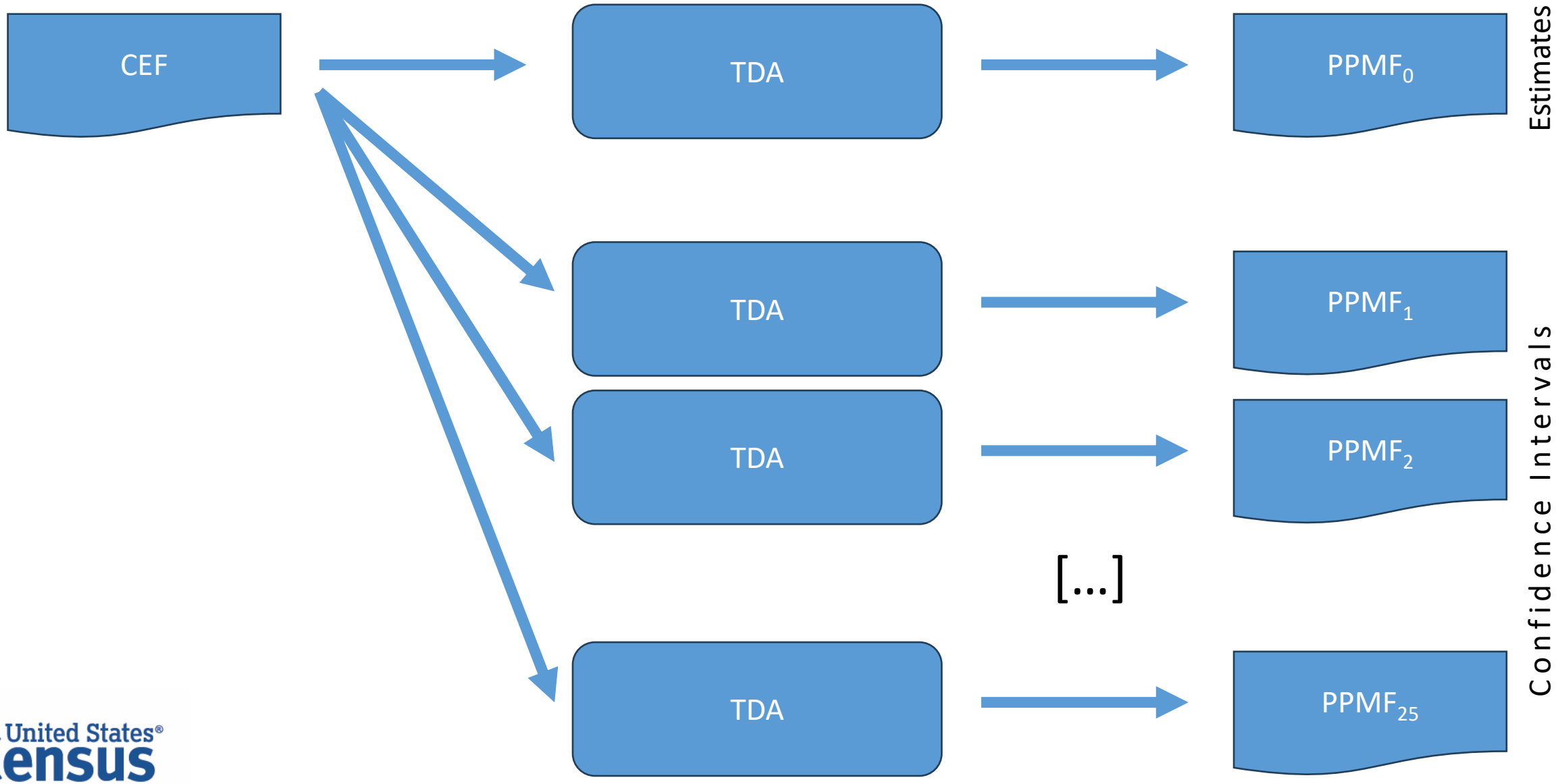
- 2020 Census Redistricting Data (P.L. 94-171) were created from the Disclosure Avoidance System's (DAS's) TopDown Algorithm (TDA) which used differential privacy (DP) to protect respondent information.
- Small amounts of "noise" (error) were infused into published tabulations
- Ideally, users can take into account the statistical properties of error to adapt their statistical analyses
  - Margins of error, confidence intervals, etc.

