Setting the Parameters of the TopDown Algorithm

Michael Hawes

Senior Advisor for Data Access and Privacy

Committee on National Statistics Workshop

June 21, 2022

United States® Census Bureau

Keeping the Public's Trust: Title 13

"To stimulate public cooperation necessary for an accurate census…Congress has provided assurances that information furnished by individuals is to be treated as confidential. **Title 13 U.S.C. §§ 8(b) and 9(a)** *explicitly provide for nondisclosure of certain census data, and* **no discretion is provided to the Census Bureau on whether or not to disclose such data***…"* (U.S. Supreme Court, Baldrige v. Shapiro, 1982)

To safeguard the public's confidential census responses, the Census Bureau has long employed a variety of statistical techniques to mitigate disclosure risk in our published data products.
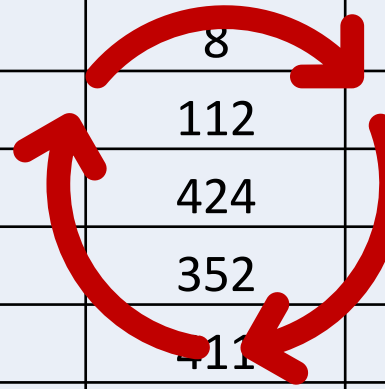
Disclosure Avoidance for Past Censuses

