Geographic Review of Differentially Private Demonstration Data

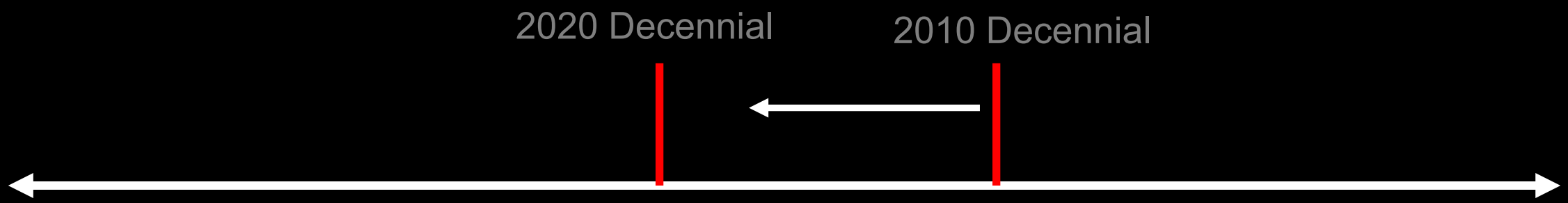
David Van Riper, University of Minnesota
Seth Spielman, University of Colorado

Agenda

- Preliminary investigations of the differences between the original 2010 data and the new DP data.
  - Various types of geography
  - Metro-scale measures of inequality
  - Spatial patterns of populations within cities
- DP Products and Process.

Big Picture: Decennial Data

Big Picture: Census2020



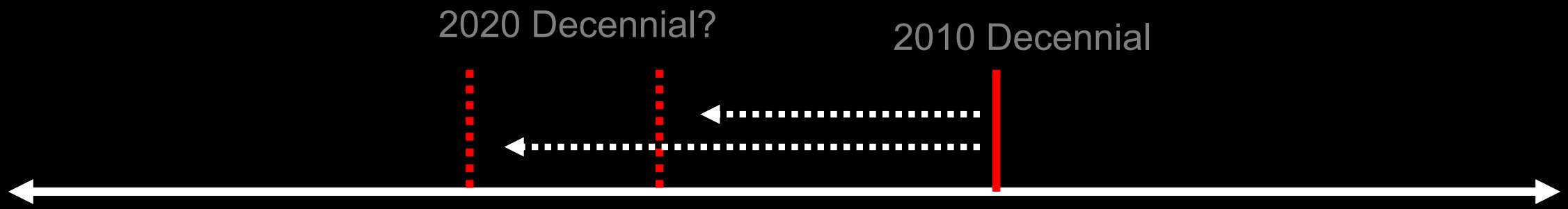
Big Picture: Census2020

Are we on the right position in this continuum?

Privacy
Private Good

To asses this question we believe you have to get under the hood and look at how these changes materially impact data users/use-cases.

Data Availability
Public Good



Big Picture: Public Debate

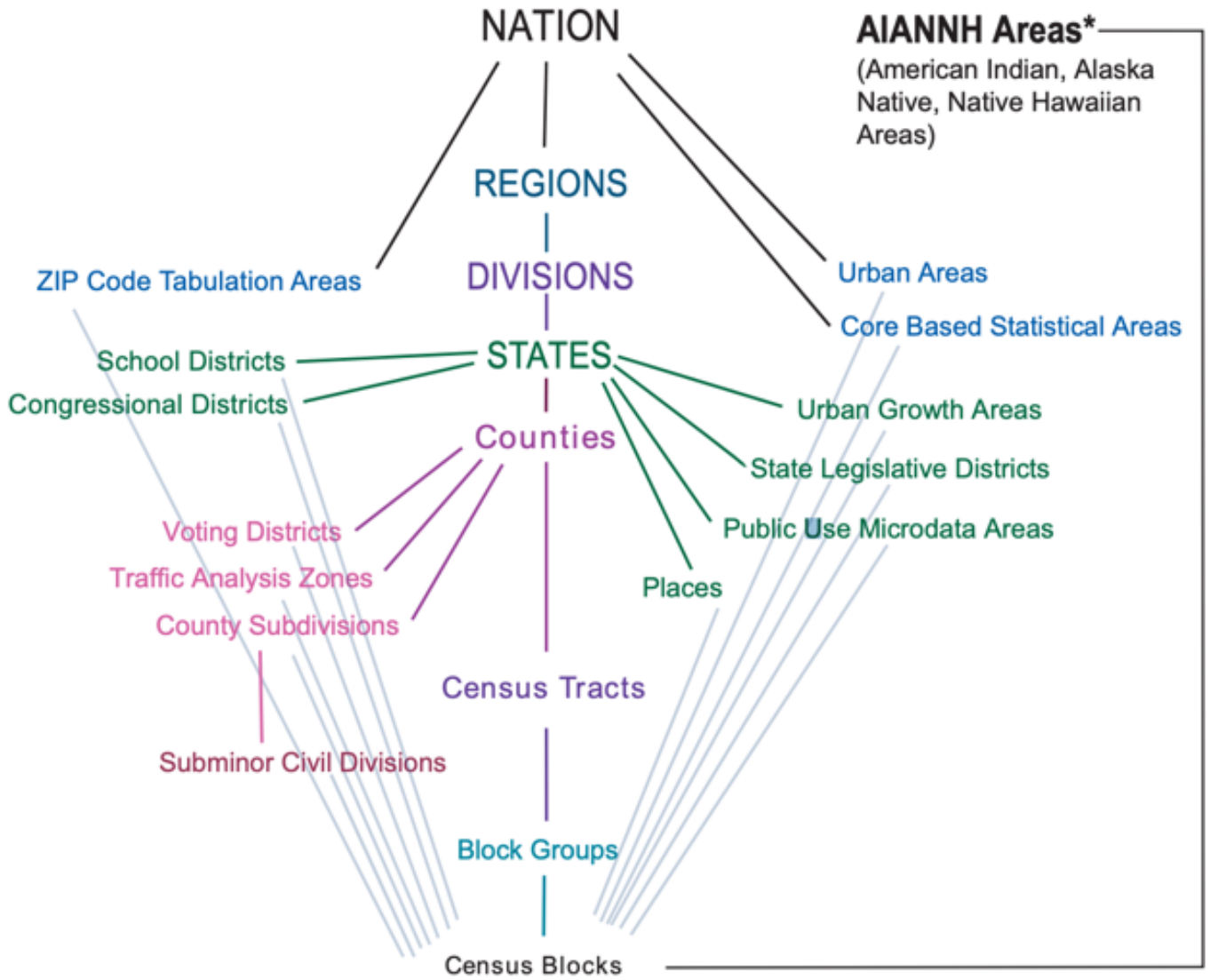
What are the changes?

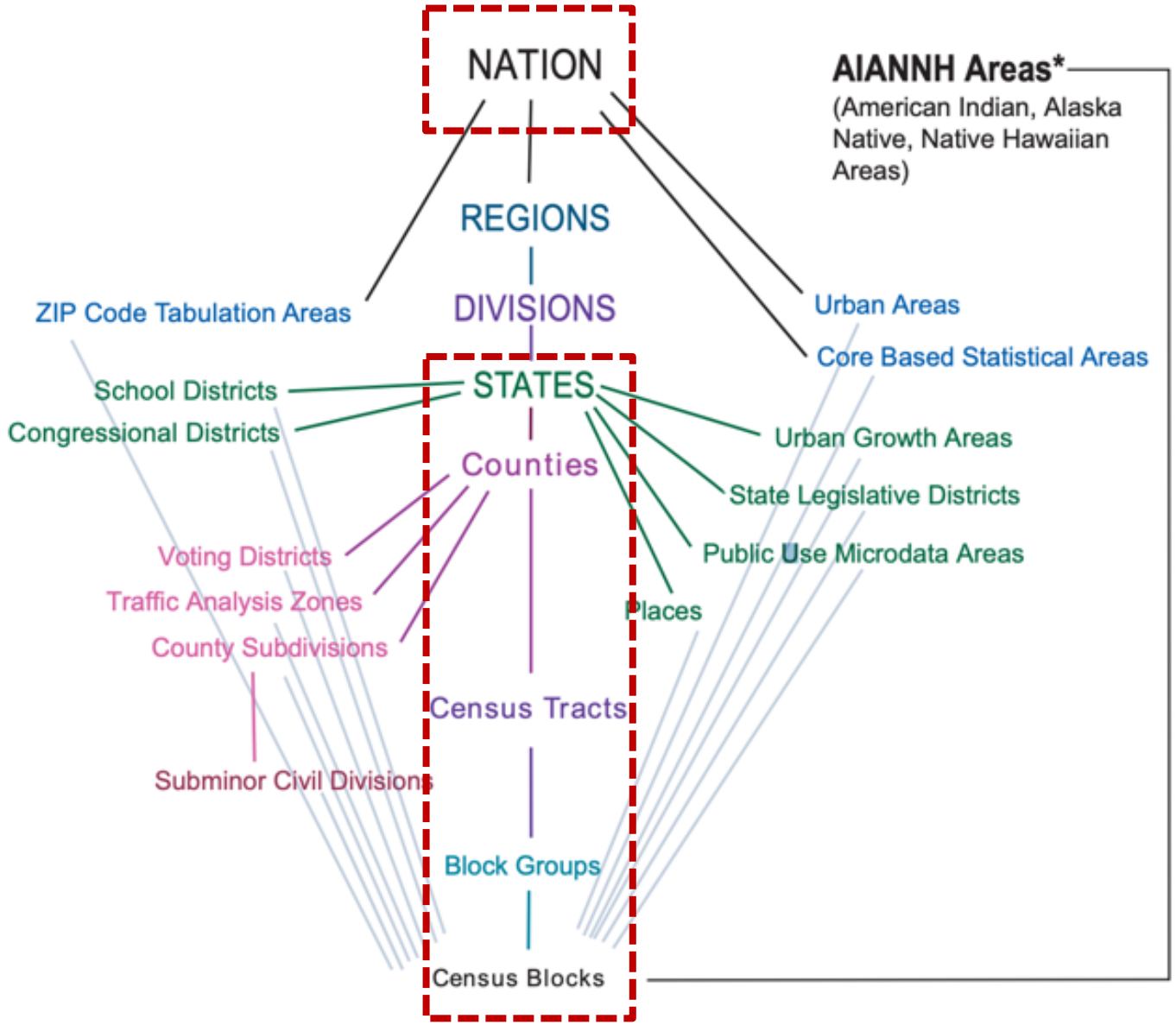
Is an empirical question. We'll share some insights, lots more to do…

Are we in the right place on the continuum?

Is a political/personal question.
We believe its very hard to honestly answer until we understand the nature of the changes and their implications for policy, planning and governance.

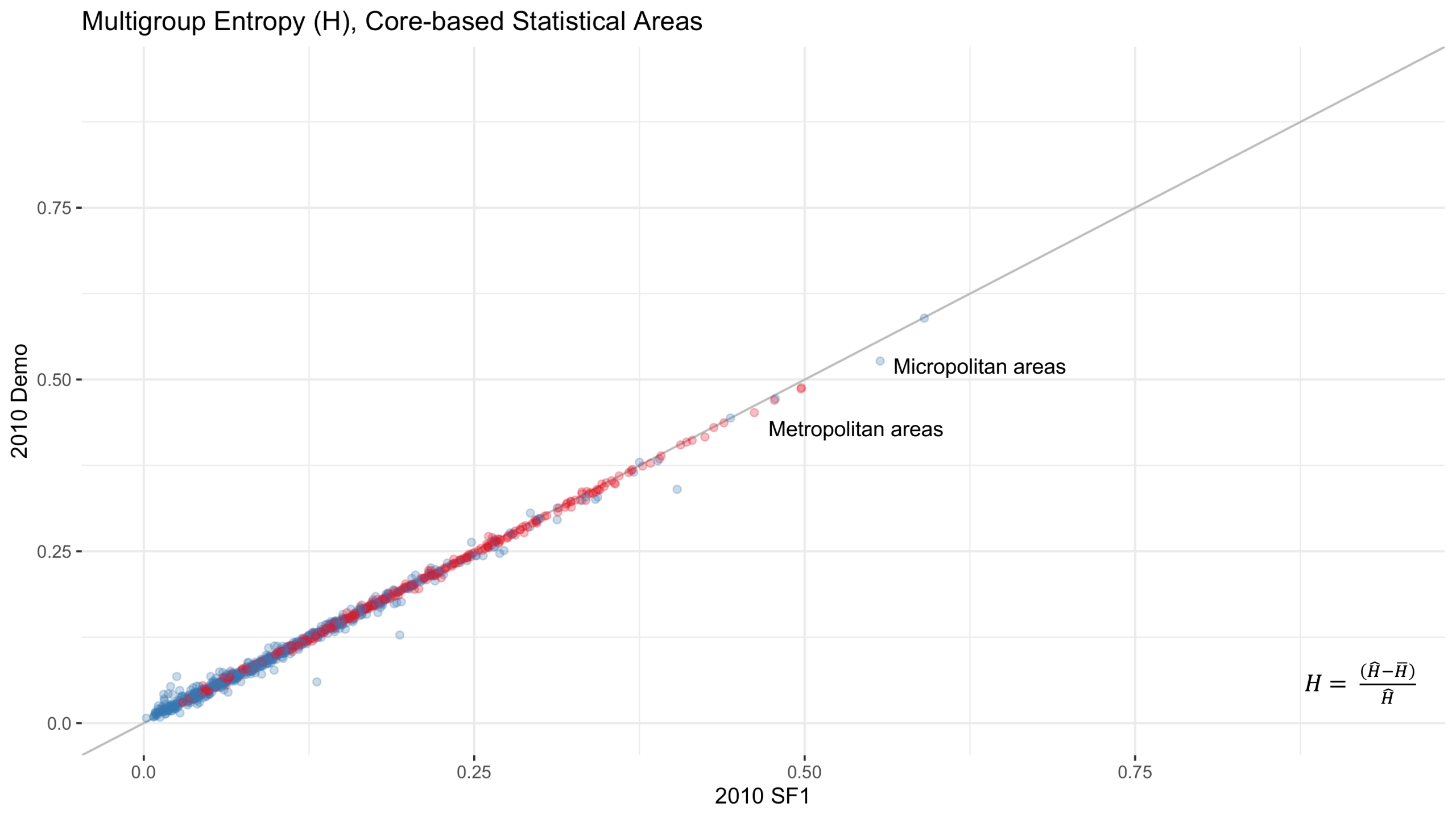
Changes in place-based data...