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/mean-squared error (MSE) formulas for tabulation queries
  - The variance/bias/MSE can depend on the underlying query (e.g. how small or large it is), not just the randomness in the noisy measurements

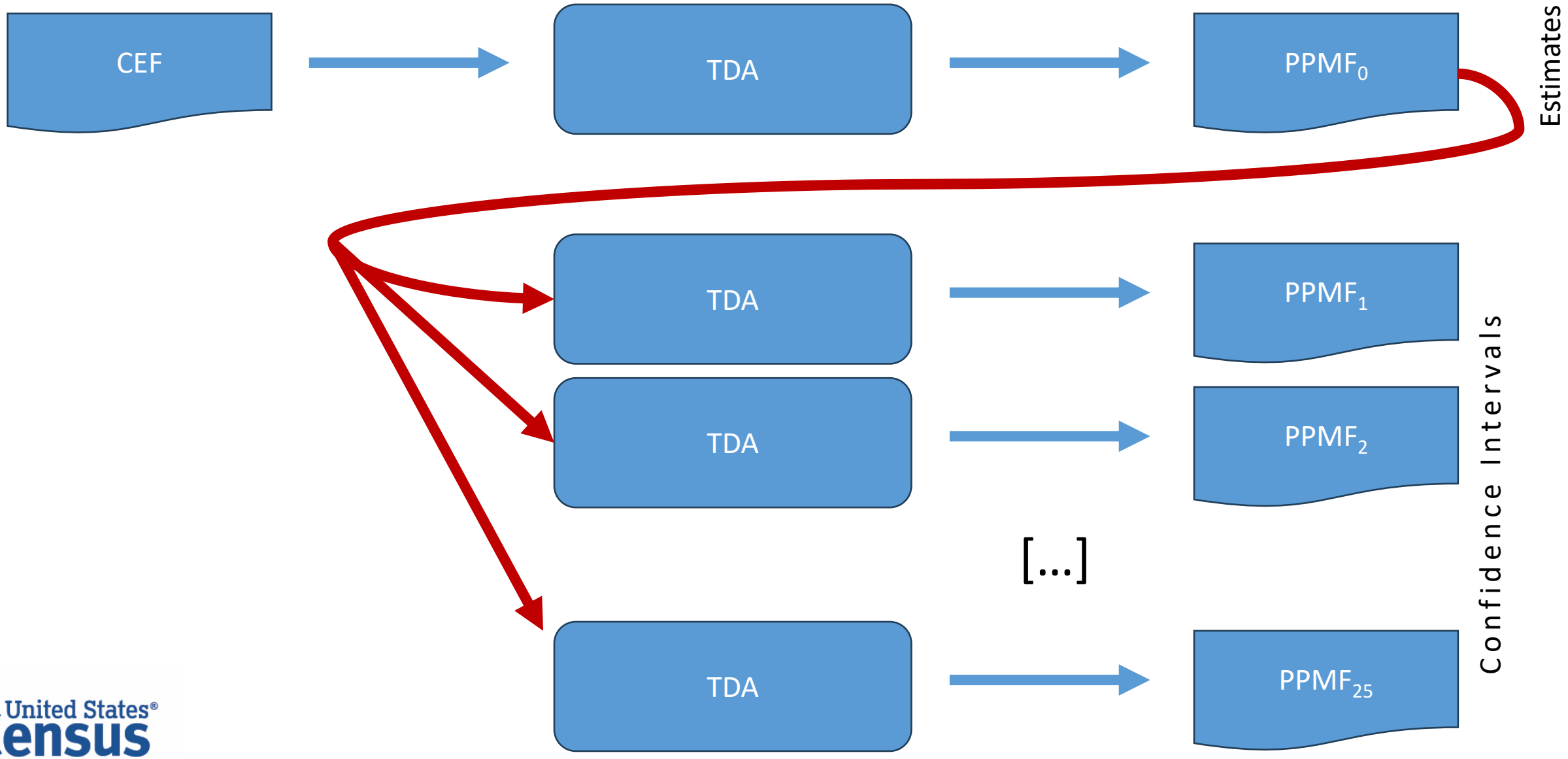
United States® Census Bureau

Standard Monte Carlo Process



United States® Census Bureau

Approximate Monte Carlo Process



United States® Census Bureau

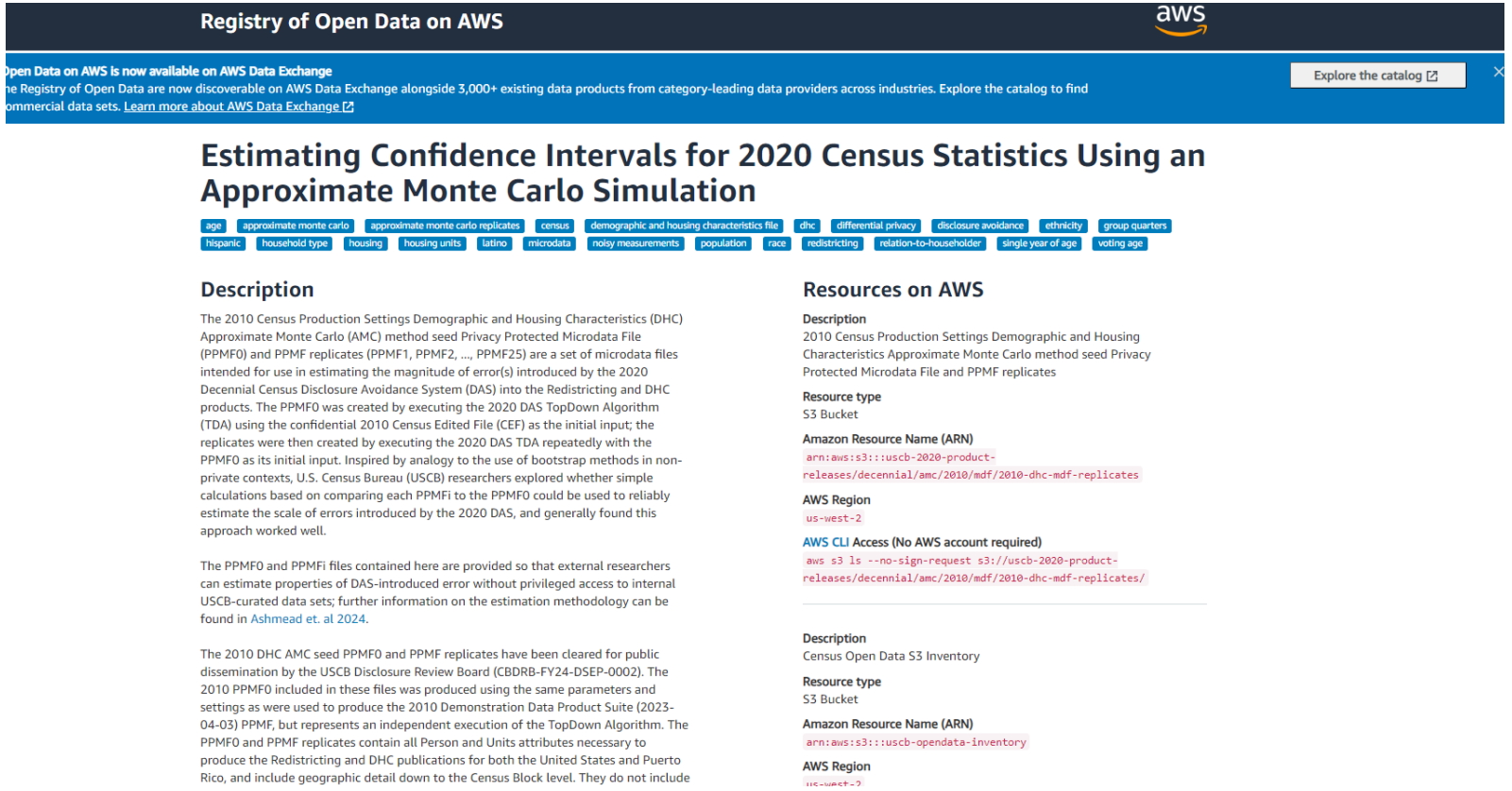
Interactive Example of Proposed Method

https://github.com/uscensusbureau/AMC_Confidence_Intervals

Proposed Solution

- Approximate Monte Carlo (AMC) Simulation Method
- Utilizes the privacy-protected microdata file (PPMF) as the starting point to simulate multiple iterations (25) of the TopDown Algorithm
- Makes calculations based on the observed variation across the iterations and relative to the PPMF
- Never accesses the confidential Census Edited File (CEF) data and thus does not contribute additional privacy-loss
- Allows for the creation of confidence intervals that reflect the uncertainty as a product of the disclosure avoidance in the PPMF-based tabulations