1970-1980 Censuses

	528	
		794
	581	
137	941	189
931		
	250	
		590

SUPPRESSION

1990-2010 Censuses

668	178	779
91	8	159
809	112	811
518	424	955
989	352	765
237	411	686
77	820	590

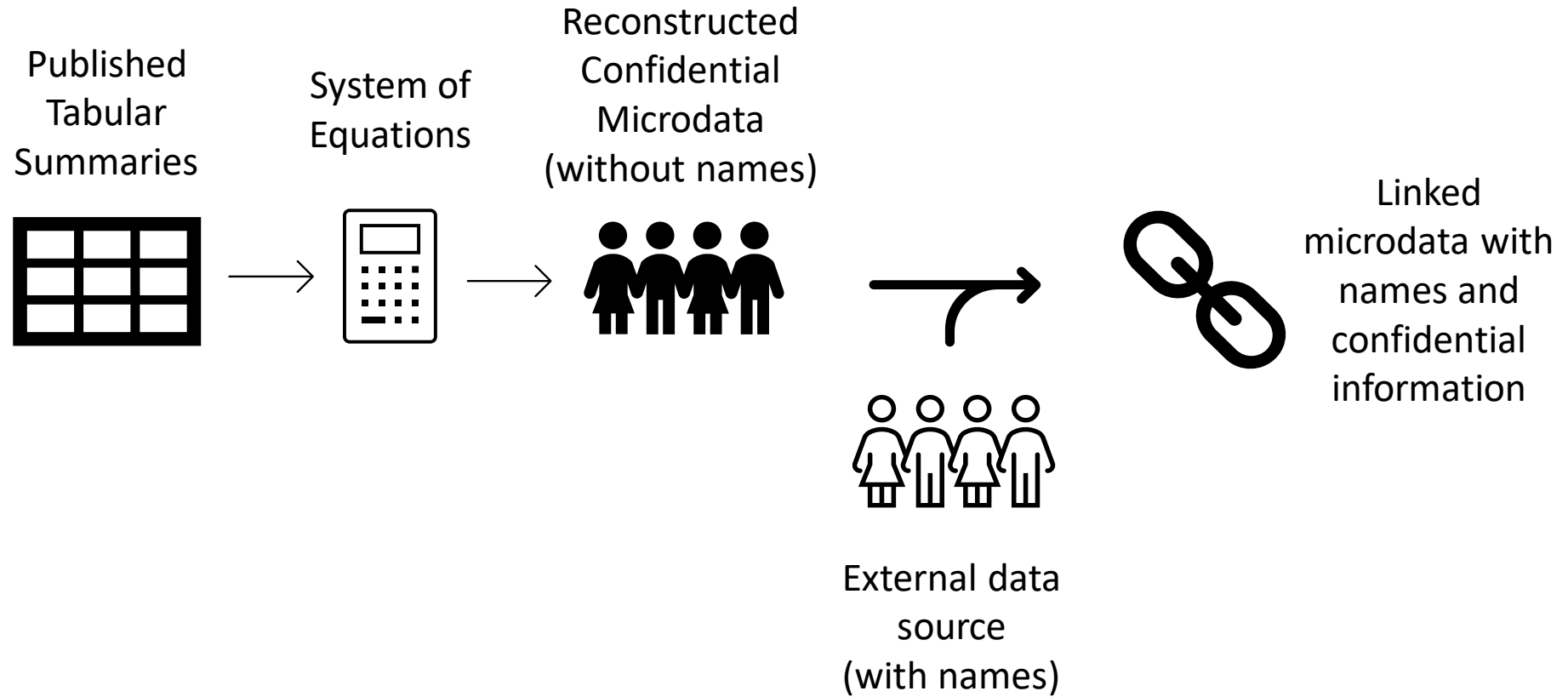## SWAPPING

The Ever-rising Risk of Disclosure

- Any data release carries some risk of disclosure.

- Improvements in computing power and the explosion of third-party data mean that disclosure risk has increased significantly.

- Protecting confidentiality means adapting and responding to these increasing threats

Simulated Re-identification Attack

Published Tabular Summaries → System of Equations → Reconstructed Confidential Microdata (without names)

External data source (with names)

Linked microdata with names and confidential information

What census data can an attacker use?

Anything published from the 2010 Census!

Our simulated attack used only a small subset:

P001 (Total Population by Block)
P006 (Total Races Tallied by Block)
P007 (Hispanic or Latino Origin by Race by Block)
P009 (Hispanic or Latino, and Not Hispanic or Latino by Race by Block)
P011 (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over by Block)
P012 (Sex by Age by Block)
P012A-I (Sex by Age by Block, iterated by Race)
P014 (Sex by Single-year-of-age for the Population under 20 Years by Block)
PCT012A-N (Sex by Single-year-of-age by Tract, iterated by Race)

United States®
Census
Bureau

Agreement rates for reconstructed microdata

Percentage of reconstructed records that exactly agree with the CEF on location, sex, age (exact/binned), race, and ethnicity

Agreement Rates	Exact Age	Exact and Binned Age
Published 2010 Tables (swapping)	46.5	91.8
High Swapping Experiment	26.5	52.1
DAS Run (ρ=3.325) (DDP DHC-P*)	15.7	33.1
DAS Run (ρ=6.65)	17.1	36.4

*DDP DHC-P: The DAS Run with ρ=3.325 is the run used to generate the 2010 DHC-P Demonstration Data Product 2022-03-16.

For additional explanation and results see 2022-03-17 CSAC Presentation at https://www2.census.gov/about/partners/cac/sac/meetings/2022-03/presentation-reconstruction-and-reidentification-of-the-dhc.pdf

CBDRB-FY22-DSEP-004

Re-identification of Population Uniques for Non-Modal Races

Re-identification statistics for "population uniques" of the linking pseudo-identifiers (those who are unique within their block on sex, and either exact age [SAB] or binned age [SAbB]) for individuals of the blocks' Non-Modal Races

Non-Modal Race	Putative Rate		Confirmation Rate		Precision Rate	
	SAB	**SAbB**	**SAB**	**SAbB**	**SAB**	**SAbB**
Published 2010 Tables (swapping) to Commercial	24.0	13.8	14.3	12.3	59.4	89.2
High Swapping Experiment to Commercial	20.6	11.5	5.0	3.5	24.4	30.6
DAS Run (ρ=3.325) to Commercial (DDP)	11.4	5.3	2.4	1.2	20.8	23.2
DAS Run (ρ=6.65) to Commercial	12.3	5.7	2.6	1.4	21.2	24.0
Published 2010 Tables (swapping) to CEF	90.6	86.2	60.4	70.2	66.7	81.4
High Swapping Experiment to CEF	71.6	65.5	20.0	21.9	27.9	33.4
DAS Run (ρ=3.325) to CEF (DDP)	34.7	25.9	7.8	6.2	22.3	24.0
DAS Run (ρ=6.65) to CEF	37.6	28.3	8.6	7.1	23.0	25.1

For additional explanation and results see 2022-03-17 CSAC Presentation at https://www2.census.gov/about/partners/cac/sac/meetings/2022-03/presentation-reconstruction-and-reidentification-of-the-dhc.pdf

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.