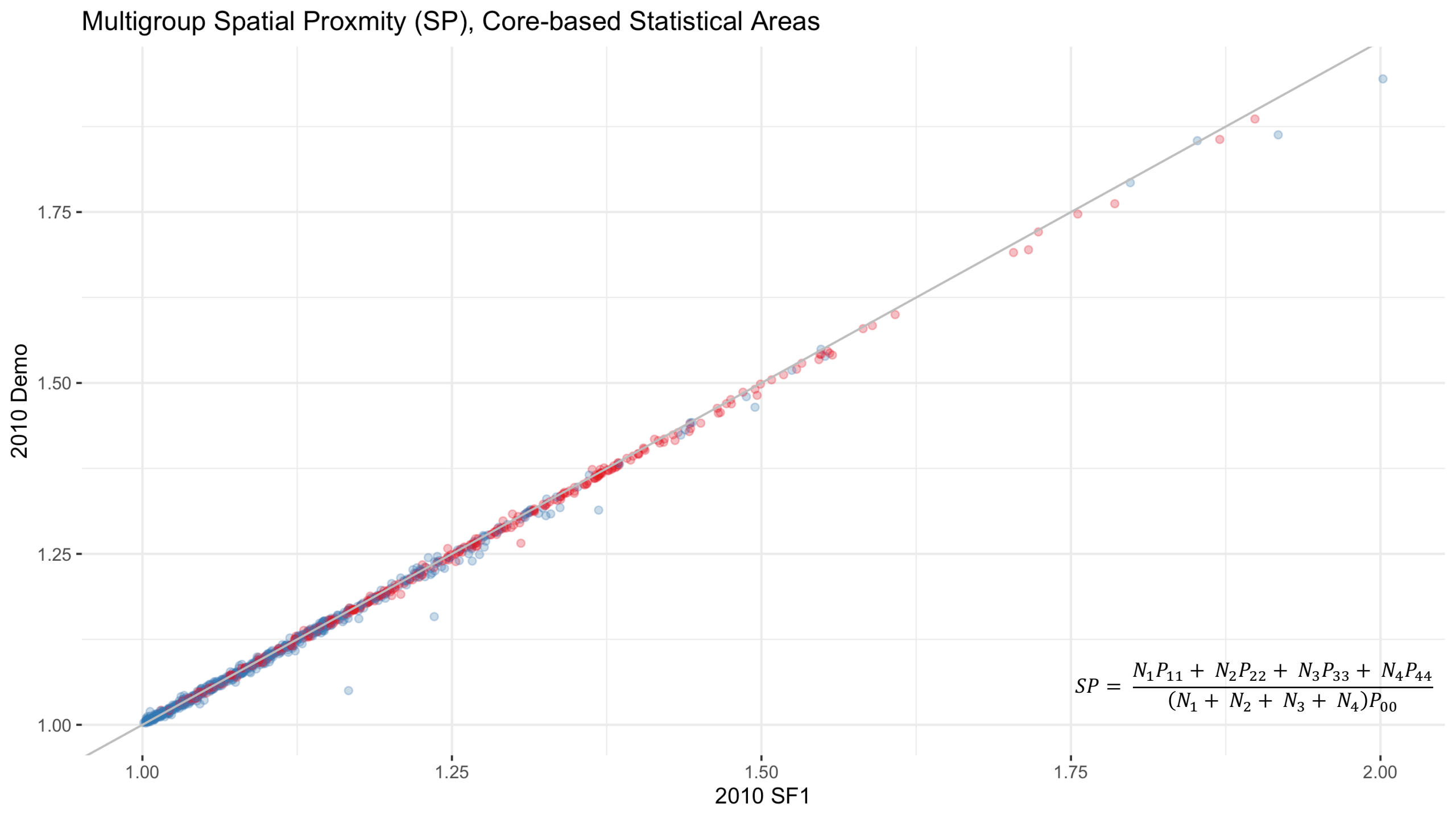

NATION
AIANNH Areas*
(American Indian, Alaska Native, Native Hawaiian Areas)
REGIONS
DIVISIONS
ZIP Code Tabulation Areas
Urban Areas
Core Based Statistical Areas
STATES
School Districts
Congressional Districts
Urban Growth Areas
Counties
State Legislative Districts
Public Use Microdata Areas
Voting Districts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

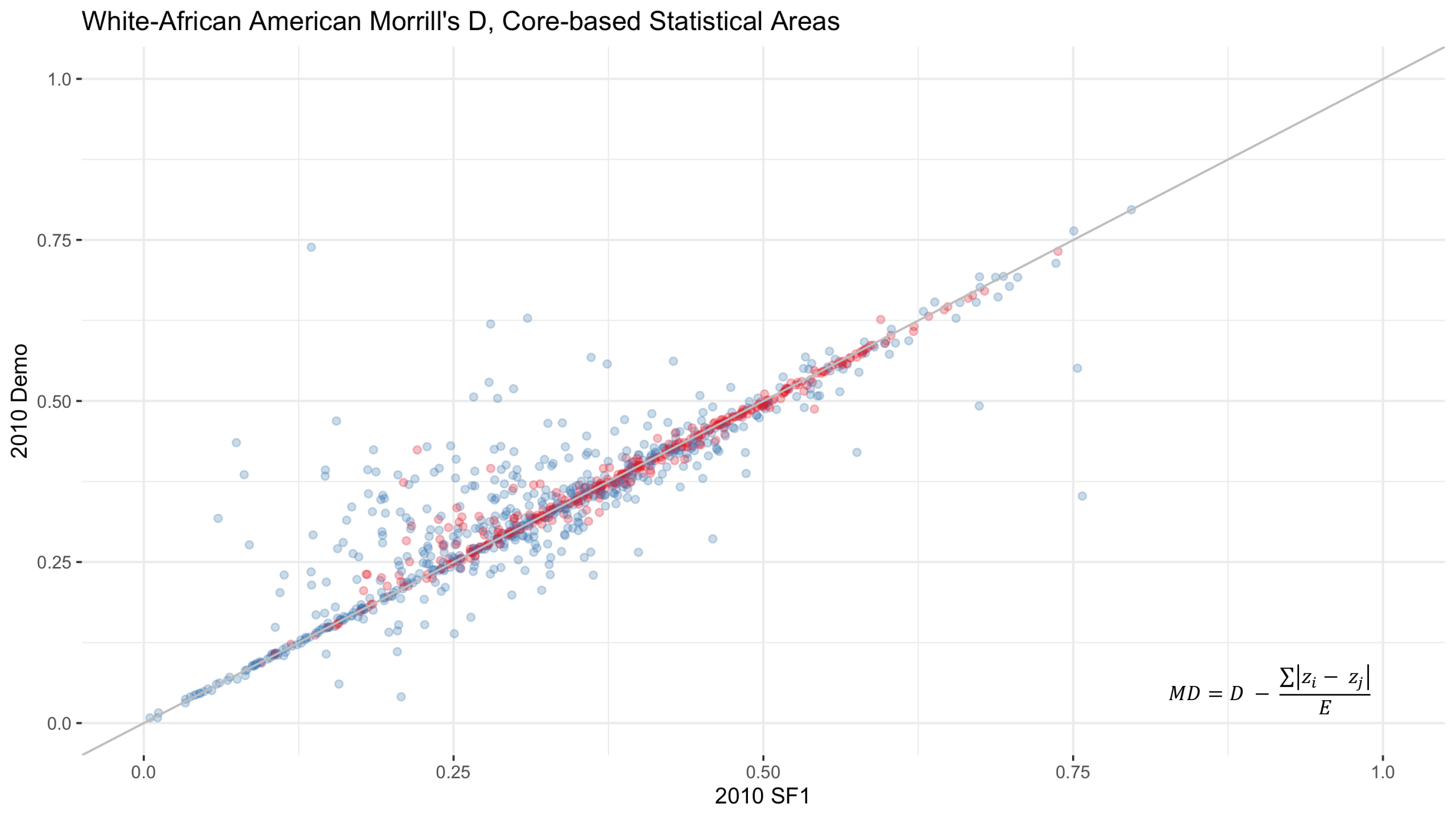

NATION
REGIONS
DIVISIONS
STATES
Counties
Census Tracts
Block Groups
Census Blocks
AIANNH Areas*
(American Indian, Alaska Native, Native Hawaiian Areas)
ZIP Code Tabulation Areas
Urban Areas
Core Based Statistical Areas
School Districts
Congressional Districts
Urban Growth Areas
State Legislative Districts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

Describing Inequality

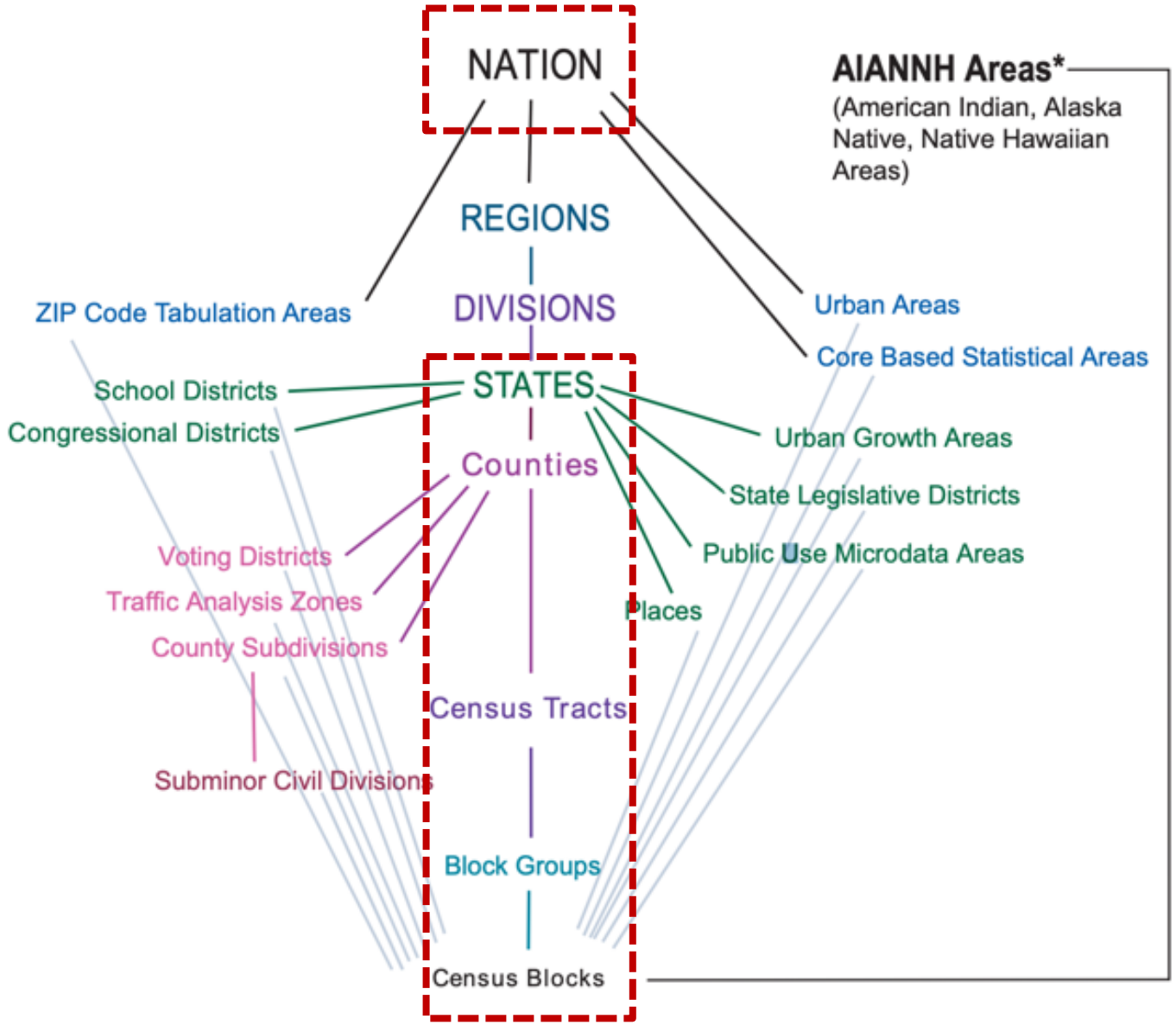
(Residential Segregation)