2010 AMC Replicates



Registry of Open Data on AWS

aws

Open Data on AWS is now available on AWS Data Exchange

The Registry of Open Data are now discoverable on AWS Data Exchange alongside 3,000+ existing data products from category-leading data providers across industries. Explore the catalog to find commercial data sets. Learn more about AWS Data Exchange

Explore the catalog

Estimating Confidence Intervals for 2020 Census Statistics Using an Approximate Monte Carlo Simulation

age | approximate monte carlo | approximate monte carlo replicates | census | demographic and housing characteristics file | dhc | differential privacy | disclosure avoidance | ethnicity | group quarters | hispanic | household type | housing | housing units | latino | microdata | noisy measurements | population | race | redistricting | relation-to-householder | single year of age | voting age

Description

The 2010 Census Production Settings Demographic and Housing Characteristics (DHC) Approximate Monte Carlo (AMC) method seed Privacy Protected Microdata File (PPMF0) and PPMF replicates (PPMF1, PPMF2, ..., PPMF25) are a set of microdata files intended for use in estimating the magnitude of error(s) introduced by the 2020 Decennial Census Disclosure Avoidance System (DAS) into the Redistricting and DHC products. The PPMF0 was created by executing the 2020 DAS TopDown Algorithm (TDA) using the confidential 2010 Census Edited File (CEF) as the initial input; the replicates were then created by executing the 2020 DAS TDA repeatedly with the PPMF0 as its initial input. Inspired by analogy to the use of bootstrap methods in non-private contexts, U.S. Census Bureau (USCB) researchers explored whether simple calculations based on comparing each PPMFi to the PPMF0 could be used to reliably estimate the scale of errors introduced by the 2020 DAS, and generally found this approach worked well.

The PPMF0 and PPMFi files contained here are provided so that external researchers can estimate properties of DAS-introduced error without privileged access to internal USCB-curated data sets; further information on the estimation methodology can be found in Ashmead et. al 2024.

The 2010 DHC AMC seed PPMF0 and PPMF replicates have been cleared for public dissemination by the USCB Disclosure Review Board (CBDRB-FY24-DSEP-0002). The 2010 PPMF0 included in these files was produced using the same parameters and settings as were used to produce the 2010 Demonstration Data Product Suite (2023-04-03) PPMF, but represents an independent execution of the TopDown Algorithm. The PPMF0 and PPMF replicates contain all Person and Units attributes necessary to produce the Redistricting and DHC publications for both the United States and Puerto Rico, and include geographic detail down to the Census Block level. They do not include

Resources on AWS

Description
2010 Census Production Settings Demographic and Housing Characteristics Approximate Monte Carlo method seed Privacy Protected Microdata File and PPMF replicates

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::uscb-2020-product-releases/decennial/amc/2010/mdf/2010-dhc-mdf-replicates`

AWS Region
`us-west-2`

AWS CLI Access (No AWS account required)
`aws s3 ls --no-sign-request s3://uscb-2020-product-releases/decennial/amc/2010/mdf/2010-dhc-mdf-replicates/`

