What should librarians know about differential privacy and the 2020 Census?

David Van Riper

vanriper@umn.edu

@dcvanriper

FDLP Academy

June 9, 2020

Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau

Fri Aug 17 2018

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY

Protecting the Confidentiality of America's Statistics: Ensuring Confidentiality and Fitness-for-Use

Tue Sep 04 2018

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY

Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census

Fri Feb 15 2019

WRITTEN BY: DR. RON JARMIN, DEPUTY DIRECTOR AND COO

Outline

- New disclosure avoidance algorithm
- Fewer invariants
- Less publicly available data
- Less consistency among data products

NEW DISCLOSURE AVOIDANCE ALGORITHM

- Suppression and swaps
- Differential privacy

- Suppression and swaps
  - Top coding for income or household size
  - Table suppression
  - Swapping
    - Identify unique HHs in a geographic area and swap with similar HHs in a different geography
  - **Absolute**
- Differential privacy

- Suppression and swaps
  - Top coding for income or household size
  - Table suppression
  - Swapping
    - Identify unique HHs in a geographic area and swap with similar HHs in a different geography
  - **Absolute**

- Differential privacy
  - Inject noise into statistics
  - Magnitude of noise depends on policy decisions
  - **Relative**

HOW IS DIFFERENTIAL PRIVACY IMPLEMENTED?

“True” microdata

	Sex	School
	Male	Never
	Male	Never
	Male	Never
x12	Male	Attending
	Male	Attending
	⋮	
	Male	Attending
x33	Male	Past
	⋮	
	Male	Past

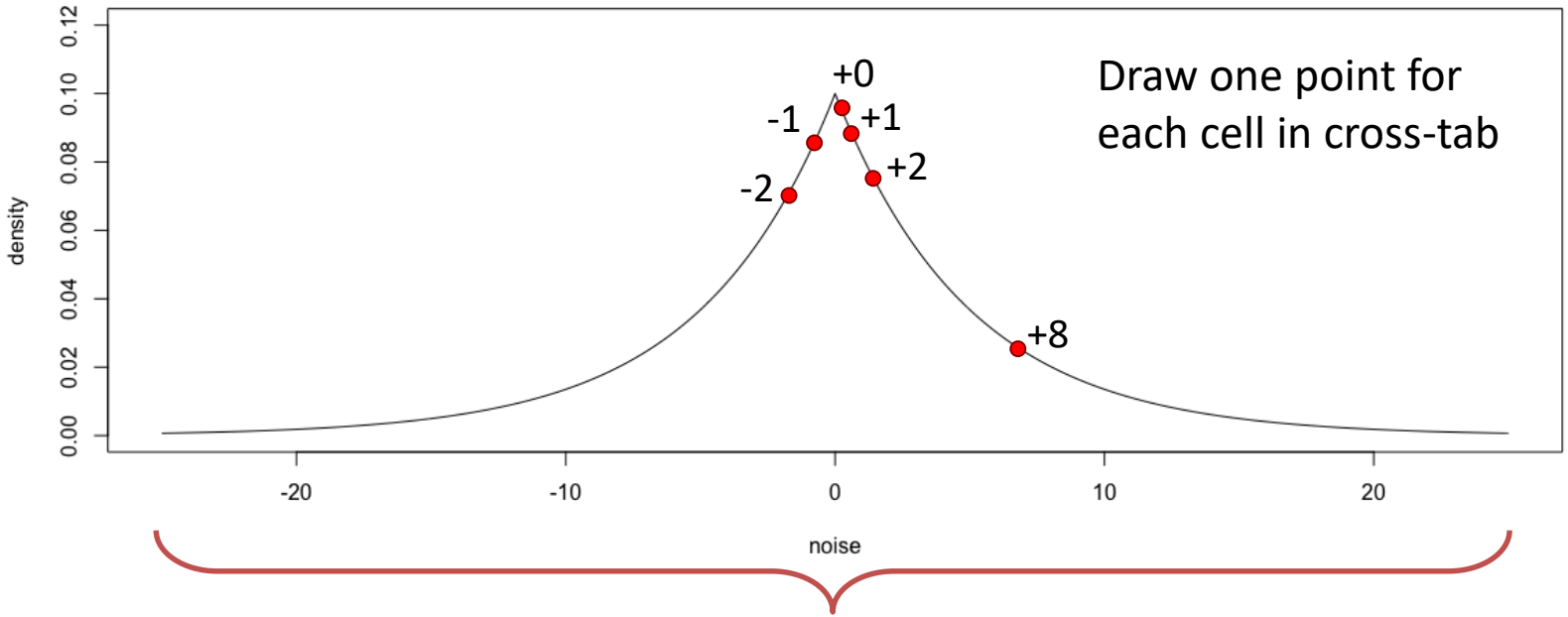
	Sex	School
x4	Female	Never
	⋮	
	Female	Never
x17	Female	Attending
	⋮	
	Female	Attending
x31	Female	Past
	⋮	
	Female	Past

Construct cross-tabs from “true” data

	School Attendance		
	Never	Attending	Past
Male	3	12	33
Female	4	17	31

Population = 100

Draw noise from Laplace distribution



Add noise to cross-tab

	School Attendance		
	Never	Attending	Past
Male	3 – 1 = **2**	12 + 0 = **12**	33 + 1 = **34**
Female	4 + 8 = **12**	17 + 2 = **19**	31 – 2 = **29**

Sum = 108

POLICY DECISIONS

Policy decisions

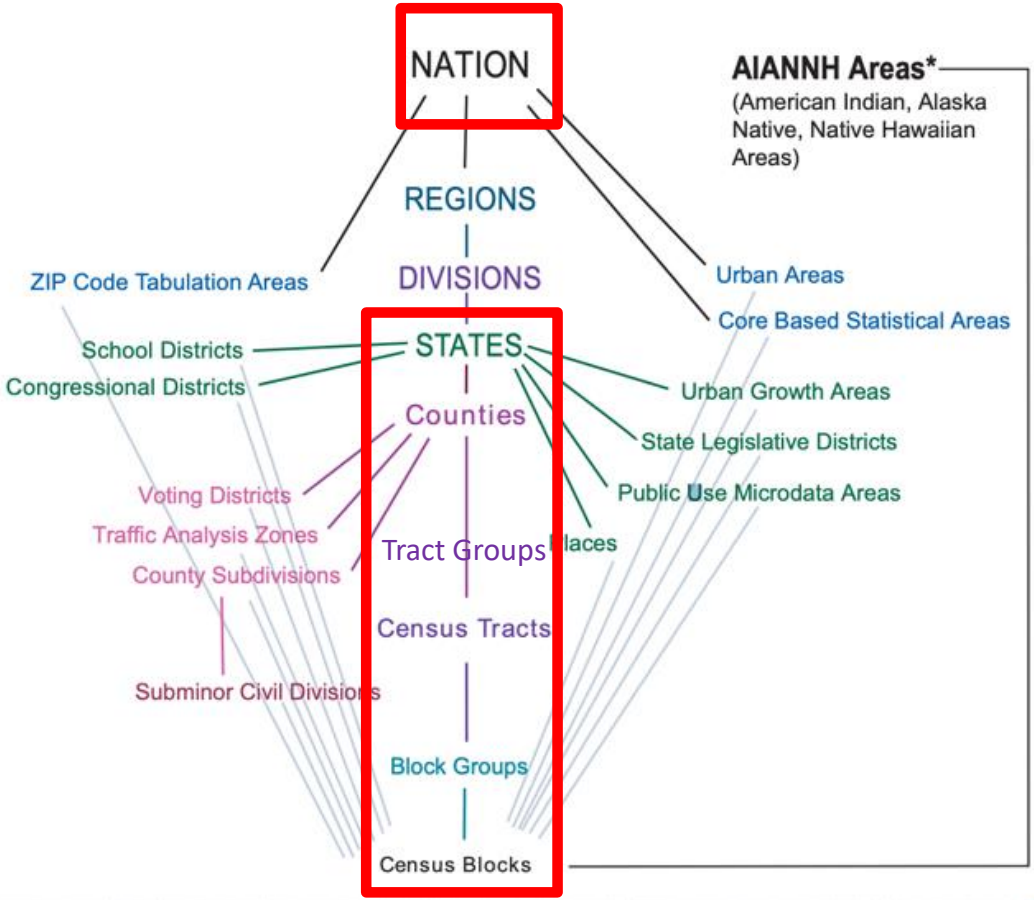
- Global privacy loss budget (**ε**)
- Fractional allocations
- Invariants and constraints