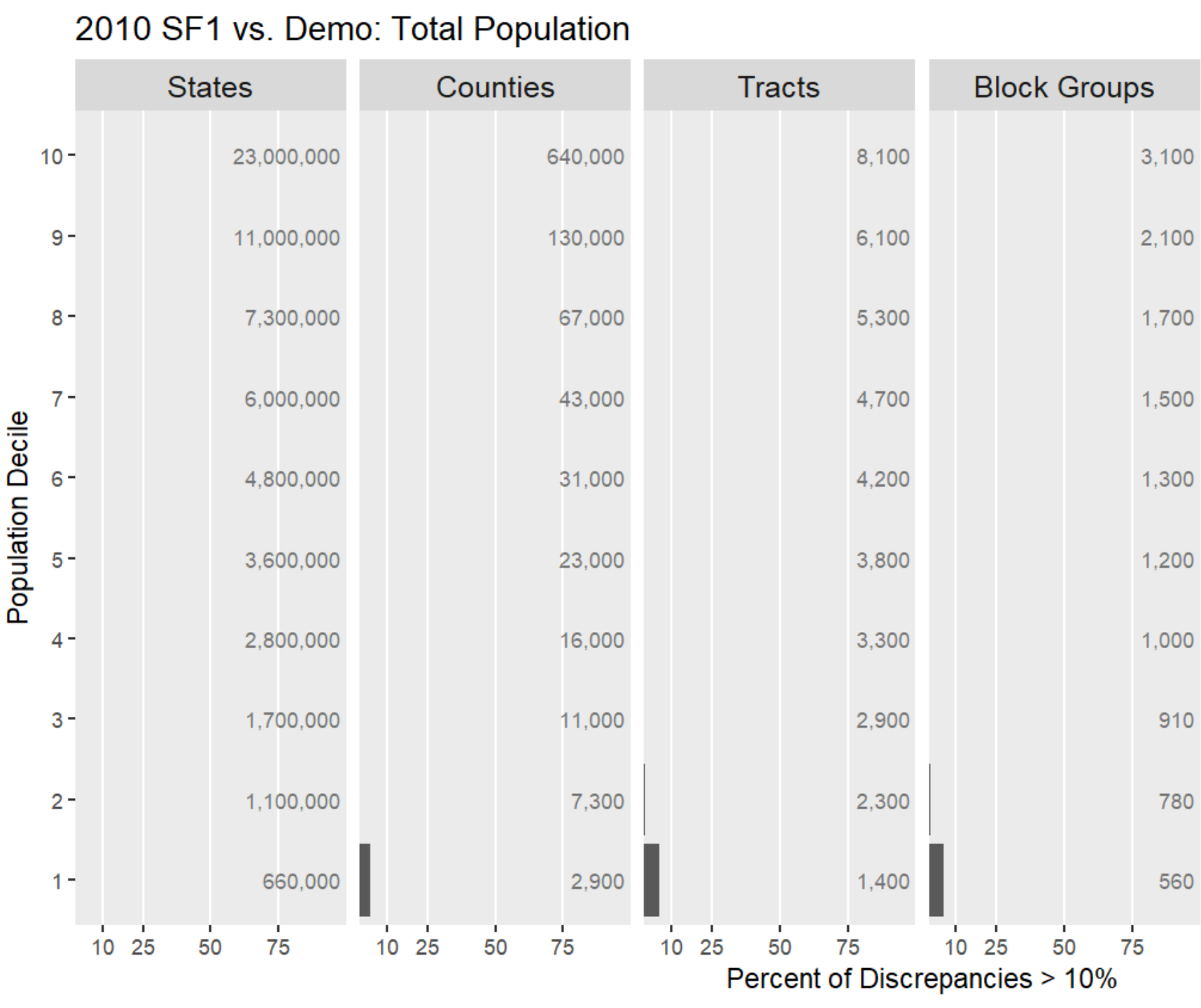

Multigroup Entropy (H), Core-based Statistical Areas
0.75
0.50
0.25
0.0
2010 Demo
0.0
0.25
0.50
0.75
2010 SF1
Micropolitan areas
Metropolitan areas
$H = \frac{(\hat{H} - \bar{H})}{\hat{H}}$

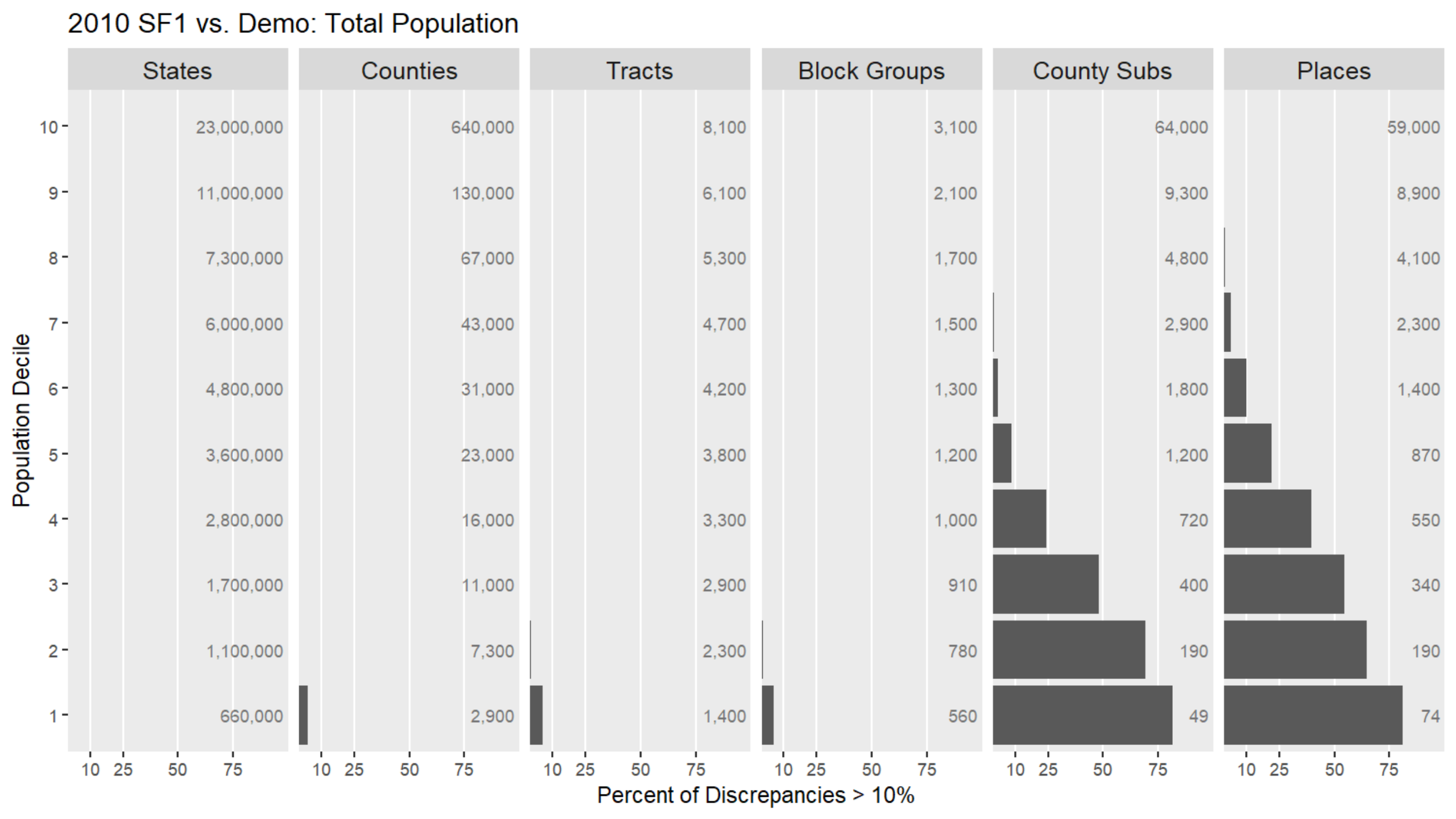

Multigroup Spatial Proxmity (SP), Core-based Statistical Areas
2010 Demo
1.00
1.25
1.50
1.75
2010 SF1
1.00
1.25
1.50
1.75
2.00
$SP = \frac{N_1P_{11} + N_2P_{22} + N_3P_{33} + N_4P_{44}}{(N_1 + N_2 + N_3 + N_4)P_{00}}$

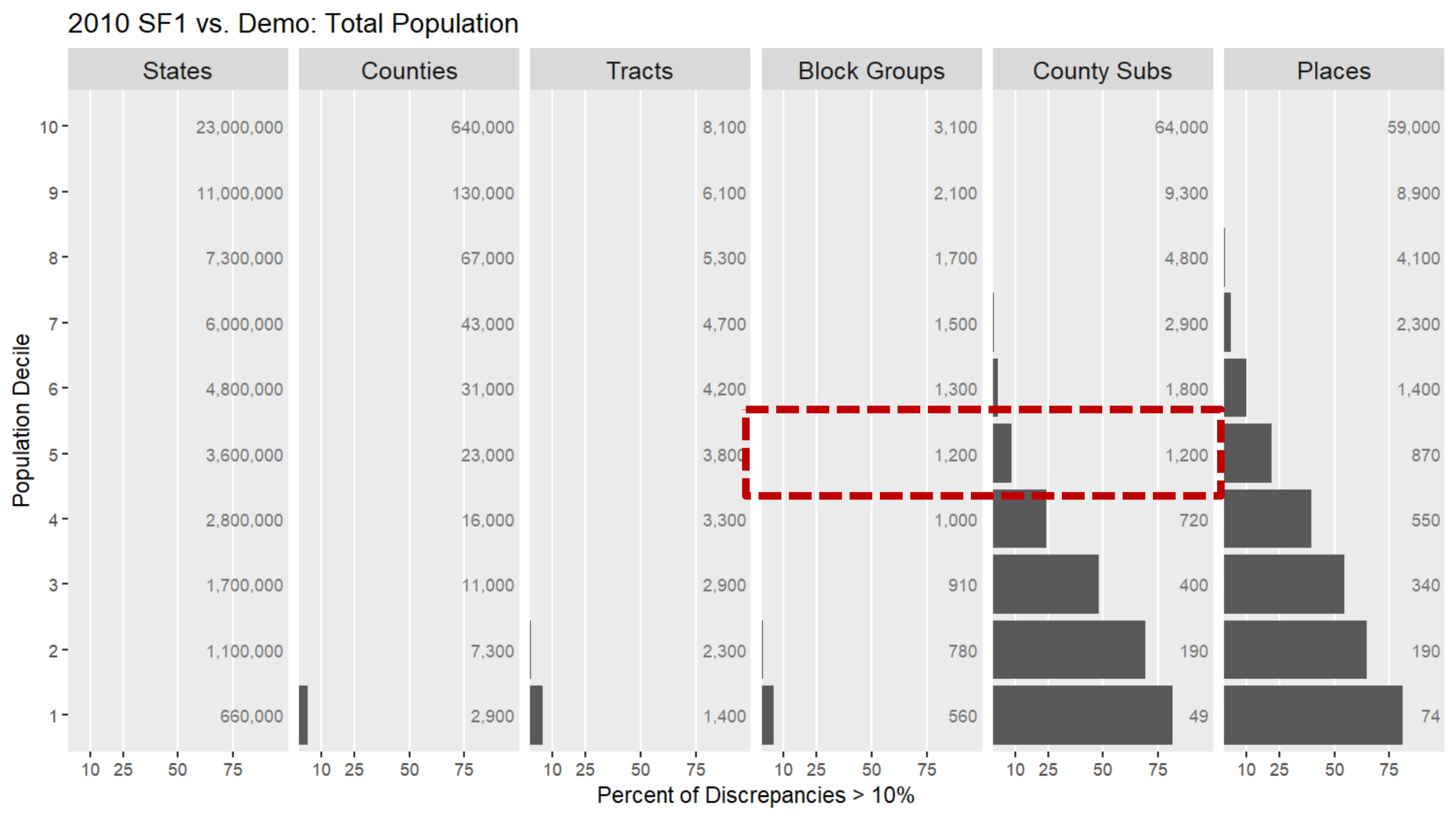

White-African American Morrill's D, Core-based Statistical Areas
1.0
0.75
0.50
0.25
0.0
2010 Demo
0.0
0.25
0.50
0.75
1.0
2010 SF1
$MD = D - \frac{\sum|z_i - z_j|}{E}$