Description
Census Open Data S3 Inventory

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::uscb-opendata-inventory`

AWS Region
`us-west-2`

https://registry.opendata.aws/census-2010-amc-mdf-replicates/

United States® Census Bureau

Summary of Working Paper

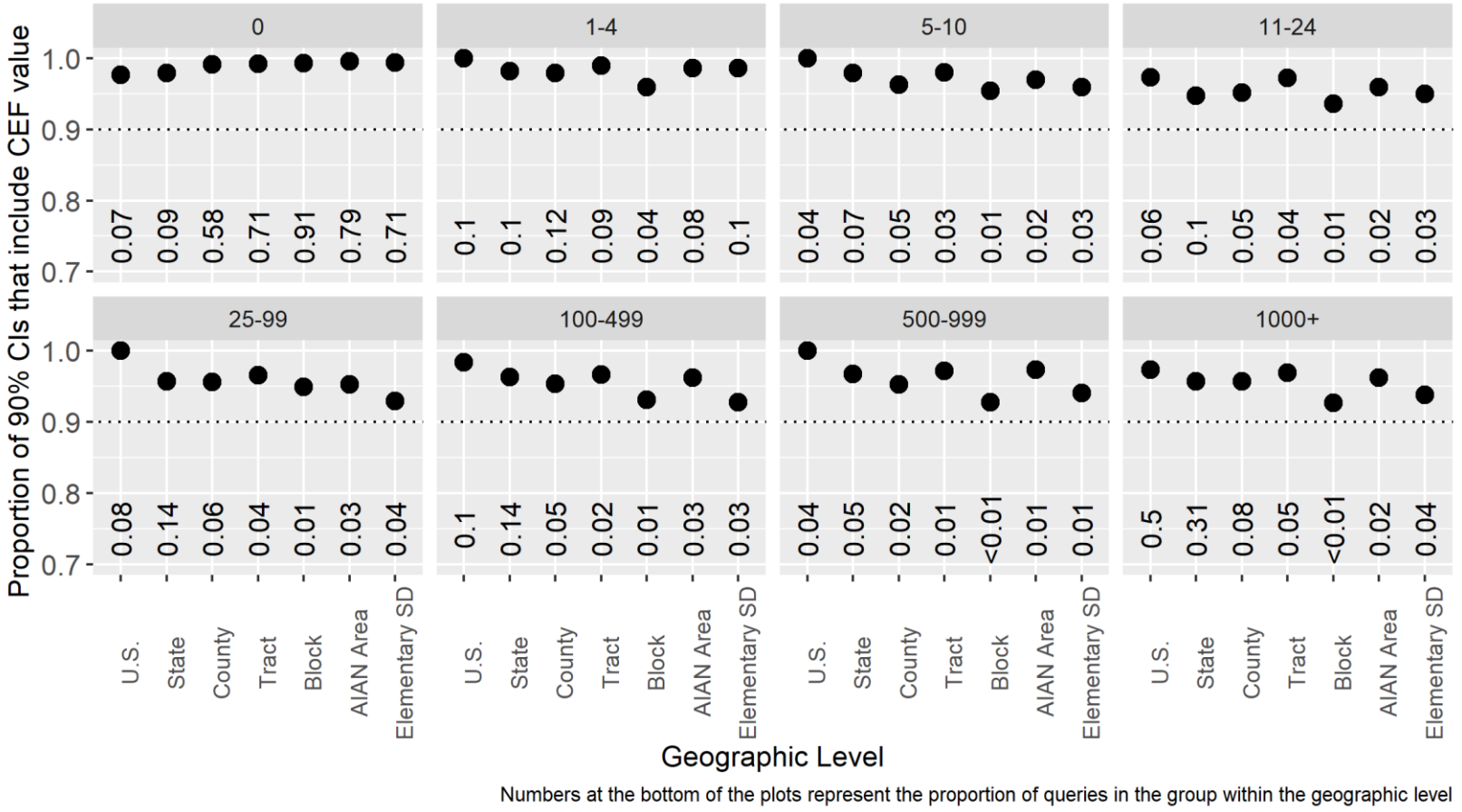
- Working Paper link
- Examined how well variance, bias, and mean-squared error are estimated for the AMC simulation method
- Proposed several possible confidence interval methods and tested their coverage and compared their widths
- We recommend users utilize the proposed conditionally bias-corrected t interval
- Overall, we found the method works very well in terms of being able to construct 90% confidence intervals with appropriate coverage
- Discuss situations when it worked less well and future directions

Redistricting Queries

- 301 total queries per geography
  - P1. Race [71 cells]
  - P2. Hispanic or Latino, and not Hispanic or Latino by Race [73 cells]
  - P3. Race for the Population 18 Years and Over [71 cells]
  - P4. Hispanic or Latino, and not Hispanic or Latino by Race for the Population 18 Years and Over [73 cells]
  - P5. Group Quarters Population by Major Group Quarters Type [10 cells]
  - H1. Occupancy Status (Housing) [3 cells]