Policy decisions

- Global privacy loss budget ($\boldsymbol{\varepsilon}$)
- Fractional allocations
- Invariants and constraints

Fractional allocations

- Geographic levels
- Queries


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

- Detailed person
  - Age * Sex * Hispanic * Race * HHGQ * Citizen
- Voting age * Hispanic * Race * Citizen
- Detailed housing
- Hispanic * Race * Size of HH * HH type

Invariants and Constraints

- Invariants are counts not subject to noise injection

2010 Demonstration Data Product	
State – total population	
Census block – total housing units	
Census block – group quarters count	
Census block – group quarters type count	

2010 Demonstration Data Product	2010 Decennial
State – total population	Census block – total population
Census block – total housing units	Census block – total housing units
Census block – group quarters count	Census block – occupied housing units
Census block – group quarters type count	Census block – voting age population
	Census block – group quarters count
	Census block – group quarters type count

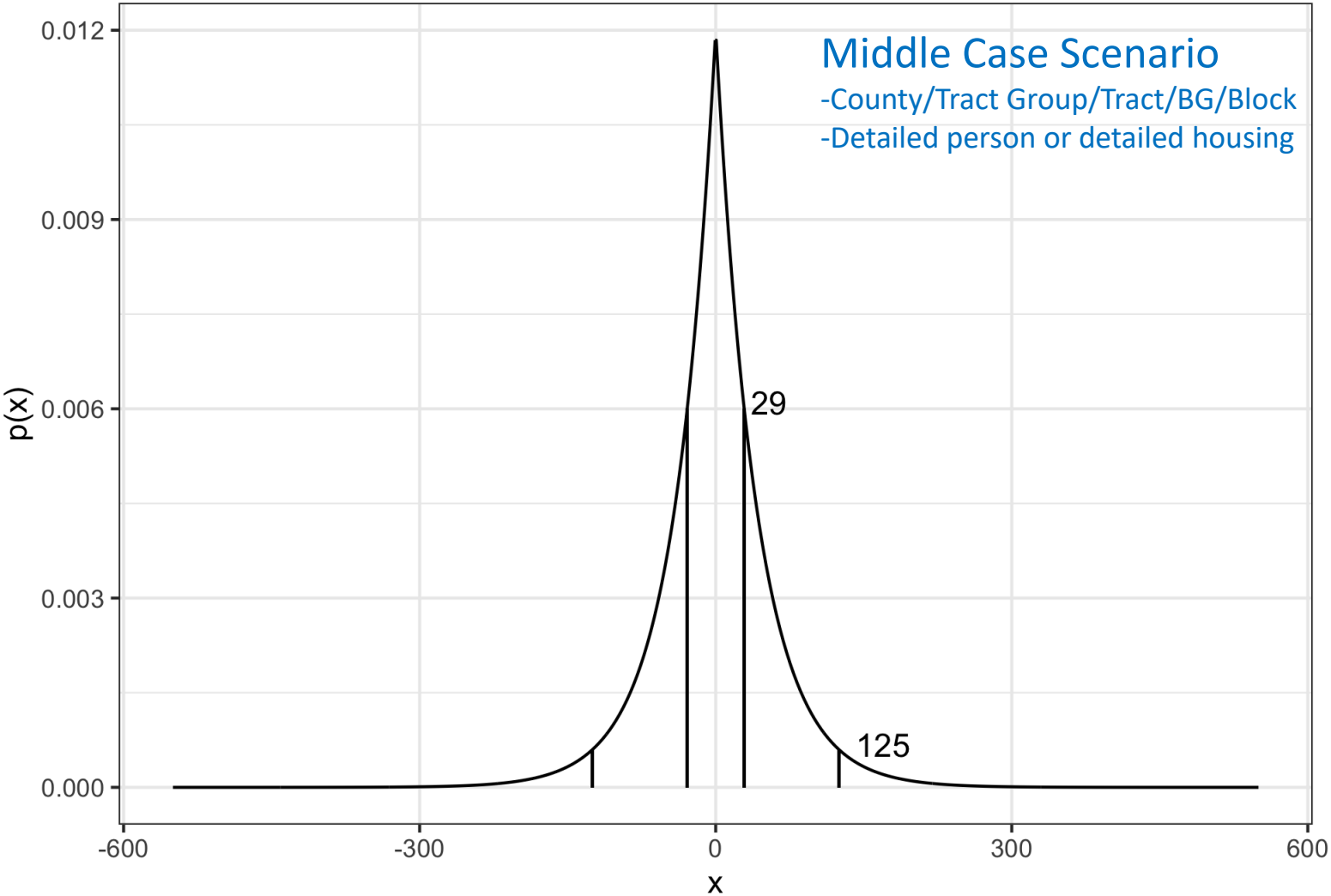
Invariants and Constraints

- Invariants are counts not subject to noise injection
- Constraints

Invariants and Constraints

- Invariants are counts not subject to noise injection
- Constraints
  - Non-negativity
  - Consistency