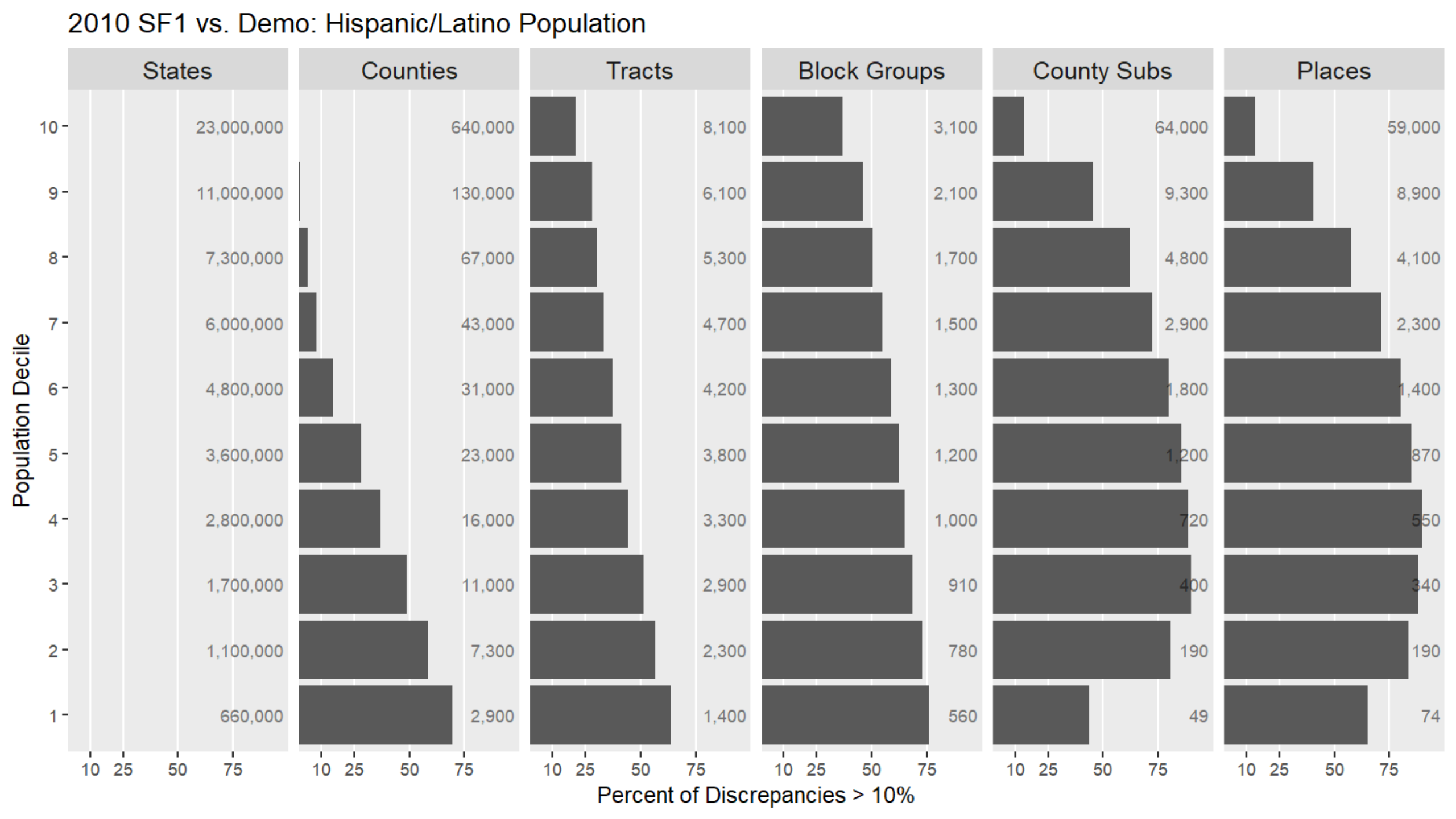
Changes along geographic hierarchy

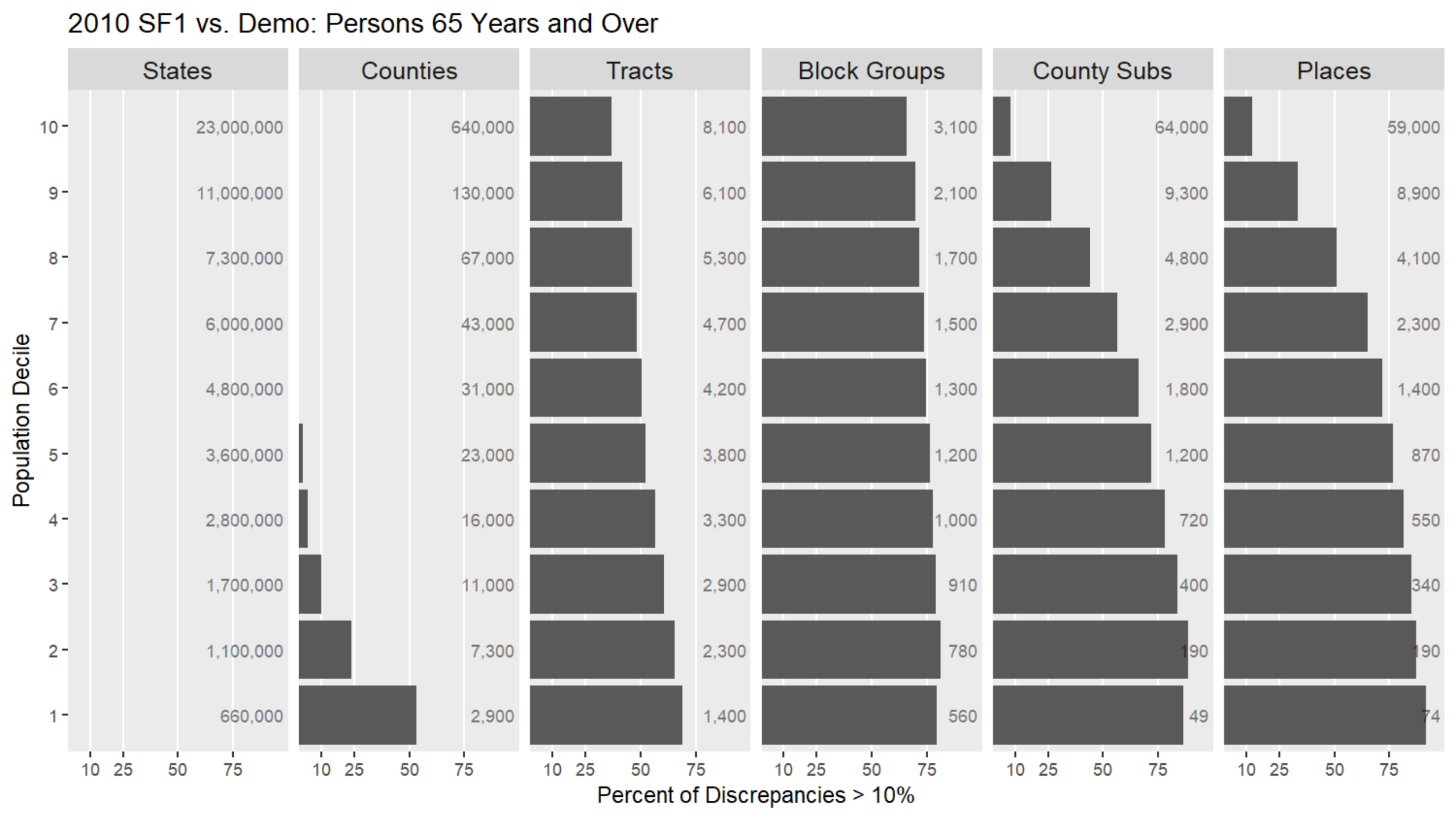

NATION
REGIONS
DIVISIONS
STATES
Counties
Census Tracts
Block Groups
Census Blocks
AIANNH Areas*
(American Indian, Alaska Native, Native Hawaiian Areas)
ZIP Code Tabulation Areas
Urban Areas
Core Based Statistical Areas
School Districts
Congressional Districts
Urban Growth Areas
State Legislative Districts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

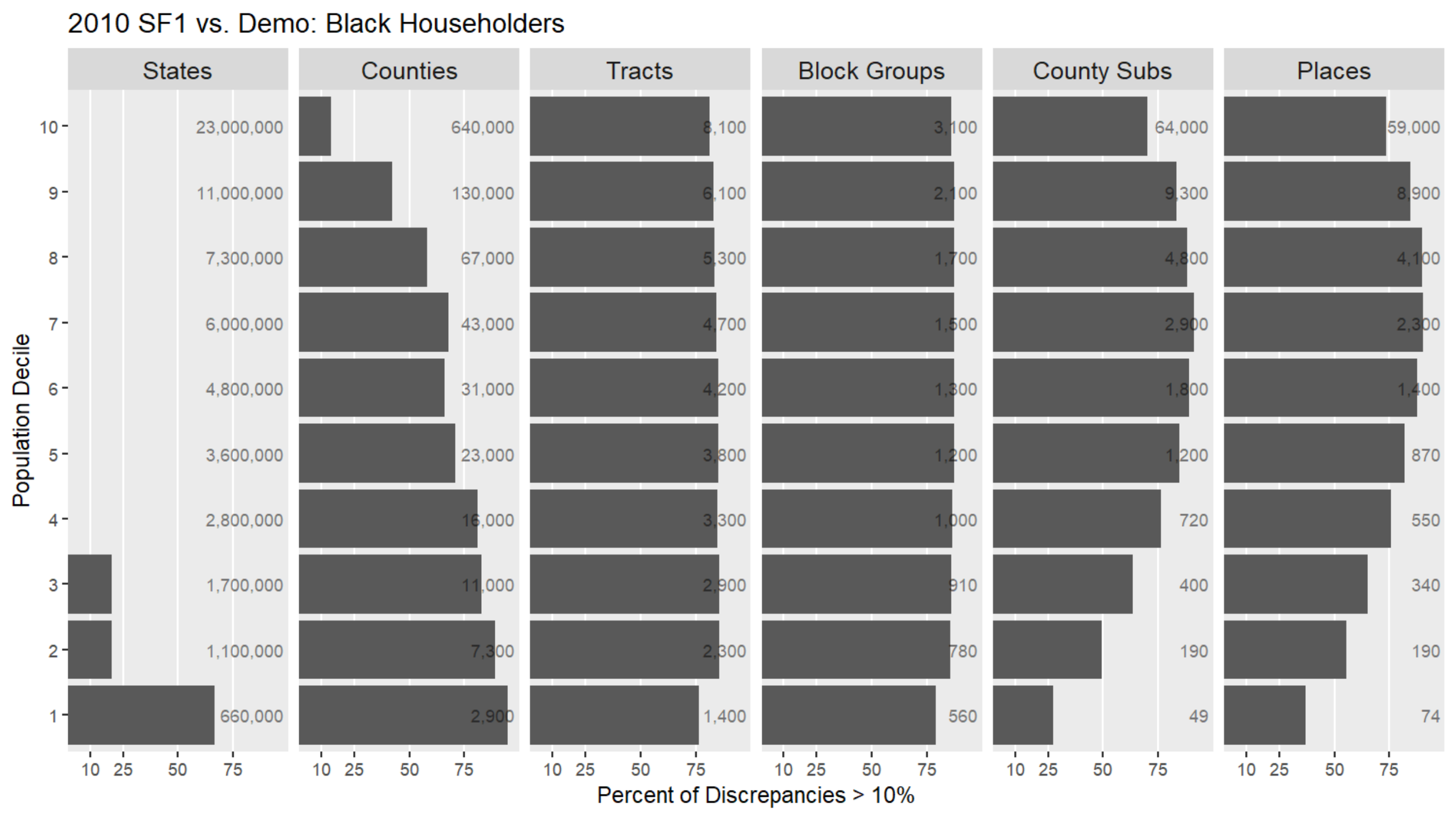

2010 SF1 vs. Demo: Total Population
States
Counties
Tracts
Block Groups
Population Decile
10
9
8
7
6
5
4
3
2
1
23,000,000
11,000,000
7,300,000
6,000,000
4,800,000
3,600,000
2,800,000
1,700,000
1,100,000
660,000
640,000
130,000
67,000
43,000
31,000
23,000
16,000
11,000
7,300
2,900
8,100
6,100
5,300
4,700
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%