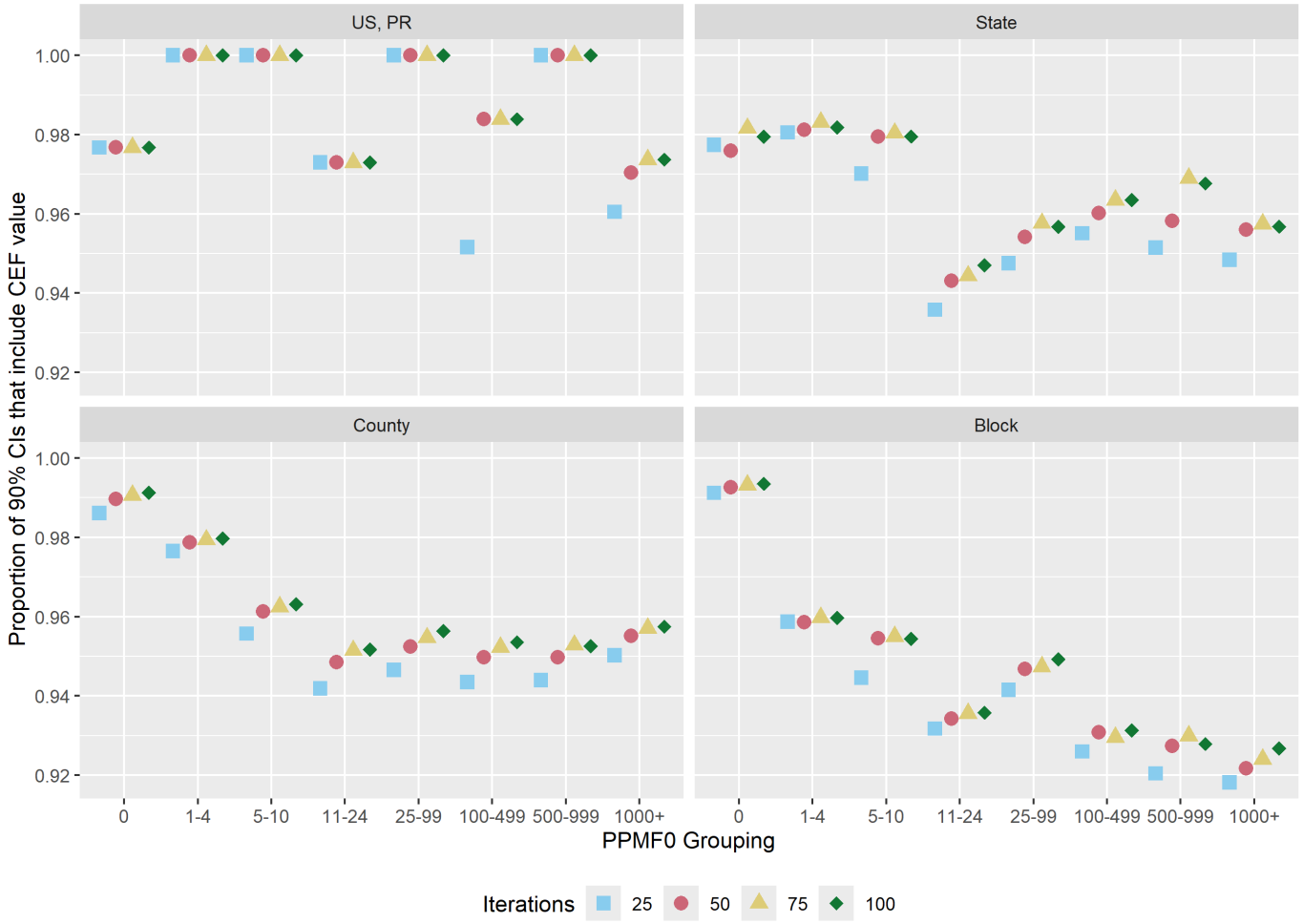
Confidence Interval Coverage for Redistricting Queries (2010)

These are nominal 90% confidence intervals using the conditional bias-correction rule and t-distribution critical value



The Effect of the Number of Iterations

- These are nominal 90% confidence intervals using the conditional bias-correction rule and t-distribution critical value
- We first examined using the results using 100 simulation iterations
- We then experimented with a smaller number and found that as few as 25 still had good performance