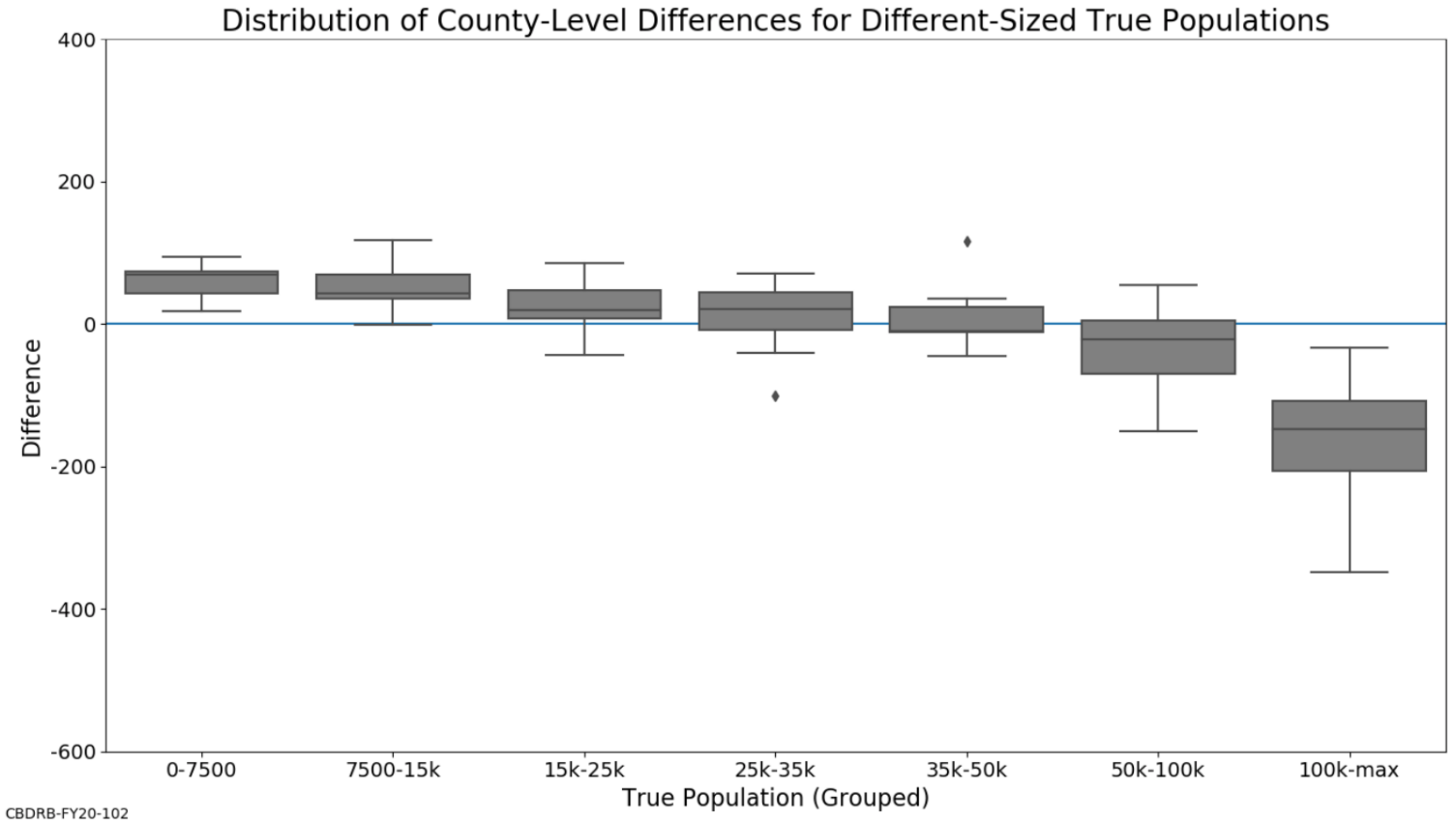
NOISE INJECTION


Middle Case Scenario
-County/Tract Group/Tract/BG/Block
-Detailed person or detailed housing
0.012
0.009
0.006
0.003
0.000
p(x)
29
125
-600
-300
0
300
600
x

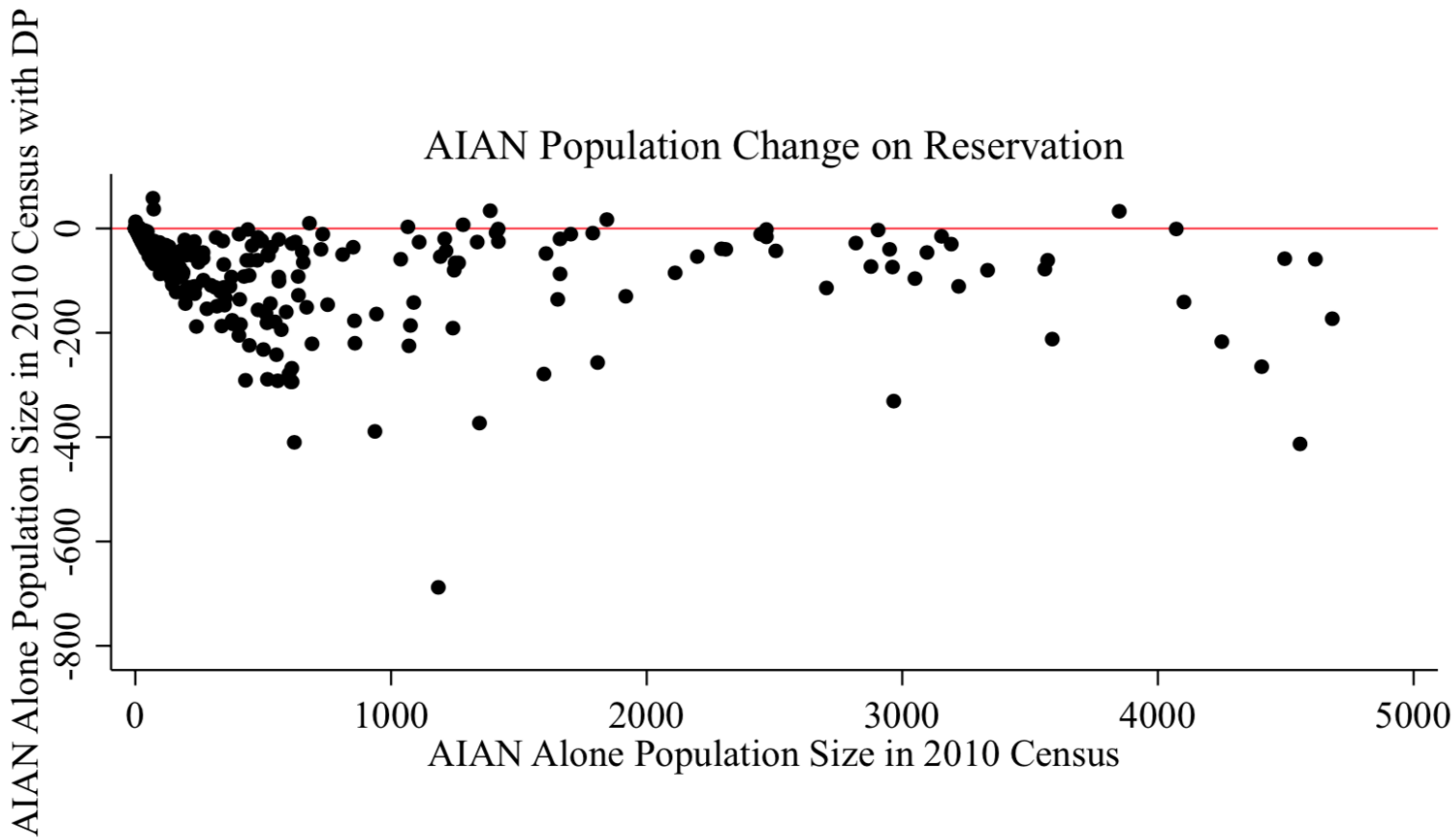
ANALYZING DIFFERENTIALLY PRIVATE 2010 CENSUS DATA