The 2020 Census Disclosure Avoidance System

TopDown Algorithm (TDA)

Produces privacy-protected microdata (Microdata Detail File) that can be ingested by Decennial tabulation systems

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

SafeTab PHSafe

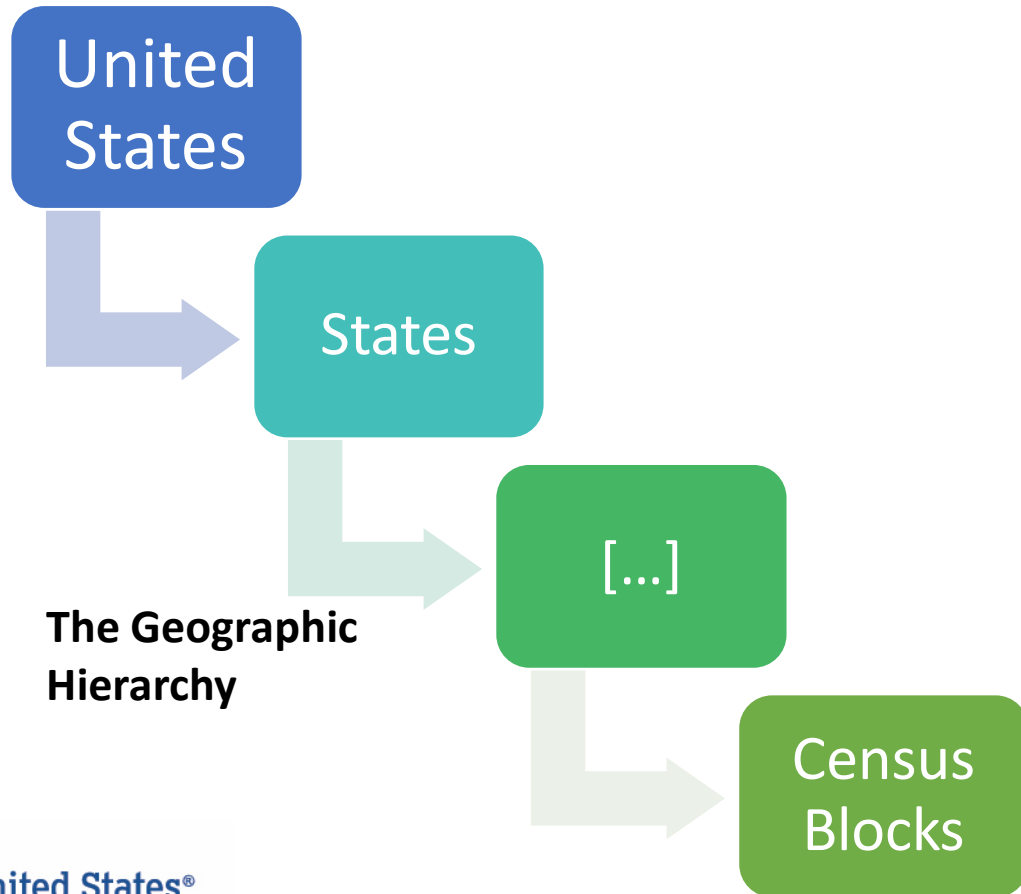Produce privacy-protected tabulations directly