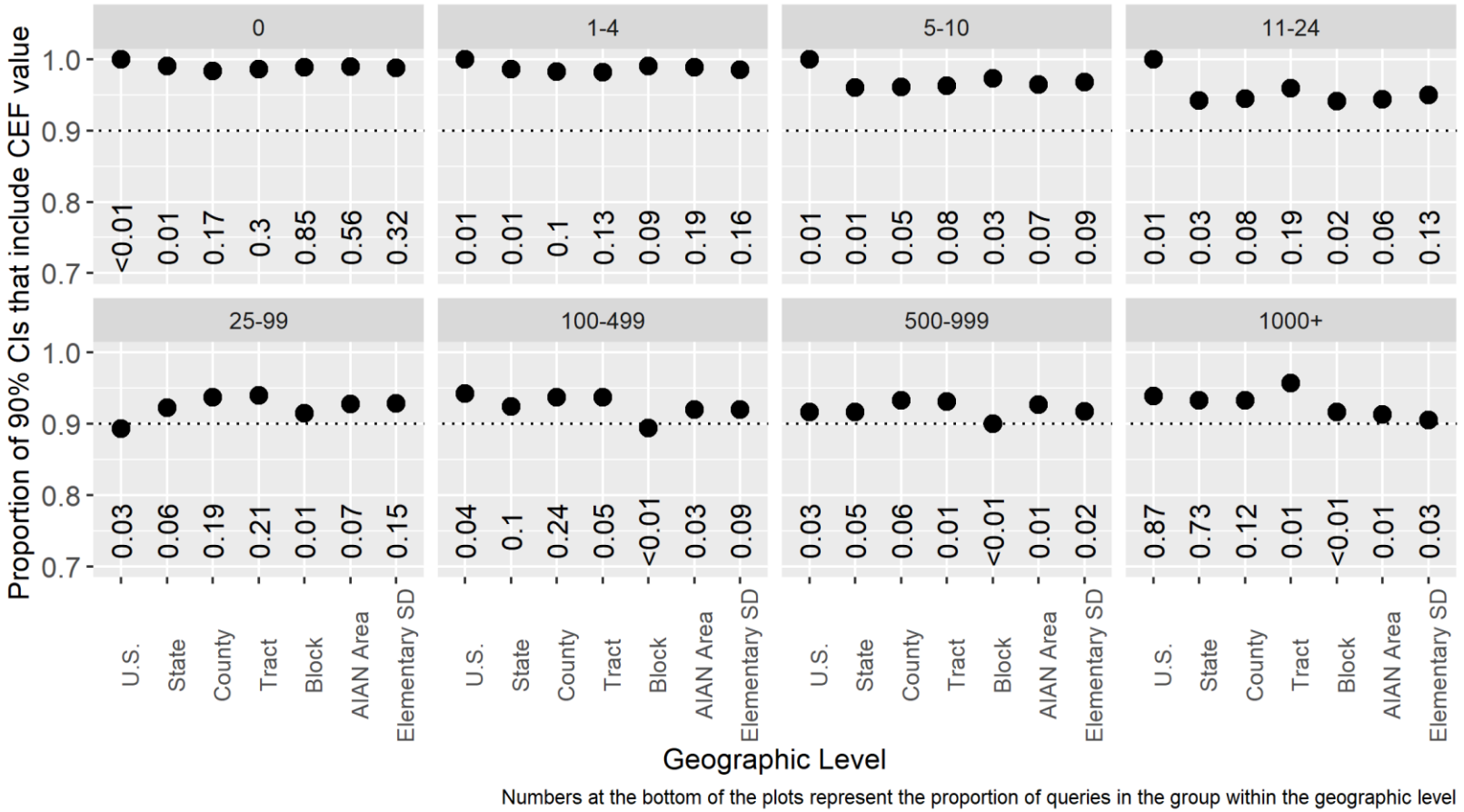
United States® Census Bureau

DHC (Person) Queries

- Subset of all DHC persons queries
- 405 total queries per geography
  - PCT8. Relationship by Age Under 18 Year [36 cells]
  - PCT9, PCT9A-I. Household Type by Relationship for the Population 65 Years and Over [16 cells x 10 iterations]
  - PCT12. Sex by Single-Year Age [209 cells]

Confidence Interval Coverage for DHC (Person) Queries (2010)

These are nominal 90% confidence intervals using the conditional bias-correction rule and t-distribution critical value