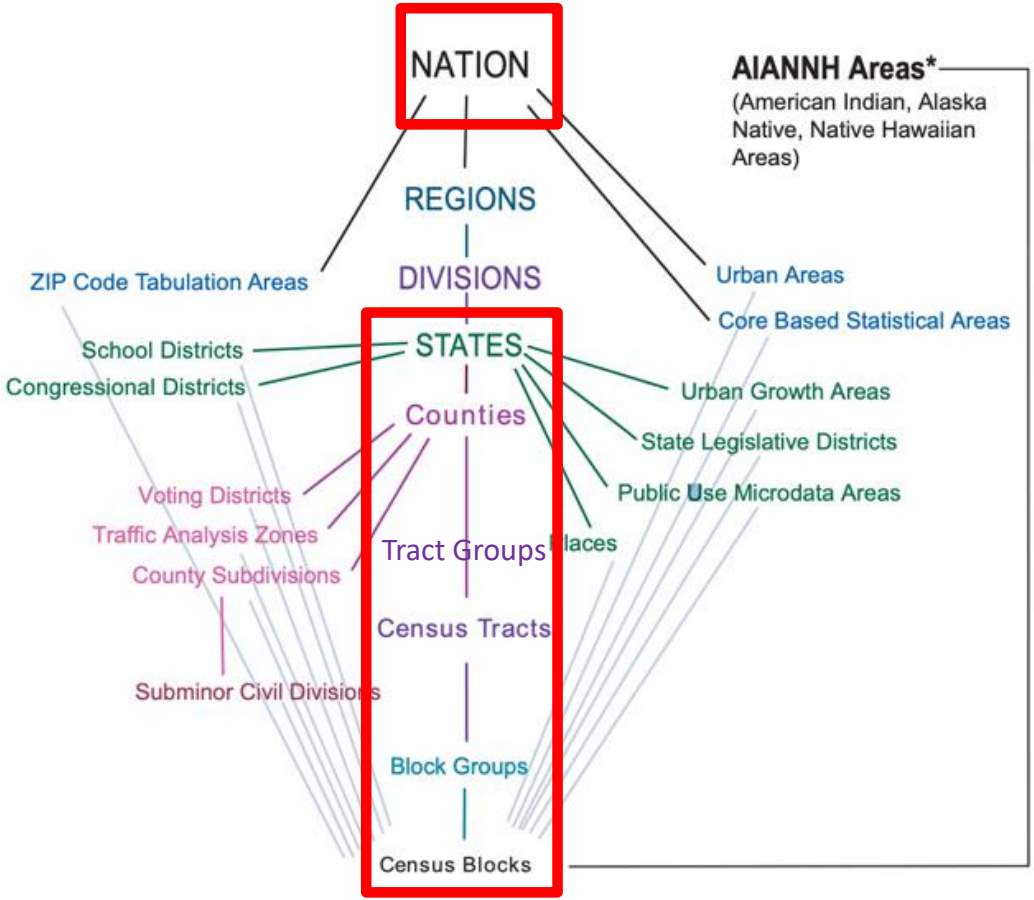
- Results based on talks given at the **Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations**
  - Hosted by CNStat
  - December 11-12, 2019
  - https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518