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

The TopDown Algorithm

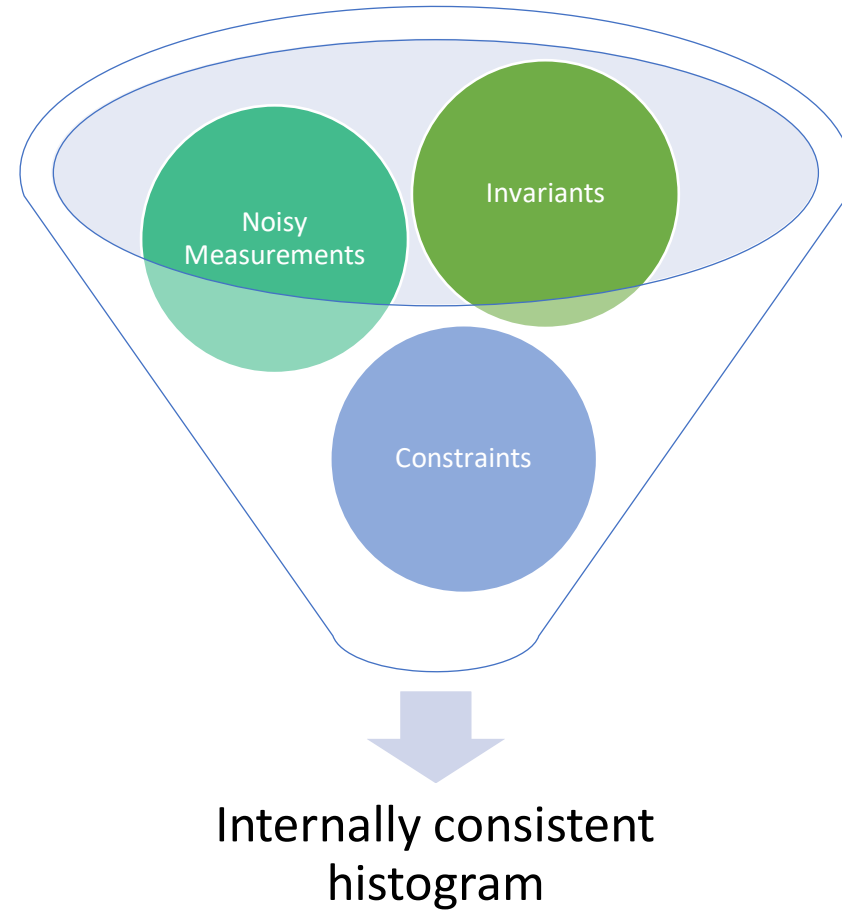| Input Microdata (CEF) & Tabulation Geographic Reference File (Tab GRF-C)	Conversion to Histogram	Noisy Measurements	Post-processing	Conversion to Microdata (MDF)

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
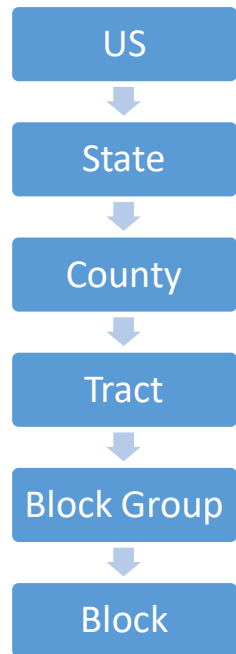
The TopDown Algorithm

United States

States

[…]

Census Blocks

The Geographic Hierarchy

At each geographic level:

Noisy Measurements

Invariants

Constraints

Internally consistent histogram

The Geographic Hierarchy ("Spine")

Standard Tabulation Hierarchy

US → State → County → Tract → Block Group → Block

TDA's AIAN Spine for Redistricting

United States

Non-AIANNH portions of 51 State/State Equivalent Areas → Non-AIANNH portions of Counties → Non-AIANNH portions of Tracts and Block Groups → Non-AIANNH Census Blocks

AIANNH portions of 37 States → AIANNH portions of Counties → AIANNH portions of Tracts and Block Groups → AIANNH Census Blocks
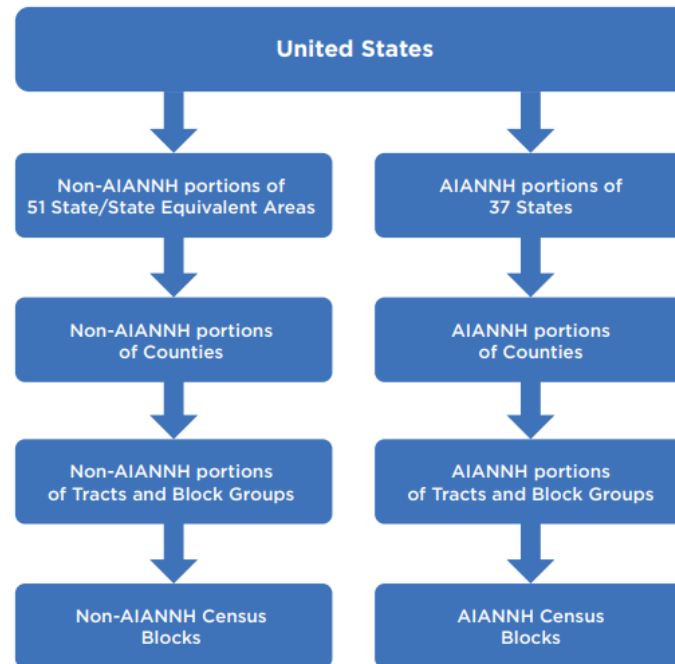
Geographic Optimization
(for Redistricting Data)