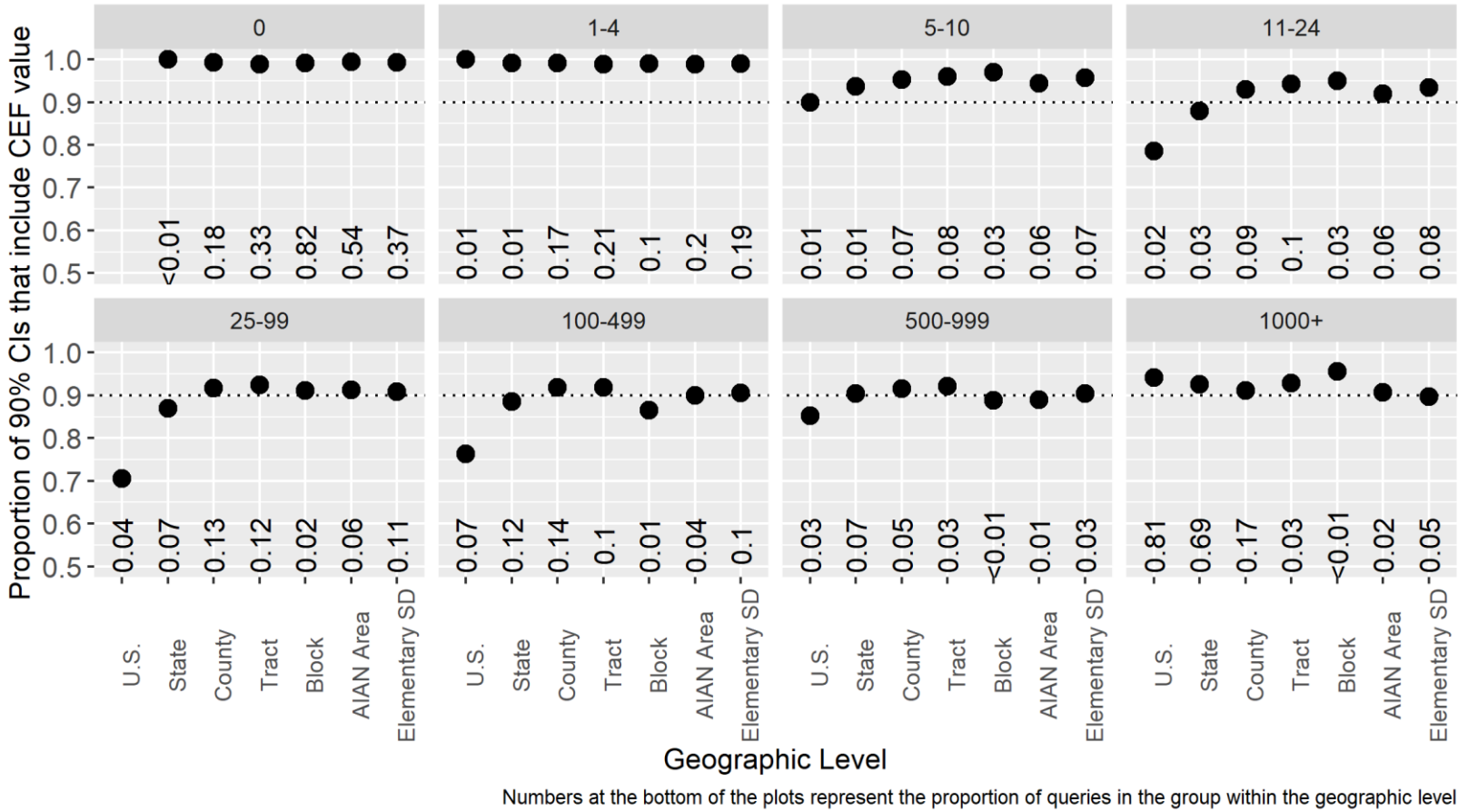
United States® Census Bureau

DHC (Unit) Queries

- Subset of all DHC persons queries
- 405 total queries per geography
  - H4. Tenure [4 cells]
  - P19. Households by Presence of People 65 Years and Over, Household Size, and Household Type [11 cells]
  - PCT7, PCT7A-I. Household Type by Household Size [16 cells x 10 iterations]
  - PCT10, PCT10A-I. Family Type by Presence and Age of Own Children [20 cells x 10 iterations]
  - PCT14, PCT14A-I. Presence of Multigenerational Households [3 cells x 10 iterations]

Confidence Interval Coverage for DHC (Unit) Queries (2010)

These are nominal 90% confidence intervals using the conditional bias-correction rule and t-distribution critical value



Next Steps

- Explore the utility of other applications of the AMC method
  - Ratios (e.g. persons per household queries)
  - Regression models
- Refine the decision rule for the conditional bias-corrected confidence interval
- Goal: Release AMC iterations for the 2020 PPMF and tools for using them

Questions?

Workshop on Using 2020 Census Data

Session V: Concluding Discussion

Cass Dorius
Research and Methodology

United States® Census Bureau

What additional resources would you like to see?

United States® Census Bureau