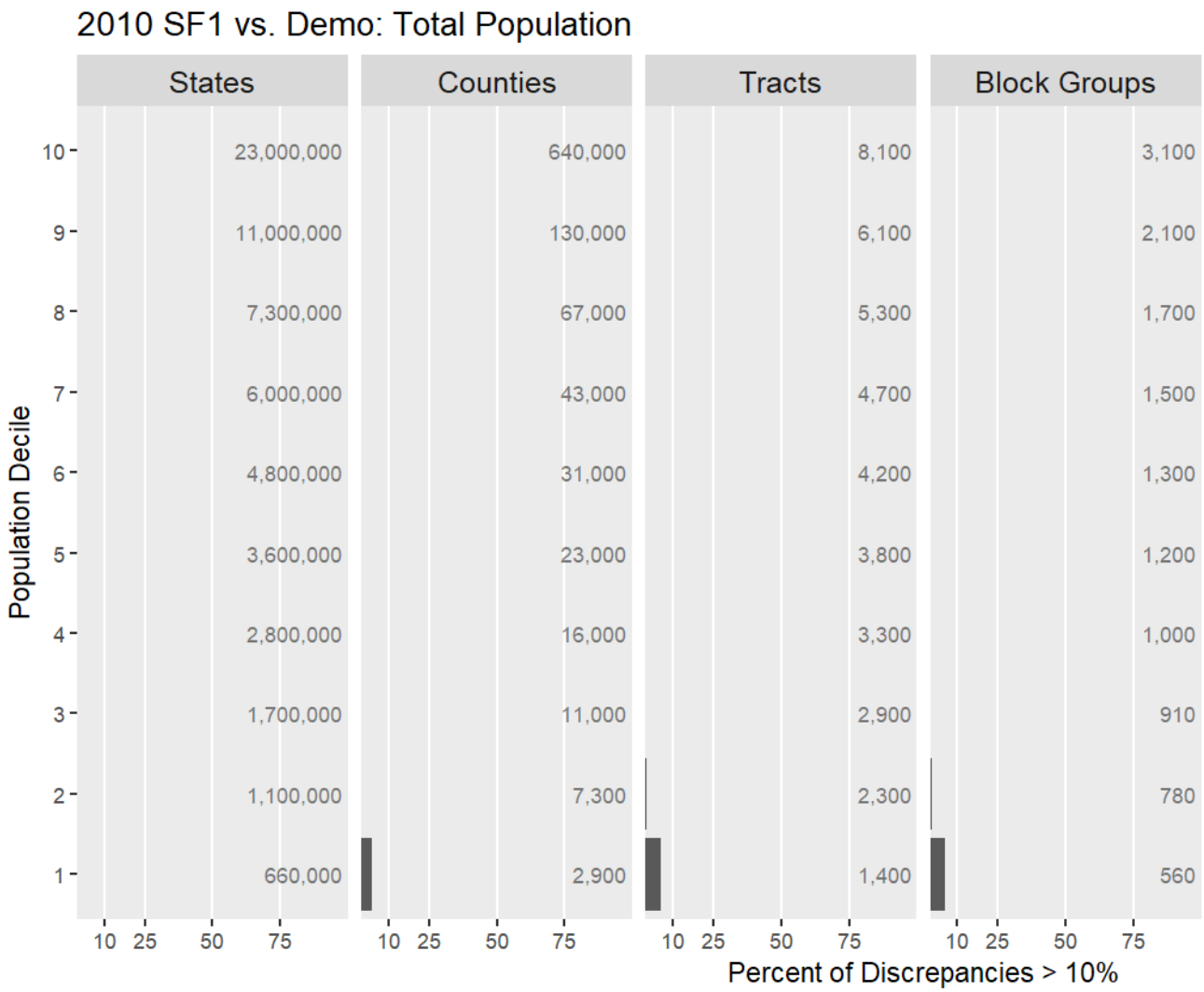
4

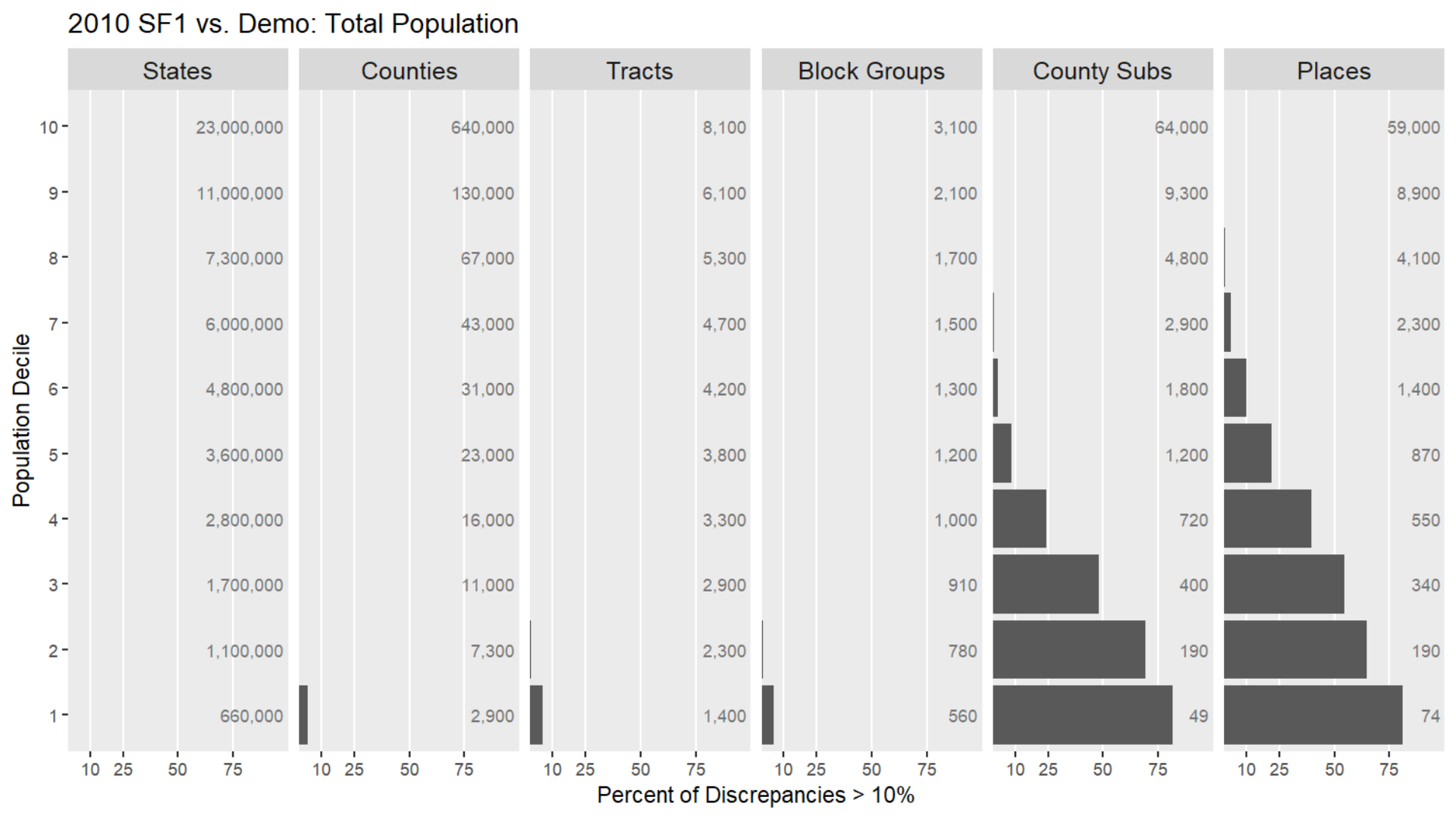


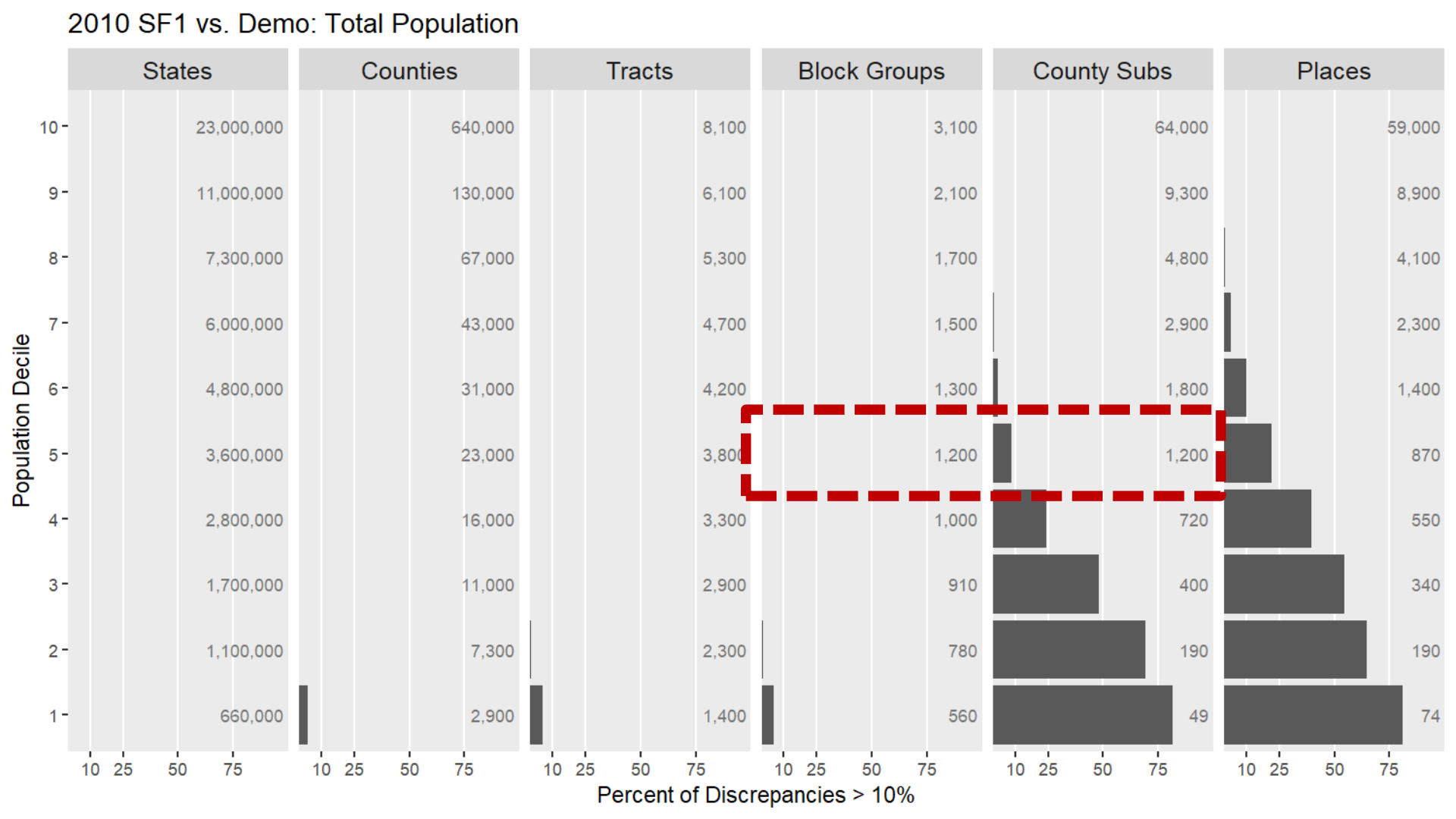
CBDRB-FY20-102

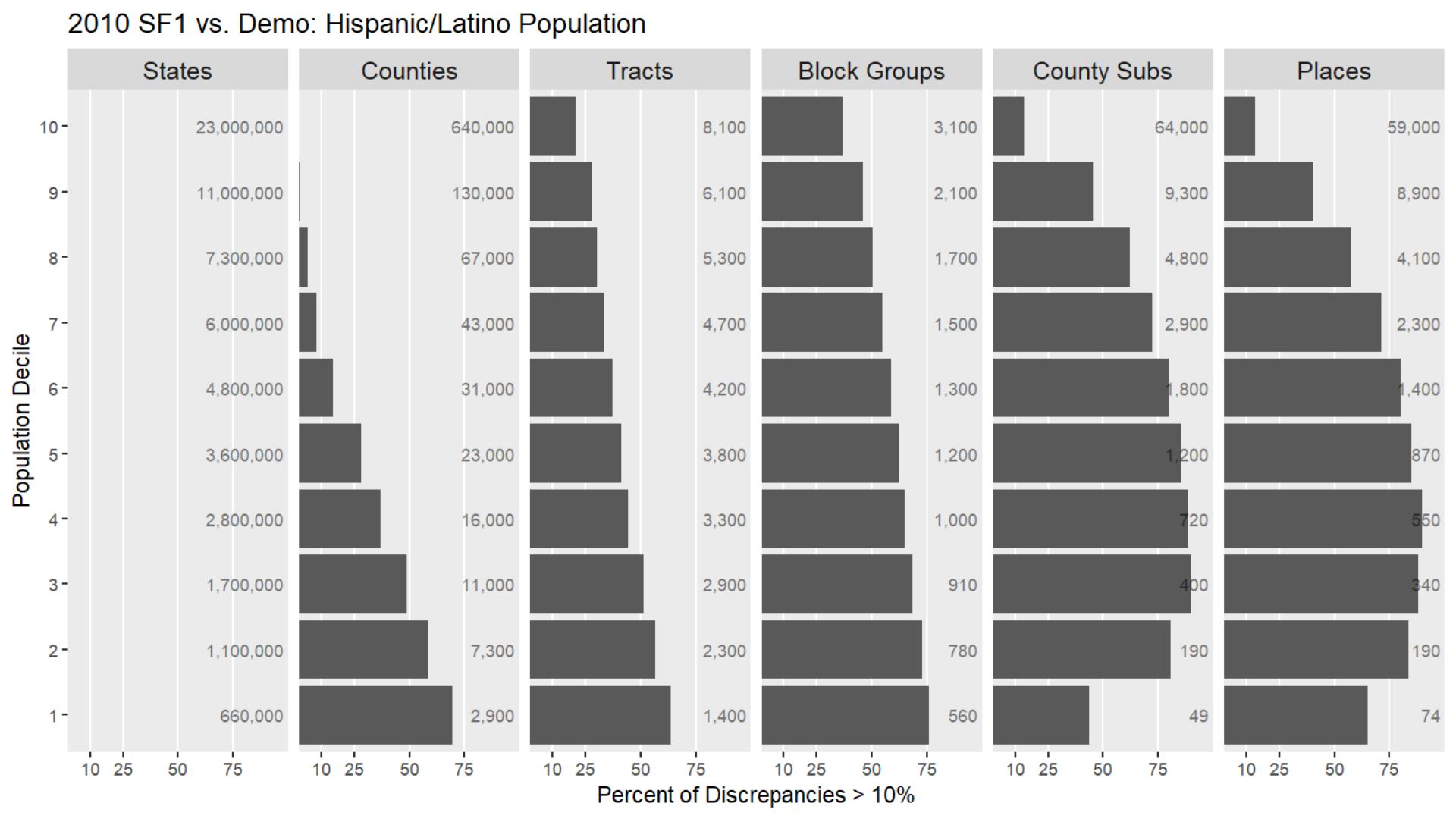

AIAN Population Change on Reservation
AIAN Alone Population Size in 2010 Census with DP
-800
-600
-400
-200
0
0
1000
2000
3000
4000
5000
AIAN Alone Population Size in 2010 Census


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

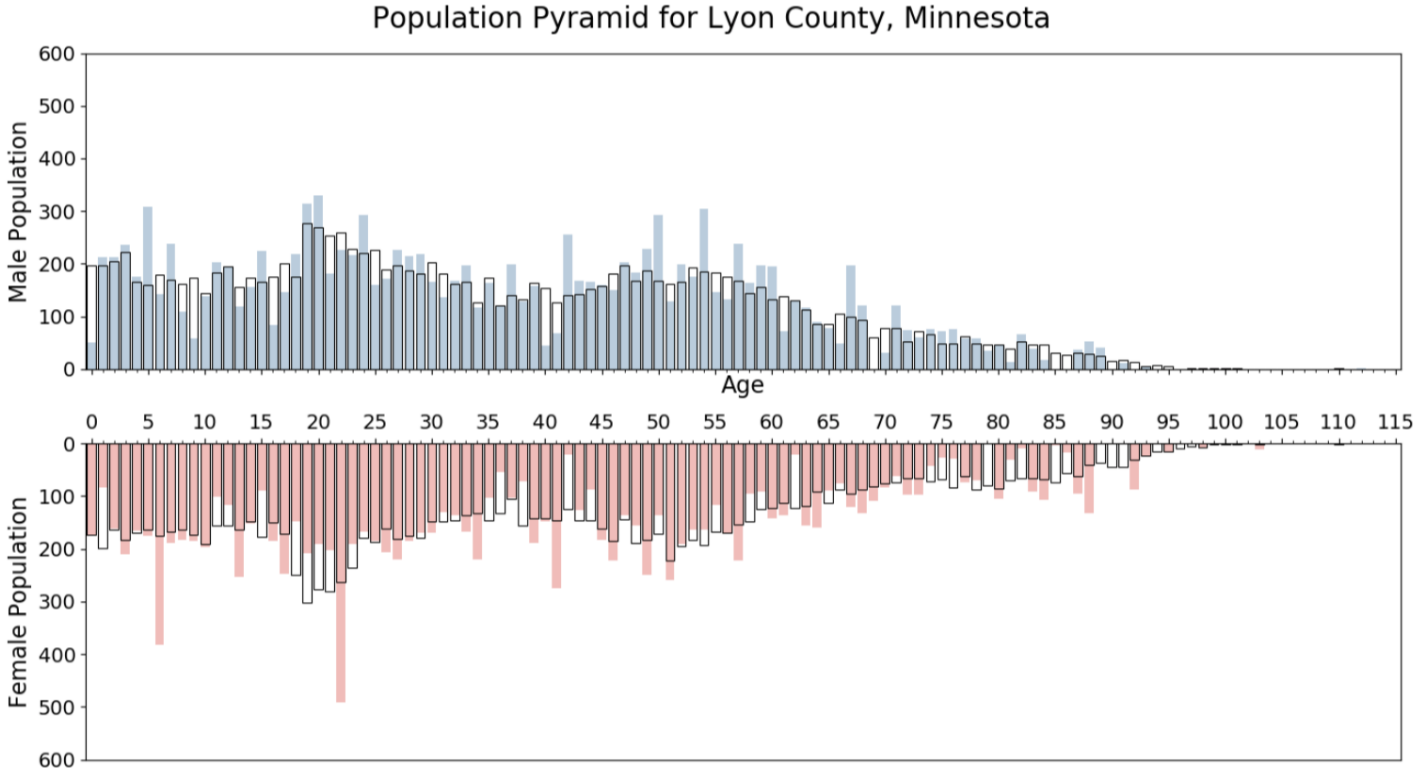

2010 SF1 vs. Demo: Total Population
States
Counties
Tracts
Block Groups
Population Decile
10
9
8
7
6
5
4
3
2
1
23,000,000
11,000,000
7,300,000
6,000,000
4,800,000
3,600,000
2,800,000
1,700,000
1,100,000
660,000
640,000
130,000
67,000
43,000
31,000
23,000
16,000
11,000
7,300
2,900
8,100
6,100
5,300
4,700
4,200
3,800
3,300
2,900
2,300
1,400
3,100
2,100
1,700
1,500
1,300
1,200
1,000
910
780
560
10 25 50 75
Percent of Discrepancies > 10%


2010 SF1 vs. Demo: Total Population
States
Counties
Tracts
Block Groups
County Subs
Places
Population Decile
10
9
8
7
6
5
4
3
2
1
23,000,000
11,000,000
7,300,000
6,000,000
4,800,000
3,600,000
2,800,000
1,700,000
1,100,000
660,000
640,000
130,000
67,000
43,000
31,000
23,000
16,000
11,000
7,300
2,900
8,100
6,100
5,300
4,700
4,200
3,800
3,300
2,900
2,300
1,400
3,100
2,100
1,700
1,500
1,300
1,200
1,000
910
780
560
64,000
9,300
4,800
2,900
1,800
1,200
720
400
190
49
59,000
8,900
4,100
2,300
1,400
870
550
340
190
74
10 25 50 75
Percent of Discrepancies > 10%