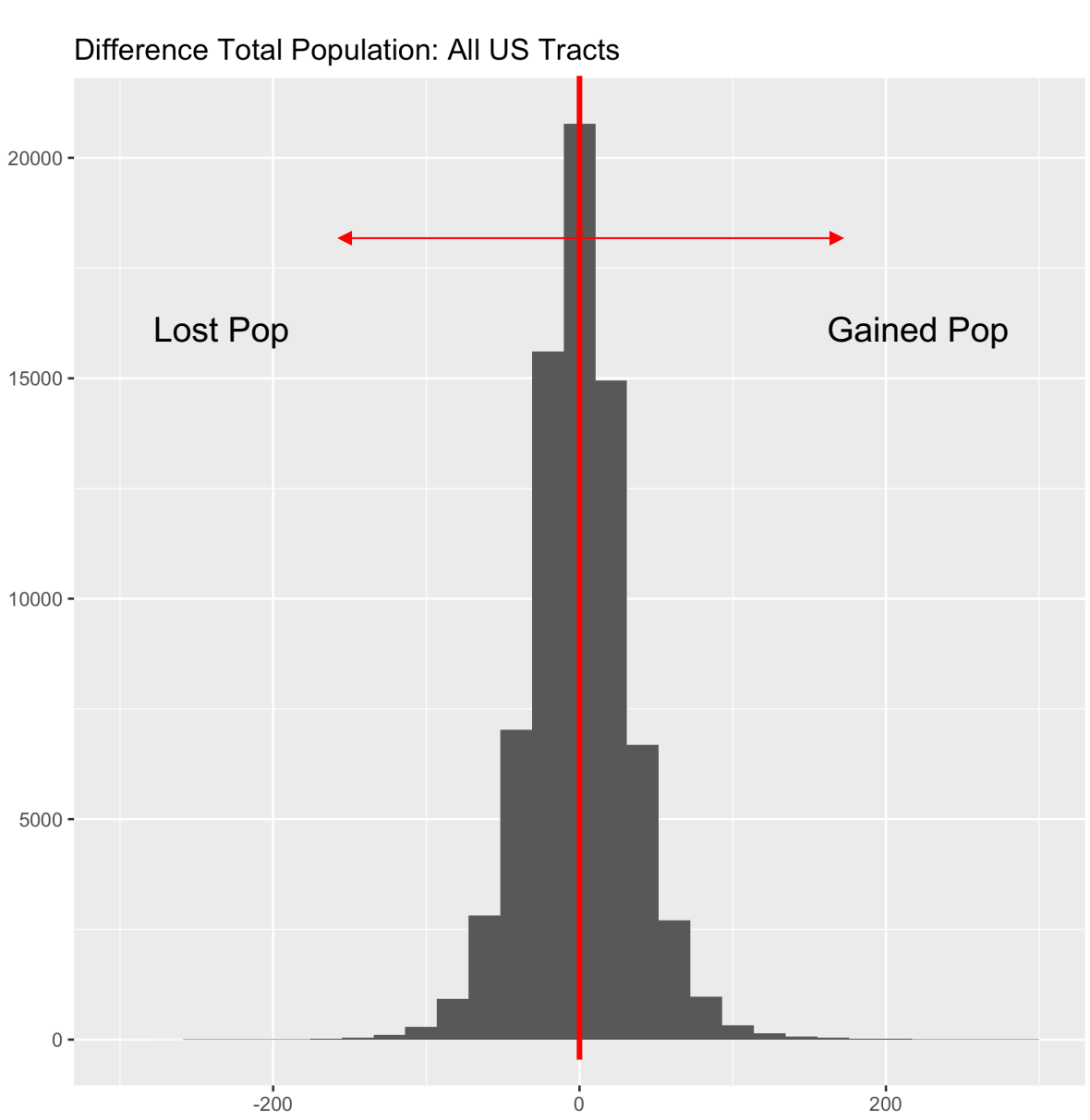

2010 SF1 vs. Demo: Persons 65 Years and Over
States
Counties
Tracts
Block Groups
County Subs
Places
Population Decile
10
9
8
7
6
5
4
3
2
1
23,000,000
11,000,000
7,300,000
6,000,000
4,800,000
3,600,000
2,800,000
1,700,000
1,100,000
660,000
640,000
130,000
67,000
43,000
31,000
23,000
16,000
11,000
7,300
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%

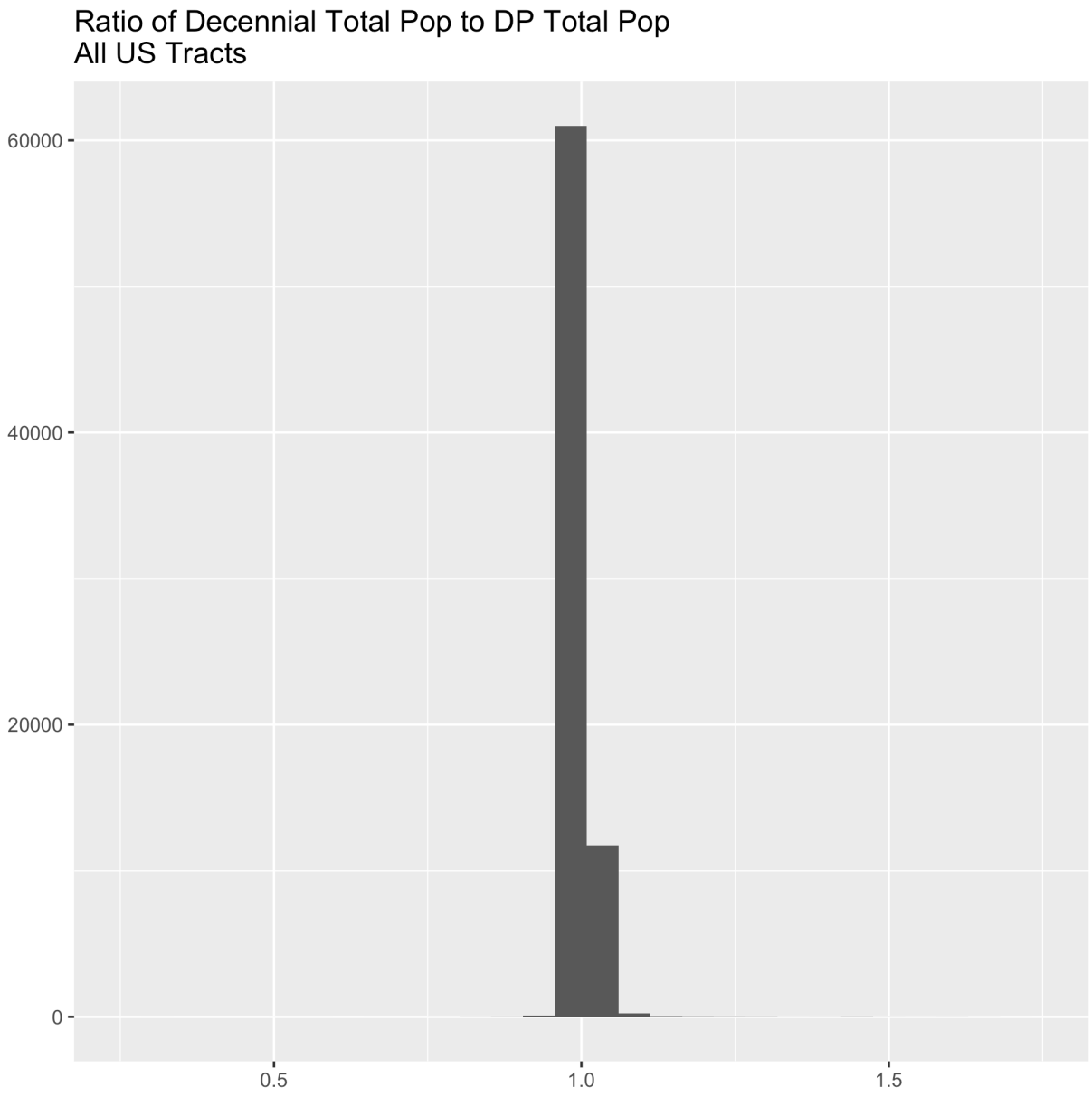

2010 SF1 vs. Demo: Black Householders
States
Counties
Tracts
Block Groups
County Subs
Places
Population Decile
10
9
8
7
6
5
4
3
2
1
23,000,000
11,000,000
7,300,000
6,000,000
4,800,000
3,600,000
2,800,000
1,700,000
1,100,000
660,000
640,000
130,000
67,000
43,000
31,000
23,000
16,000
11,000
7,300
2,900
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%

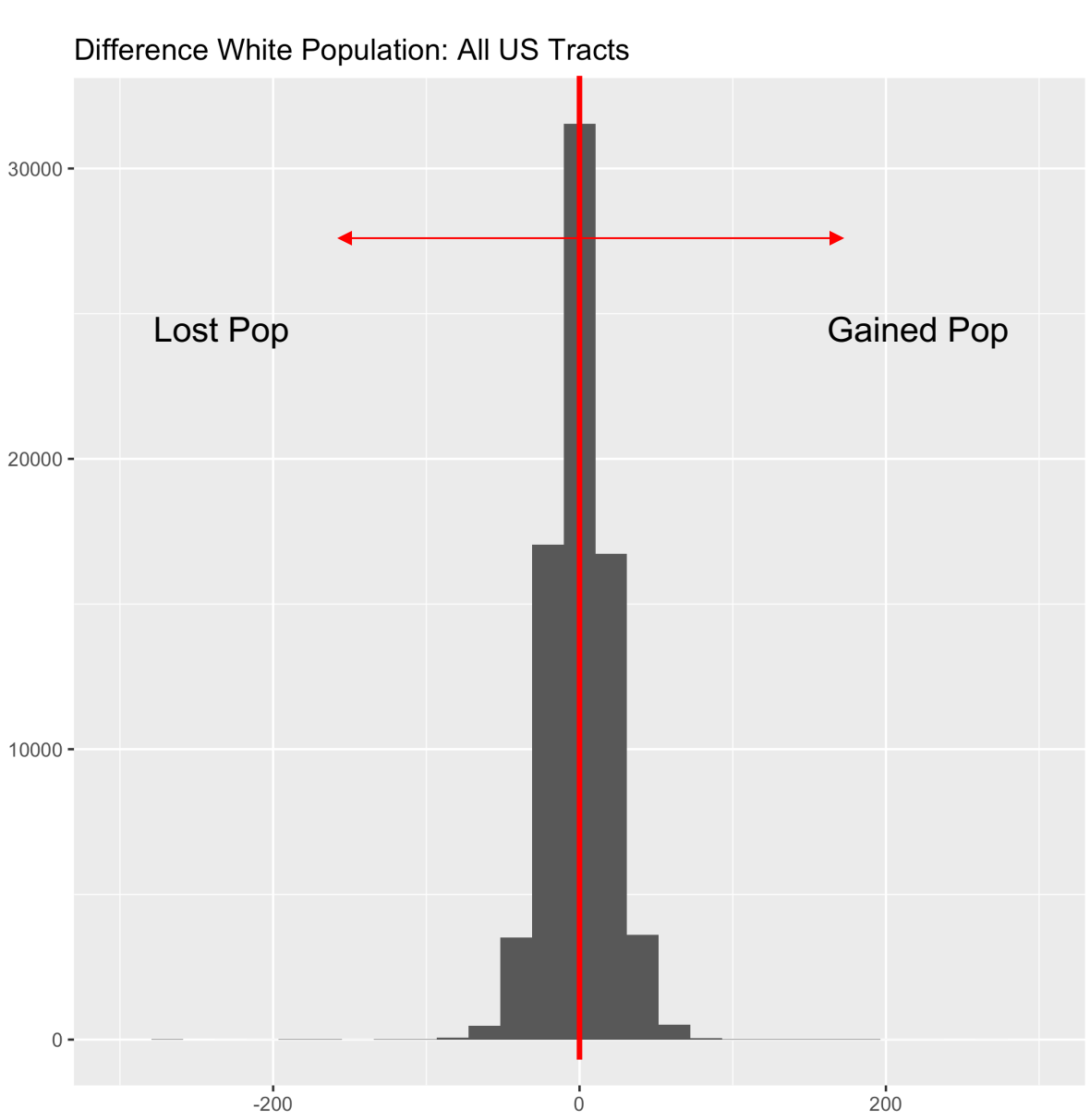
Summary: Changes by Geographic Resolution

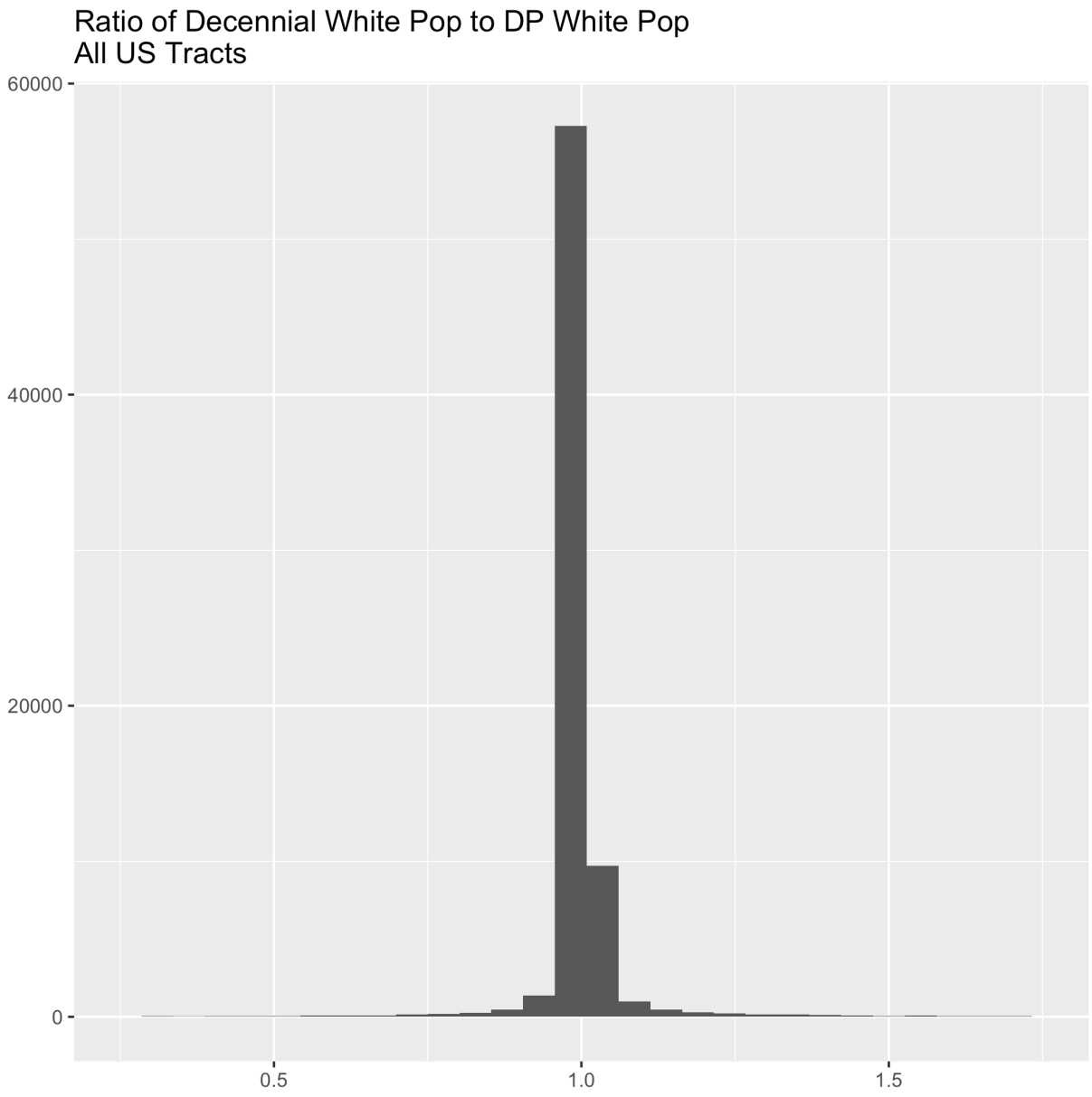
- Off-spine geographies are generally worse than on-spine geographies.
- Impact varies by place, population sub-group.