Block Group → "Optimized" Block Group
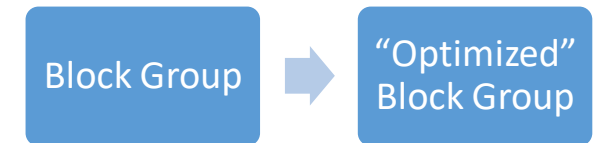
Reconfigured TDA's definition of block groups to bring MCDs, Places, and individual AIAN areas closer to the spine. Optimized block groups are used inside TDA. Tabulation block groups from the standard hierarchy are used for all published data tables.

Population Estimates Primitive Geographies

```
US
 ↓
State
 ↓
County
 ↓
Population Estimates
Primitive Geography
 ↓
Tract Subset Group
 ↓
Tract Subset
 ↓
Optimized Block
Group
 ↓
Block
```
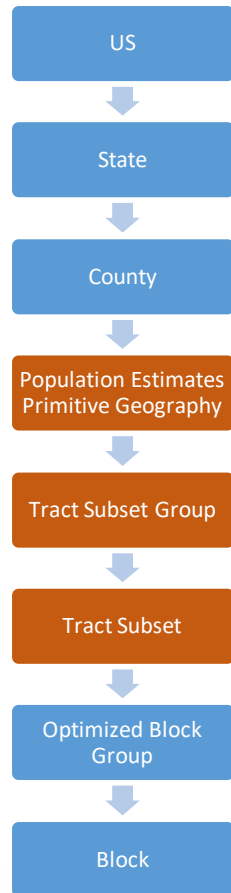
For the DHC, the TDA Geographic Hierarchy was further modified to include "Population Estimates Primitive Geographies" onto the spine.

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.

TDA Query Structure

TDA only takes noisy measurements for defined queries (tabulations) at particular geographic levels. Adjusting the queries asked and/or the share of PLB assigned to those queries determine the resulting amount of noise injected into the DHC statistics derived from those queries.

DHC-P PLB allocations by geographic level and query as reflected in the 2022-03-16 Demonstration Data Product

Global *rho*	3.325
Global *epsilon*	20.01
delta	10^{-10}

	rho Allocation by Geographic Level
US	1.95%
State	27.07%
County	8.42%
Population Estimates Primitive Geography[†]	12.93%
Tract Subset Group[‡]	12.93%
Tract Subset[‡]	23.46%
Optimized Block Group[◊]	12.93%
Block	0.30%

	Per Query *rho* Allocation by Geographic Level							
Query	US	State	County	Population Estimates Primitive Geography[†]	Tract Subset Group[‡]	Tract Subset[‡]	Optimized Block Group[◊]	Block
AGE (3 bins) * HHGQ (4 Levels) (12 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
AGE (3 bins) * SEX (6 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
AGE (13 bins) * SEX (26 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
HISPANIC * SEX (4 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
SEX * HHGQ (4 levels) (8 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
HISPANIC * SEX * AGE (13 bins) * HHGQ (8 levels) * CENRACE (26,208 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
HHGQ (8 levels) * AGE (23 bins) * HISPANIC * CENRACE * SEX (46,368 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
RELGQ * AGE (23 bins) * HISPANIC * CENRACE * SEX (243,432 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
RELGQ * SEX * AGE (116 bins) * HISPANIC * CENRACE (1,227,744 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%