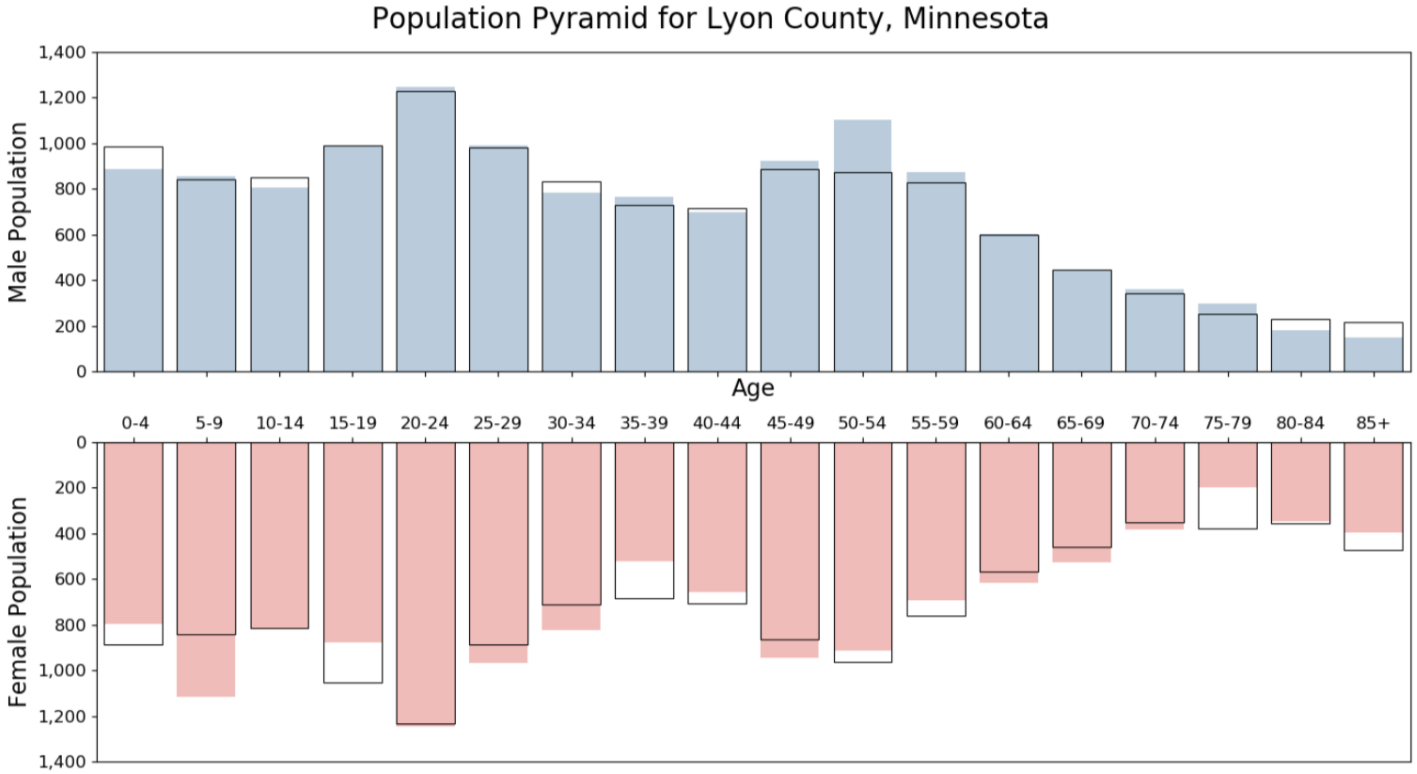

2010 SF1 vs. Demo: Total Population
States
Counties
Tracts
Block Groups
County Subs
Places
Population Decile
10
9
8
7
6
5
4
3
2
1
23,000,000
11,000,000
7,300,000
6,000,000
4,800,000
3,600,000
2,800,000
1,700,000
1,100,000
660,000
640,000
130,000
67,000
43,000
31,000
23,000
16,000
11,000
7,300
2,900
8,100
6,100
5,300
4,700
4,200
3,800
3,300
2,900
2,300
1,400
3,100
2,100
1,700
1,500
1,300
1,200
1,000
910
780
560
64,000
9,300
4,800
2,900
1,800
1,200
720
400
190
49
59,000
8,900
4,100
2,300
1,400
870
550
340
190
74
10 25 50 75
Percent of Discrepancies > 10%