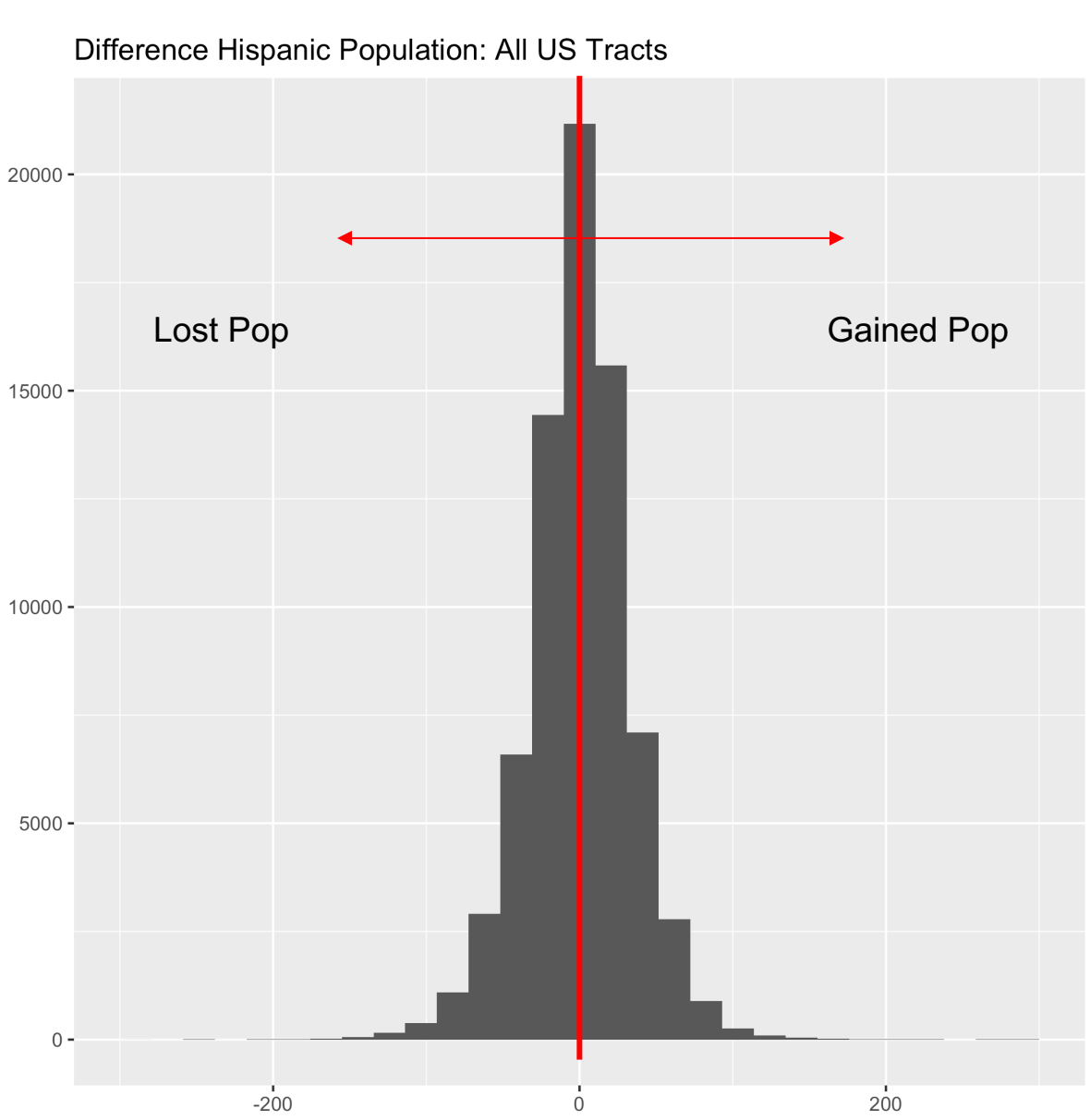
Spatial Patterns In Changes

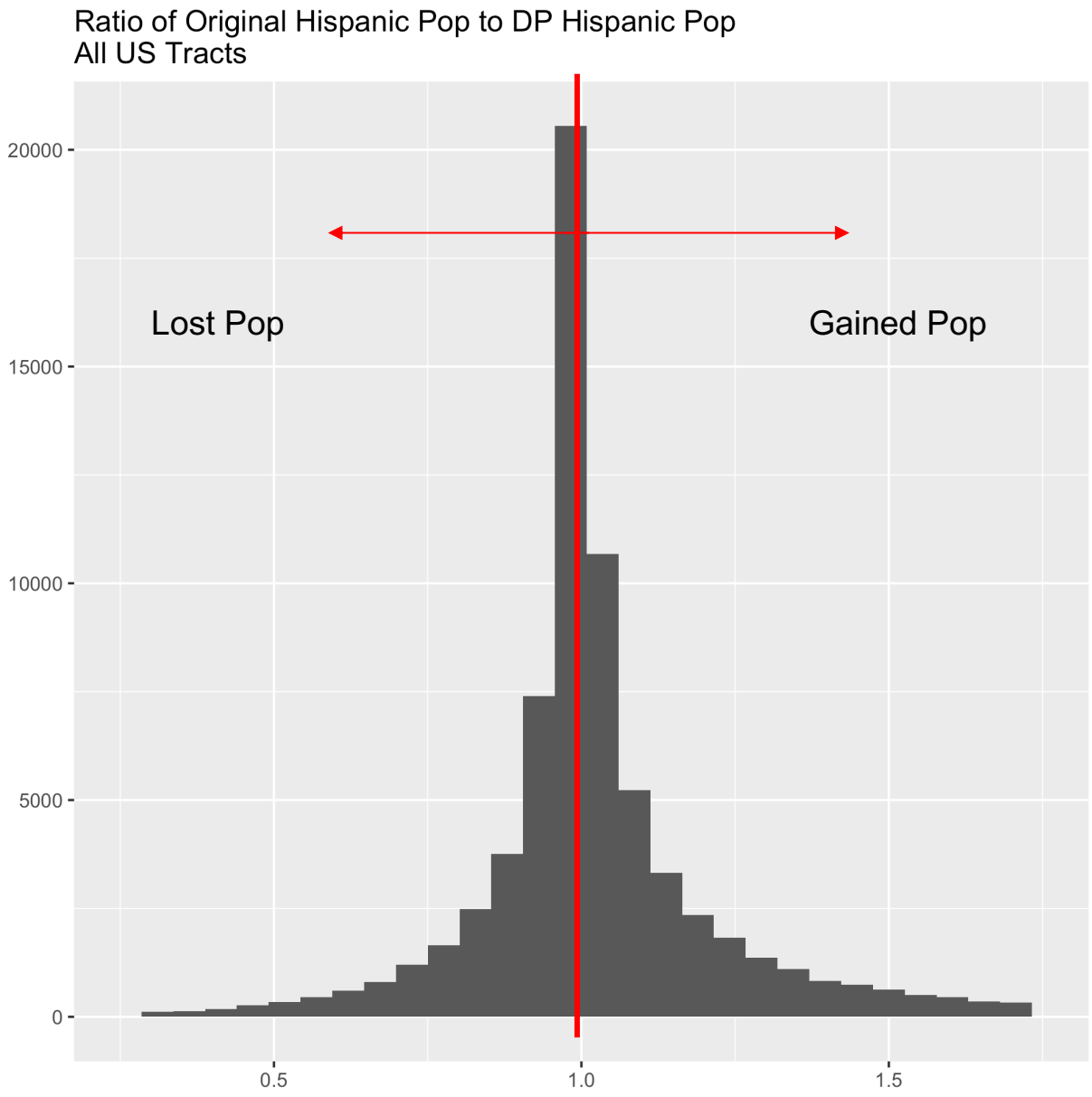

Difference Total Population: All US Tracts
20000
15000
10000
5000
0
Lost Pop
Gained Pop
-200
0
200


Ratio of Decennial Total Pop to DP Total Pop
All US Tracts
60000
40000
20000
0
0.5
1.0
1.5


Difference White Population: All US Tracts
30000
20000
10000
0
Lost Pop
Gained Pop
-200
0
200

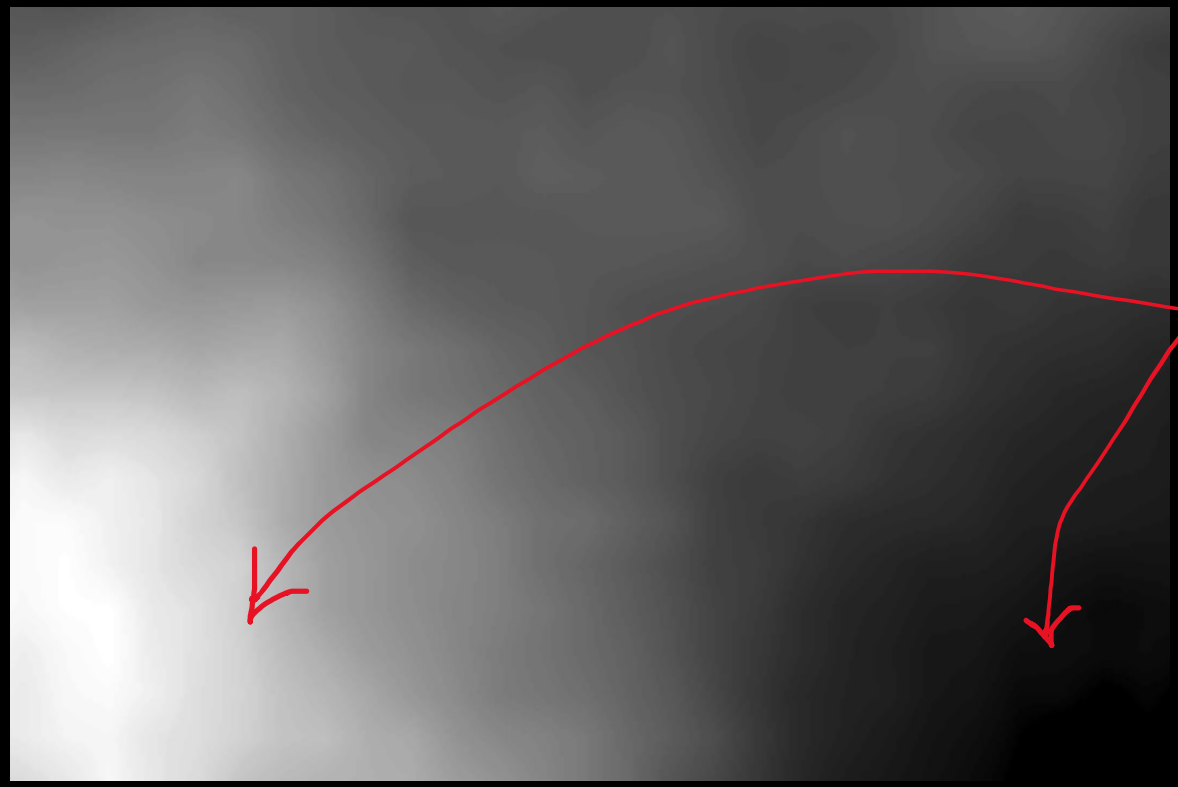

Ratio of Decennial White Pop to DP White Pop
All US Tracts
60000
40000
20000
0
0.5
1.0
1.5

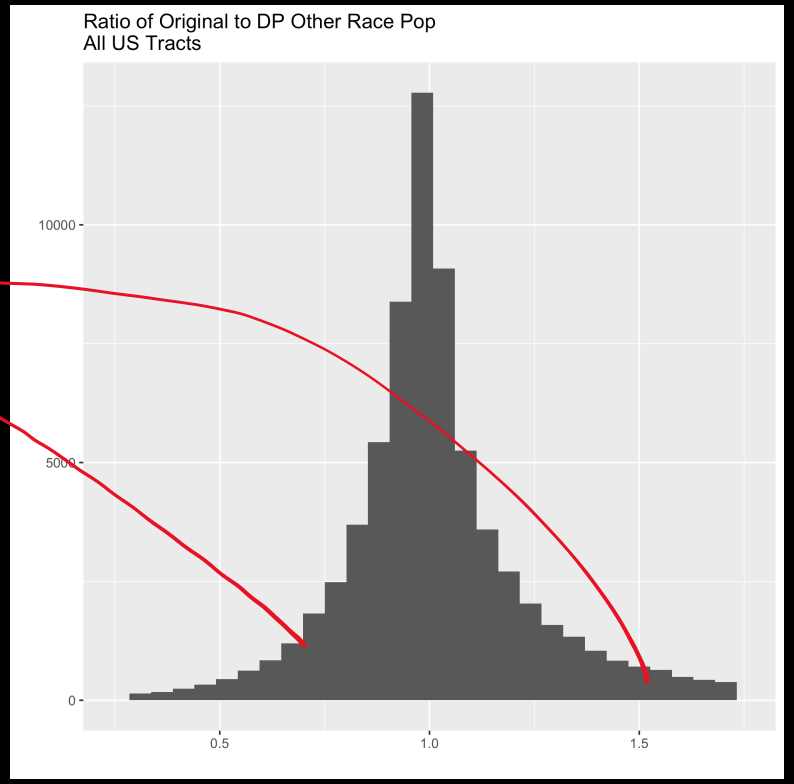

Difference Hispanic Population: All US Tracts
20000
15000
10000
5000
0
Lost Pop
Gained Pop
-200
0
200