Query Structure for the DHC-H File

Query structure for the DHC-H tabulations of the
2022-03-16 Demonstration Data File

| |
|---|
| SEX * HISPANIC * HH_TENURE * RACE * FAMILY_NONFAMILY_SIZE (728 cells) |
| SEX * HISPANIC * HH_TENURE * RACE * HH_AGE * FAMILY_NONFAMILY_SIZE (6,552 cells) |
| SEX * HH_AGE * HISPANIC * RACE * ELDERLY * HH_TENURE * HH_TYPE (1,052,352 cells) |
| TENVACGQ (35 cells) |
| MULTG * HISPANIC * HH_TENURE (8 cells) |
| PARTNER_TYPE_OWN_CHILD_STATUS * SEX * HH_TENURE (24 cells) |
| COUPLED_HH_TYPE * HISPANIC * HH_TENURE (20 cells) |
| SEX * HISPANIC * HH_TENURE * RACE * DETAILED_COUPLETYPE_MULTG_OWNCHILD_SIZE (5,544 cells) |
| SEX * HISPANIC * HH_TENURE * RACE * HH_AGE * DETAILED_COUPLETYPE_MULTG_OWNCHILD_SIZE (49,896 cells) |

DHC-H PLB Allocation Error in the 2022-03-16 Demonstration Data Product

An error in the DHC-H configuration file introduced as part of our rapid-cycle experimental tuning runs inadvertently reversed the order of the PLB allocations by geographic level for the 2022-03-16 Demonstration Data Product DHC-H file.

DHC-H DDP Intended PLB Allocations

	rho Allocation by Geographic Level
US	6.84%
State	28.39%
County	11.10%
Population Estimates Primitive Geography*	11.10%
Tract Subset Group‡	11.10%
Tract Subset‡	20.13%
Optimized Block Group◊	11.10%
Block	0.26%
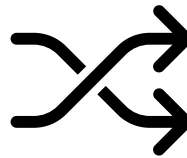
DHC-H DDP Actual PLB Allocations

	rho Allocation by Geographic Level
US	0.26%
State	11.10%
County	20.13%
Population Estimates Primitive Geography*	11.10%
Tract Subset Group‡	11.10%
Tract Subset‡	11.10%
Optimized Block Group◊	28.39%
Block	6.84%

Impact on the Demonstration Data Product

In general:

- US and State-level tabulations in the demonstration product were <u>significantly less accurate</u> than intended.

- County-level tabulations in the demonstration product were <u>slightly more accurate</u> than intended.

- Tabulations for Incorporated Places were <u>comparable</u>

- Tract-level tabulations in the demonstration product were <u>slightly less accurate</u> than intended.

Implications

- The disclosure risk implications of the unintended PLB expended on block-level tabulations require close scrutiny.

- Feedback on the published 2022-03-16 Demonstration Data Product is still enormously valuable in informing the setting of use-case-based accuracy targets for the second DHC Demonstration Data Product.

- We do not plan to re-issue a corrected version of the 2022-03-16 Demonstration Data Product. We will release the Detailed Summary Metrics for the "as intended" run.

- The forthcoming second DHC Demonstration Data Product will reflect improvements based on extensive internal analysis and external use-case-derived accuracy targets

- For tabulations above the block-level that were released with greater accuracy than intended in the 2202-03-16 Demonstration Data Product, the Census Bureau commits to maintaining (or improving) that level of accuracy in the second DHC Demonstration Data Product and in the 2020 Census DHC production run, consistent with ongoing confidentiality assessments.

Examples (State-level Tabulations):

DHC-H DDP (as released)

DHC Use Case Table 2.a: Tenure by Age of Householder for states - MAE, RMSE, MAPE, CV, MALPE, and outliers

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of geographies where the absolute percent difference exceeds 5%
Owner occupied							
Householder 15 to 24 years	51	488.37	611.23	5.90	3.58	5.90	16
Householder 25 to 34 years	51	127.06	214.45	0.14	0.14	0.04	-
Householder 35 to 54 years	51	566.61	613.50	0.20	0.10	(0.20)	-
Householder 55 to 64 years	51	173.49	206.36	0.12	0.06	0.11	-
Householder 65 years and over	51	198.55	371.62	0.10	0.09	(0.01)	-
Renter occupied							
Householder 15 to 24 years	51	213.78	402.84	0.39	0.45	0.26	-
Householder 25 to 34 years	51	268.06	311.16	0.34	0.15	(0.33)	-
Householder 35 to 54 years	51	490.25	560.41	0.35	0.19	(0.35)	-
Householder 55 to 64 years	51	381.82	408.62	1.05	0.43	1.04	-
Householder 65 years and over	51	258.08	309.25	0.54	0.27	0.51	-