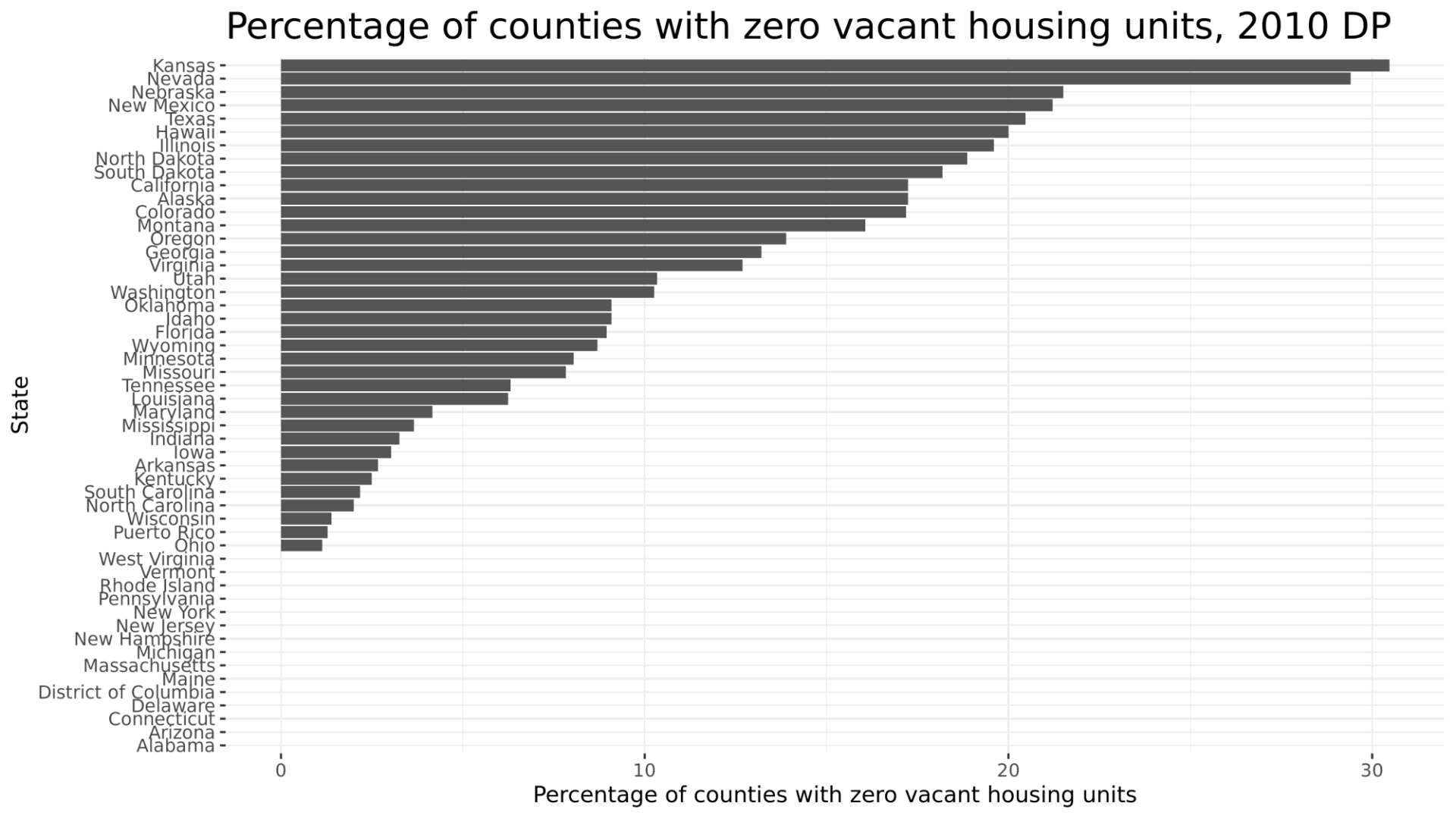

2010 SF1 vs. Demo: Hispanic/Latino Population
States
Counties
Tracts
Block Groups
County Subs
Places
Population Decile
10
9
8
7
6
5
4
3
2
1
23,000,000
11,000,000
7,300,000
6,000,000
4,800,000
3,600,000
2,800,000
1,700,000
1,100,000
660,000
640,000
130,000
67,000
43,000
31,000
23,000
16,000
11,000
7,300
2,900
8,100
6,100
5,300
4,700
4,200
3,800
3,300
2,900
2,300
1,400
3,100
2,100
1,700
1,500
1,300
1,200
1,000
910
780
560
64,000
9,300
4,800
2,900
1,800
1,200
720
400
190
49
59,000
8,900
4,100
2,300
1,400
870
550
340
190
74
10 25 50 75
Percent of Discrepancies > 10%

Age & Sex: Median County



Age & Sex: Median County, 5-Year




Percentage of counties with zero vacant housing units, 2010 DP
State
Kansas
Nevada
Nebraska
New Mexico
Texas
Hawaii
Illinois
North Dakota
South Dakota
California
Alaska
Colorado
Montana
Oregon
Georgia
Virginia
Utah
Washington
Oklahoma
Idaho
Florida
Wyoming
Minnesota
Missouri
Tennessee
Louisiana
Maryland
Mississippi
Indiana
Iowa
Arkansas
Kentucky
South Carolina
North Carolina
Wisconsin
Puerto Rico
Ohio
West Virginia
Vermont
Rhode Island
Pennsylvania
New York
New Jersey
New Hampshire
Michigan
Massachusetts
Maine
District of Columbia
Delaware
Connecticut
Arizona
Alabama
0
10
20
30
Percentage of counties with zero vacant housing units

LESS PUBLICLY AVAILABLE DATA

Census 2010 – 2020 Crosswalk

- https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/2020-census-data-products-planning-crosswalk.xlsx

2020 Census Demographic and Housing Characteristics File Proposed List of Tables						
2010 Census Table Number	2020 Census Table Number	Title	Lowest Level of Geography Proposed for 2020	Proposed for Inclusion in 2020 DHC	Lowest Level of Geography Proposed for 2010 Demonstration Data Product	Proposed for Inclusion in 2010 Demonstration Data Product
P35A.	N/A	FAMILIES (WHITE ALONE HOUSEHOLDER)	N/A	No	N/A	No
P35B.	N/A	FAMILIES (BLACK OR AFRICAN AMERICAN ALONE HOUSEHOLDER)	N/A	No	N/A	No
P35C.	N/A	FAMILIES (AMERICAN INDIAN AND ALASKA NATIVE ALONE HOUSEHOLDER)	N/A	No	N/A	No
P35D.	N/A	FAMILIES (ASIAN ALONE HOUSEHOLDER)	N/A	No	N/A	No
P35E.	N/A	FAMILIES (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE HOUSEHOLDER)	N/A	No	N/A	No
P35F.	N/A	FAMILIES (SOME OTHER RACE ALONE HOUSEHOLDER)	N/A	No	N/A	No
P35G.	N/A	FAMILIES (TWO OR MORE RACES HOUSEHOLDER)	N/A	No	N/A	No
P35H.	N/A	FAMILIES (HISPANIC OR LATINO HOUSEHOLDER)	N/A	No	N/A	No
P35I.	N/A	FAMILIES (WHITE ALONE, NOT HISPANIC OR LATINO HOUSEHOLDER)	N/A	No	N/A	No

2020 Census Demographic and Housing Characteristics File Proposed List of Tables						
2010 Census Table Number	2020 Census Table Number	Title	Lowest Level of Geography Proposed for 2020	Proposed for Inclusion in 2020 DHC	Lowest Level of Geography Proposed for 2010 Demonstration Data Product	Proposed for Inclusion in 2010 Demonstration Data Product
P38A.	PCO11A.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (WHITE ALONE HOUSEHOLDER)	County	Yes	Block	Yes
P38B.	PCO11B.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (BLACK OR AFRICAN AMERICAN ALONE HOUSEHOLDER)	County	Yes	Block	Yes
P38C.	PCO11C.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (AMERICAN INDIAN AND ALASKA NATIVE ALONE HOUSEHOLDER)	County	Yes	Block	Yes
P38D.	PCO11D.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (ASIAN ALONE HOUSEHOLDER)	County	Yes	Block	Yes
P38E.	PCO11E.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE HOUSEHOLDER)	County	Yes	Block	Yes
P38F.	PCO11F.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (SOME OTHER RACE ALONE HOUSEHOLDER)	County	Yes	Block	Yes
P38G.	PCO11G.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (TWO OR MORE RACES HOUSEHOLDER)	County	Yes	Block	Yes
P38H.	PCO11H.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (HISPANIC OR LATINO HOUSEHOLDER)	County	Yes	Block	Yes
P38I.	PCO11I.	FAMILY TYPE BY PRESENCE AND AGE OF OWN CHILDREN (WHITE ALONE, NOT HISPANIC OR LATINO HOUSEHOLDER)	County	Yes	Block	Yes

Concept	Finest 2010 geog	Finest 2020 geog (proposed)
Race	Block	Block and TBD
Households	Block	County
Families	Block	N/A
Group quarters	Block/tract	County/state

LESS CONSISTENCY AMONG PRODUCTS

Group 1 products

- Apportionment
- PL94-171
- Demographic and Housing Characteristics (DHC)
  - Replaces SF1
- Demographic Profile
- Congressional District DHC

Group 2

- Detailed race/ethnicity
- American Indian and Alaska Native Summary File
- Person—household joins

Less consistency in counts

- AI/AN_{DHC} != $AI/AN_{summary\ file}$
- $County_{DHC}$!= $County_{Detailed\ race/ethnicity\ summary}$

Next steps

Next steps

- Census modifying its algorithm to try and fix issues found in the 2010 demonstration data

Next steps

- Census modifying its algorithm to try and fix issues found in the 2010 demonstration data
- Timeline is short
  - I'm not sure Census has time to address all issues and create usable data

Next steps

- Census modifying its algorithm to try and fix issues found in the 2010 demonstration data
- Timeline is short
  - I'm not sure Census has time to address all issues and create usable data
- Need another demonstration dataset

Resources

- Census Bureau Disclosure avoidance
  - https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html
- IPUMS Differential Privacy
  - https://ipums.org/changes-to-census-bureau-data-products
- New York Times editorial
  - https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html