Ratio of Original Hispanic Pop to DP Hispanic Pop
All US Tracts
20000
15000
10000
5000
0
Lost Pop
Gained Pop
0.5
1.0
1.5

Spatial Patterns

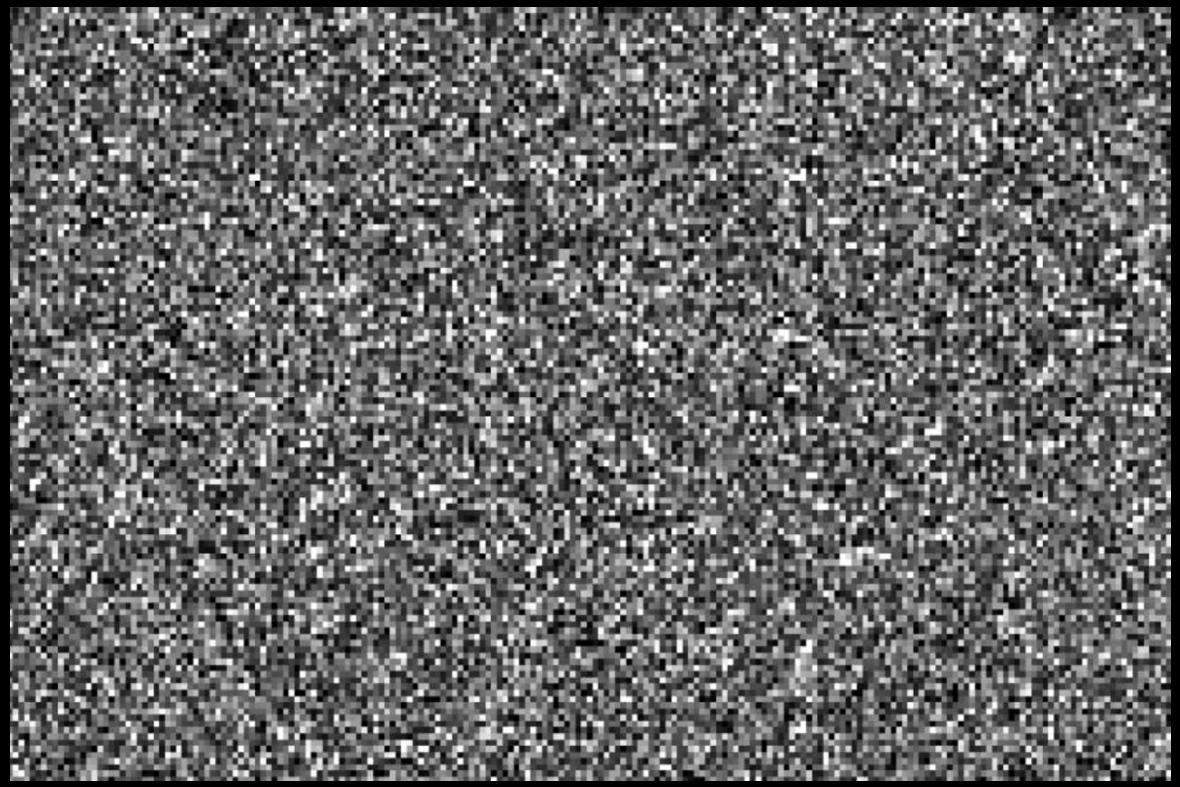
We can have a balanced histogram biased pattern

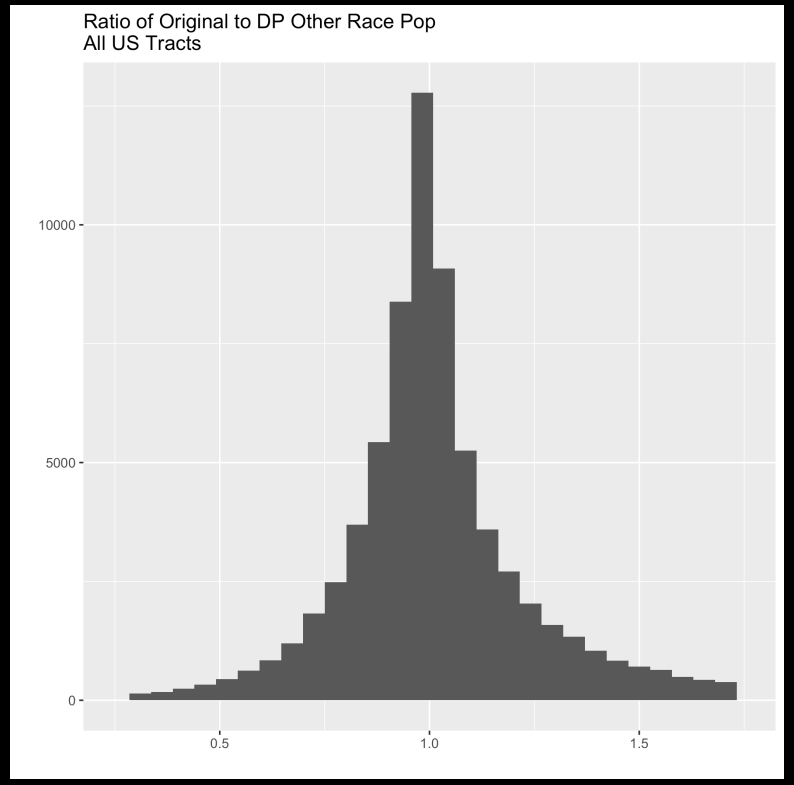




Spatial Patterns

We can have a balanced histogram no pattern





White Population diff
-231.0000 - -16.0000
-16.0000 - -4.0000
-4.0000 - 4.0000
4.0000 - 15.0000
15.0000 - 274.0000

Spatial Patterns: Moran's I

Are the changes randomly distributed in space?

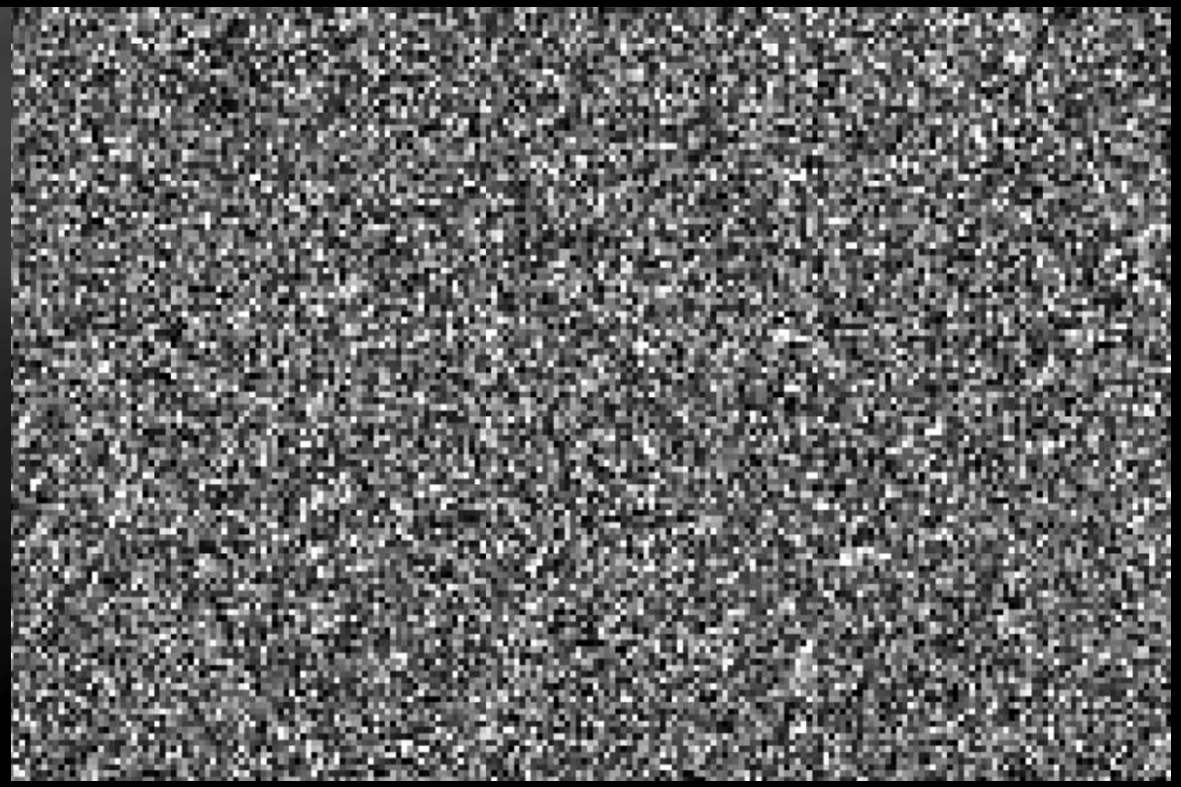
Are particular parts of the map experiencing more infused noise than others?

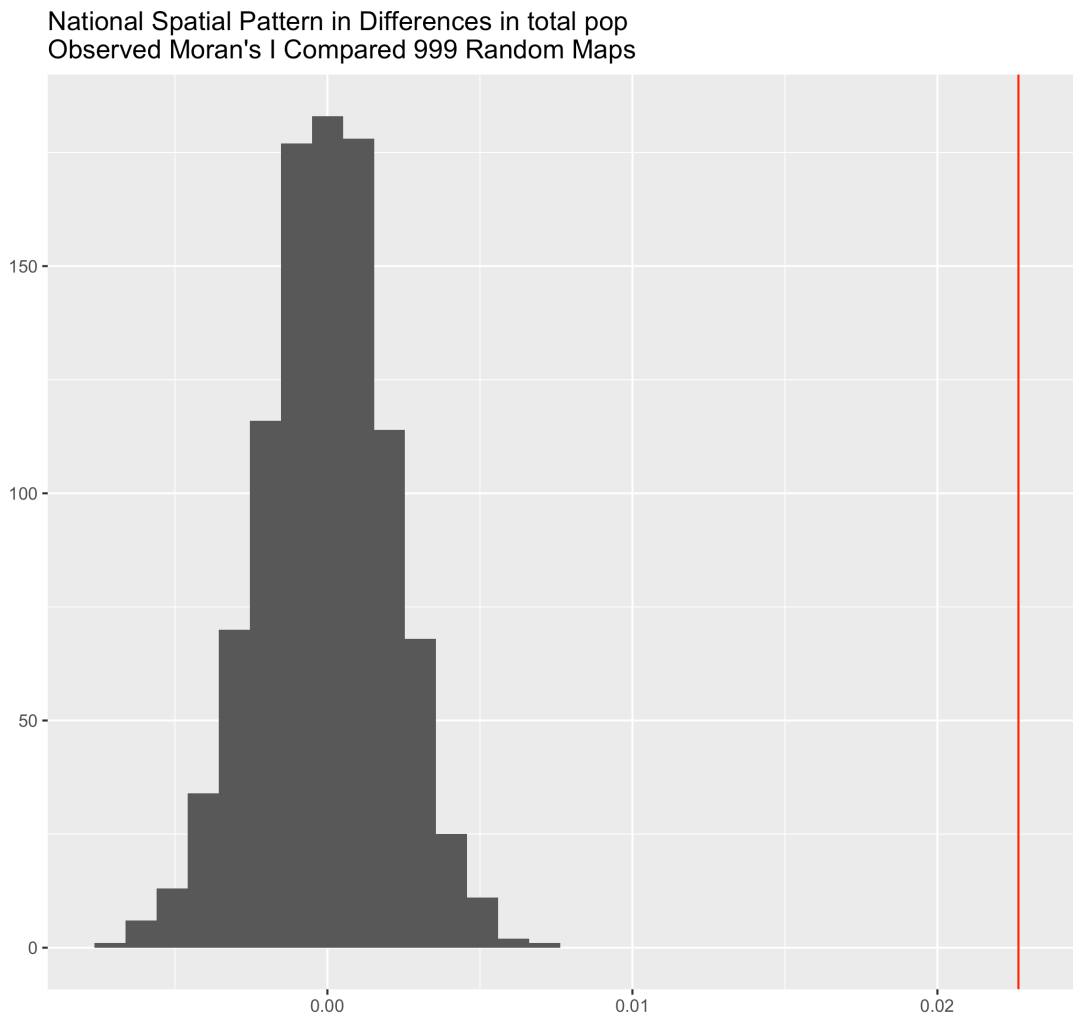
We investigate this using a statistic called Moran's I.
It is simply the correlation between the value observed for each tract and its neighbors.
In this case the value of interest is the difference between the DP and the original population estimate.
We expect these differences to be spatially random…