DHC-H DDP (as intended)

DHC Use Case Table 2.a: Tenure by Age of Householder for states - MAE, RMSE, MAPE, CV, MALPE, and outliers

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of geographies where the absolute percent difference exceeds 5%
Owner occupied							
Householder 15 to 24 years	51	125.82	174.33	2.18	1.02	2.15	6
Householder 25 to 34 years	51	71.82	106.58	0.10	0.07	0.01	-
Householder 35 to 54 years	51	157.65	177.54	0.08	0.03	(0.07)	-
Householder 55 to 64 years	51	103.08	129.03	0.07	0.04	0.05	-
Householder 65 years and over	51	128.33	226.09	0.07	0.06	-	-
Renter occupied							
Householder 15 to 24 years	51	118.88	188.10	0.26	0.21	0.08	-
Householder 25 to 34 years	51	131.76	175.33	0.15	0.09	(0.13)	-
Householder 35 to 54 years	51	128.08	148.76	0.12	0.05	(0.11)	-
Householder 55 to 64 years	51	148.25	182.04	0.45	0.19	0.42	-
Householder 65 years and over	51	127.57	206.38	0.24	0.18	0.18	-

Examples (County-level Tabulations):

DHC-H DDP (as released)

DHC Use Case Table 13.b: Coupled Household Type by Hispanic or Latino Origin of Householder for counties- MAE, RMSE, MAPE, CV, MALPE, and outliers

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of geographies where the absolute percent difference exceeds 5%
Opposite-sex married couple household	3,143	6.01	7.72	0.21	0.04	(0.03)	5
Householder who is Hispanic or Latino	3,143	4.46	6.51	11.40	0.30	3.24	1,115
Householder who is not Hispanic or Latino	3,143	5.60	7.58	0.21	0.05	(0.03)	5
Same-sex married couple household	3,143	3.56	4.71	20.75	4.24	8.15	1,897
Householder who is Hispanic or Latino	3,143	1.60	2.60	67.04	16.21	35.21	1,910
Householder who is not Hispanic or Latino	3,143	3.06	3.97	20.51	4.17	6.82	1,884
Opposite-sex unmarried partner household	3,143	5.09	6.55	2.24	0.30	0.32	271
Householder who is Hispanic or Latino	3,143	3.12	4.22	34.26	1.15	17.27	1,833
Householder who is not Hispanic or Latino	3,143	4.29	5.53	2.27	0.31	0.19	258
Same-sex unmarried partner households	3,143	3.45	4.59	31.31	2.61	16.86	1,932
Householder who is Hispanic or Latino	3,143	1.52	2.50	67.88	12.83	35.74	1,843
Householder who is not Hispanic or Latino	3,143	2.99	3.89	30.18	2.49	13.88	1,874

DHC-H DDP (as intended)

DHC Use Case Table 13.b: Coupled Household Type by Hispanic or Latino Origin of Householder for counties- MAE, RMSE, MAPE, CV, MALPE, and outliers

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of geographies where the absolute percent difference exceeds 5%
Opposite-sex married couple household	3,143	6.99	8.96	0.27	0.05	(0.01)	9
Householder who is Hispanic or Latino	3,143	4.99	7.02	12.78	0.33	4.13	1,222
Householder who is not Hispanic or Latino	3,143	6.24	8.14	0.27	0.05	(0.03)	11
Same-sex married couple household	3,143	4.36	5.70	24.05	5.13	8.73	2,056
Householder who is Hispanic or Latino	3,143	1.82	3.11	69.28	19.42	34.72	1,906
Householder who is not Hispanic or Latino	3,143	3.75	4.78	24.50	5.03	8.05	2,005
Opposite-sex unmarried partner household	3,143	6.12	7.88	2.64	0.36	0.41	359
Householder who is Hispanic or Latino	3,143	3.69	4.96	39.38	1.35	19.13	1,942
Householder who is not Hispanic or Latino	3,143	5.18	6.61	2.68	0.37	0.19	356
Same-sex unmarried partner households	3,143	4.28	5.64	37.14	3.21	20.20	2,079
Householder who is Hispanic or Latino	3,143	1.78	2.99	70.81	15.35	38.53	1,856
Householder who is not Hispanic or Latino	3,143	3.71	4.79	35.99	3.06	17.08	2,014