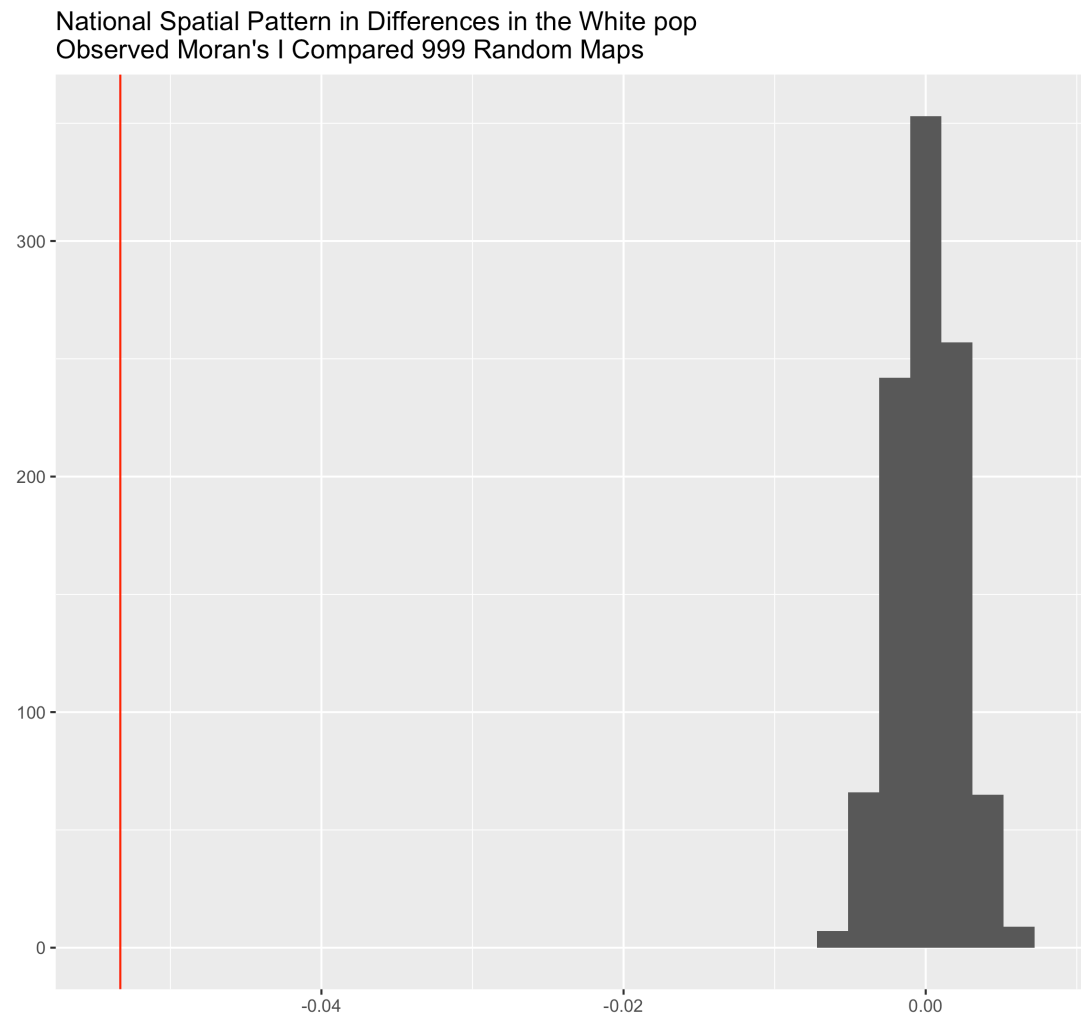
Spatial Patterns: Moran's I

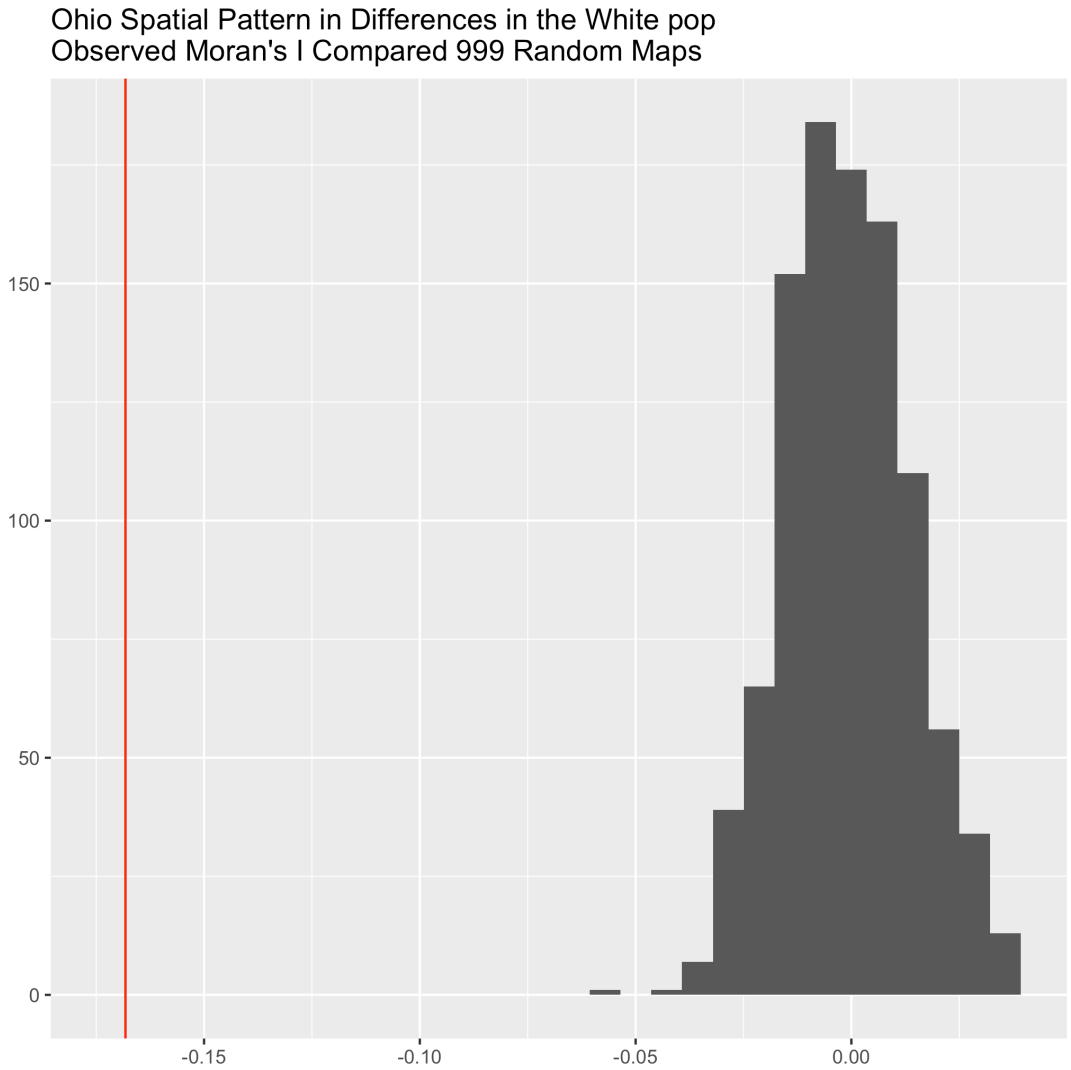
Spatial Pattern
High Moran’s I

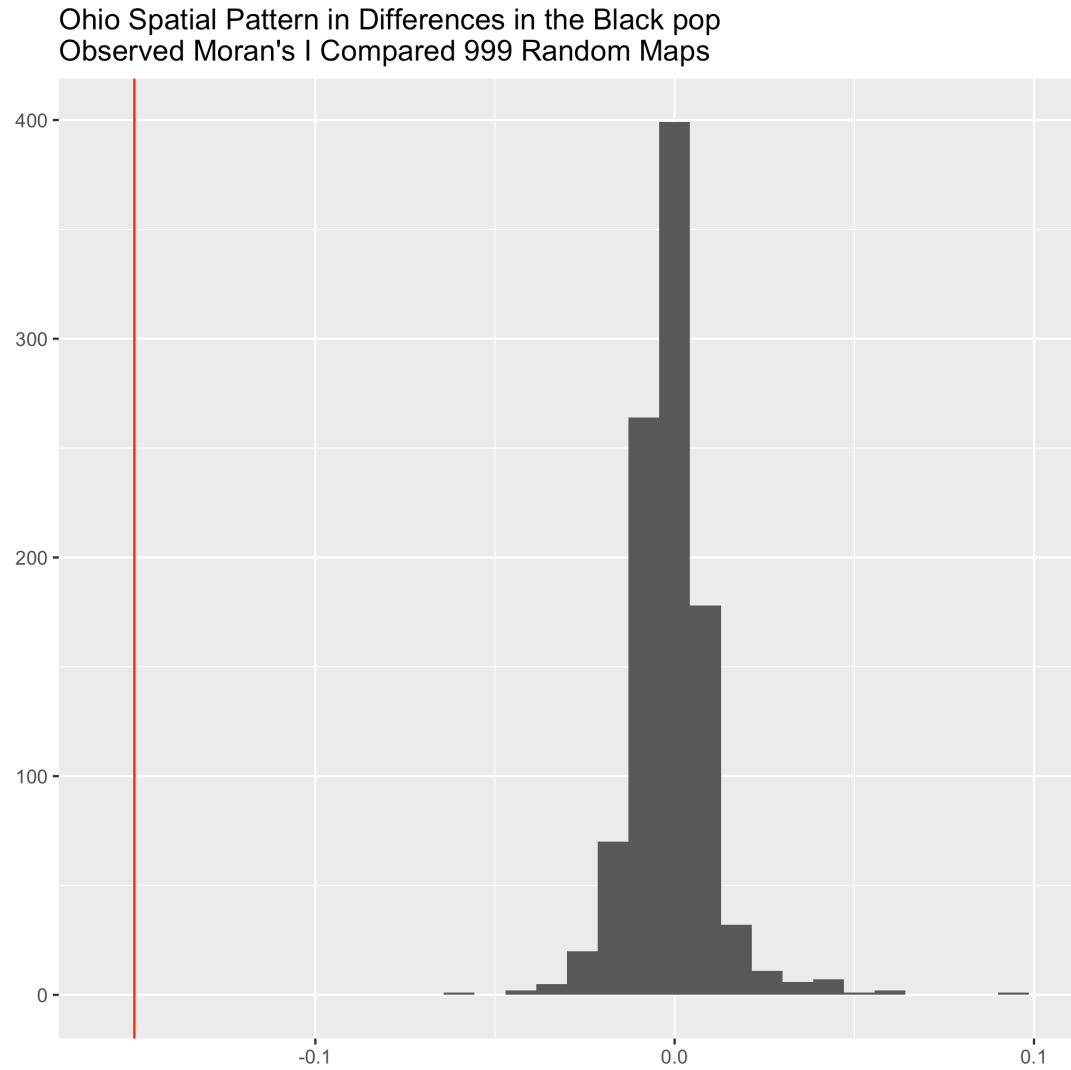
Spatial Randomness




National Spatial Pattern in Differences in total pop
Observed Moran's I Compared 999 Random Maps
150
100
50
0
0.00
0.01
0.02


National Spatial Pattern in Differences in the White pop
Observed Moran's I Compared 999 Random Maps
300
200
100
0
-0.04
-0.02
0.00


Ohio Spatial Pattern in Differences in the White pop
Observed Moran's I Compared 999 Random Maps
150
100
50
0
-0.15
-0.10
-0.05
0.00


Ohio Spatial Pattern in Differences in the Black pop
Observed Moran's I Compared 999 Random Maps
400
300
200
100
0
-0.1
0.0
0.1

Summary: Spatial Patterns

- Tract level changes seem unbiased and normally distributed.
- We see evidence that the noise infusion is not spatially random.
- This means that for some local communities and data users impacts of differential privacy are more pronounced than we'd expect.
- Needs more investigation.

Process and Products

Process: Questions

- Going forward what is the resolution process? What are the acceptance criteria for the final DP parameters?
- How will the Bureau interact with users? What is the plan for incorporating user feedback into the DP parameters? How does the Bureau plan to education users about these data?

Products: Wish list

- More realizations of 2010 data (or synthetic data)
  - Difficult to make recommendations based on N of 1
  - Lots of people with FSRDC access willing and able to analyze more versions
- Off-spine allocation – importance of administrative units
  - It's possible to trace top-down path through hierarchy and include off-spine levels
- More invariants
  - Block-level total population has always been invariant
  - Empirical analysis of privacy loss when block-level total pop invariant?
- Uncertainty metric(s)
  - If you want users to adopt more robust statistical techniques, then they need these

Summary: Overall

- Aim is to understand the impact of the DP changes on the data we use to study places in the US.
- We examined multiple scales and spatial patterns.
- Our initial evaluations suggest more investigation of the “trade off” is warranted.
- There is a lot more to do.

Summary: Overall

- With this differential privacy proposal there is a tension between important public and private goods. Different people will assign different values to these goods.
- To understand how much of the public good (useable/useful data) we are trading for the private good, we need to evaluate the data.
- We need a participatory decision-making process.

Questions or feedback:

vanriper@umn.edu
seth.spielman@colorado.edu

All code and data:

github.com/geoss/CNSTAT_DIFF_PRIVACY