Examples (Place-level Tabulations):

DHC-H DDP (as released)

DHC Use Case Table 12.c: Presence of Own Children Under 6 for incorporated place size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of geographies where the absolute percent difference exceeds 5%
All incorporated places	19,540	12.79	33.57	22.65	6.77	8.80	12,480
Incorporated places with total population less than 500	6,168	3.15	4.15	51.21	39.10	29.29	5,320
Incorporated places with total population 500 to 999	3,066	5.49	6.89	18.47	20.04	1.32	2,486
Incorporated places with total population 1,000 to 4,999	5,672	9.15	11.92	9.43	10.39	(1.55)	3,627
Incorporated places with total population 5,000 to 9,999	1,664	15.91	20.69	4.66	5.78	(1.58)	631
Incorporated places with total population 10,000 to 49,999	2,265	29.74	39.85	3.38	3.58	(0.79)	400
Incorporated places with total population 50,000 to 99,999	432	60.33	81.87	1.71	2.29	0.29	13
Incorporated places with total population of 100,000 or more	273	153.38	224.45	1.34	1.46	0.90	3

DHC-H DDP (as intended)

DHC Use Case Table 12.c: Presence of Own Children Under 6 for incorporated place size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of geographies where the absolute percent difference exceeds 5%
All incorporated places	19,540	12.56	30.99	22.09	6.25	8.24	12,514
Incorporated places with total population less than 500	6,168	3.11	4.12	49.70	38.83	26.86	5,302
Incorporated places with total population 500 to 999	3,066	5.37	6.83	18.03	19.87	2.20	2,455
Incorporated places with total population 1,000 to 4,999	5,672	9.14	11.90	9.38	10.38	(1.27)	3,648
Incorporated places with total population 5,000 to 9,999	1,664	16.24	21.07	4.71	5.89	(1.44)	654
Incorporated places with total population 10,000 to 49,999	2,265	30.26	40.25	3.38	3.62	(0.86)	438
Incorporated places with total population 50,000 to 99,999	432	55.26	74.61	1.59	2.09	0.03	15
Incorporated places with total population of 100,000 or more	273	140.63	199.78	1.26	1.30	0.71	2

Examples (Tract-level Tabulations):

DHC-H DDP (as released)

DHC Use Case Table 1.d: Tenure for tracts - MAE, RMSE, and outliers				
	Count of Units (N)	MAE	RMSE	Count of geographies where the absolute percent difference exceeds 5%
Owned with a mortgage	73,057	2.31	2.98	1,338
Owned free and clear	73,057	2.29	2.96	3,915
Renter-occupied	73,057	2.30	2.97	2,090

DHC-H DDP (as intended)

DHC Use Case Table 1.d: Tenure for tracts - MAE, RMSE, and outliers				
	Count of Units (N)	MAE	RMSE	Count of geographies where the absolute percent difference exceeds 5%
Owned with a mortgage	73,057	1.91	2.48	1,147
Owned free and clear	73,057	1.89	2.47	2,992
Renter-occupied	73,057	1.91	2.49	1,599

United States® Census Bureau

CBDRB-FY22-DSEP-003

What we hope to get out of this workshop

- Releasing demonstration data assists our internal subject-matter experts in setting accuracy targets and evaluating different parameter settings of the TDA for the next demonstration product and for the final production run of the 2020 Census DHC.

- In response to the feedback we have already received and that we will receive at this workshop, we may adjust:
  - TDA's geographic hierarchy
  - TDA's noisy measurement query structure
  - Allocation of PLB by query and/or geographic level
  - Overall PLB allocation to the DHC-P and/or DHC-H

Thank You