Analyzing the impact of differential privacy on the accuracy of decennial census data

David Van Riper

vanriper@umn.edu

Brookings Institution
September 26, 2019

Outline

- What is differential privacy?
- Applying differential privacy to data
- Implementing differential privacy for census
- Analyzing impact of differential privacy

WHAT IS DIFFERENTIAL PRIVACY?

Differential privacy is…

- A formal (mathematical) definition of privacy

$$\frac{\Pr[M(D) \in S]}{\Pr[M(D') \in S]} \leq e^{\varepsilon}$$

Differential privacy is not…

- An algorithm for disclosure control

Differential privacy is not…

- An algorithm for disclosure control
- An absolute guarantee against disclosure risk

APPLYING DIFFERENTIAL PRIVACY

“True” microdata

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

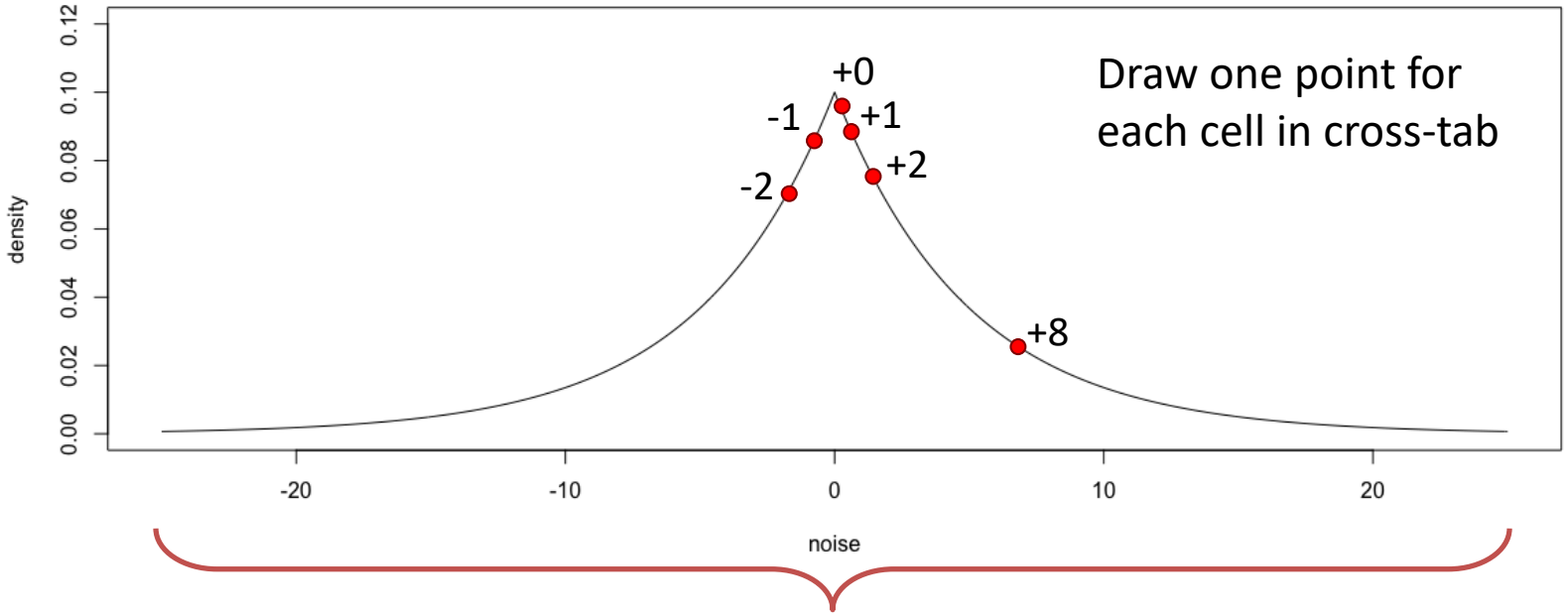
Sex	School
x4 Female	Never
⋮	
Female	Never
x17 Female	Attending
⋮	
Female	Attending
x31 Female	Past
⋮	
Female	Past

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

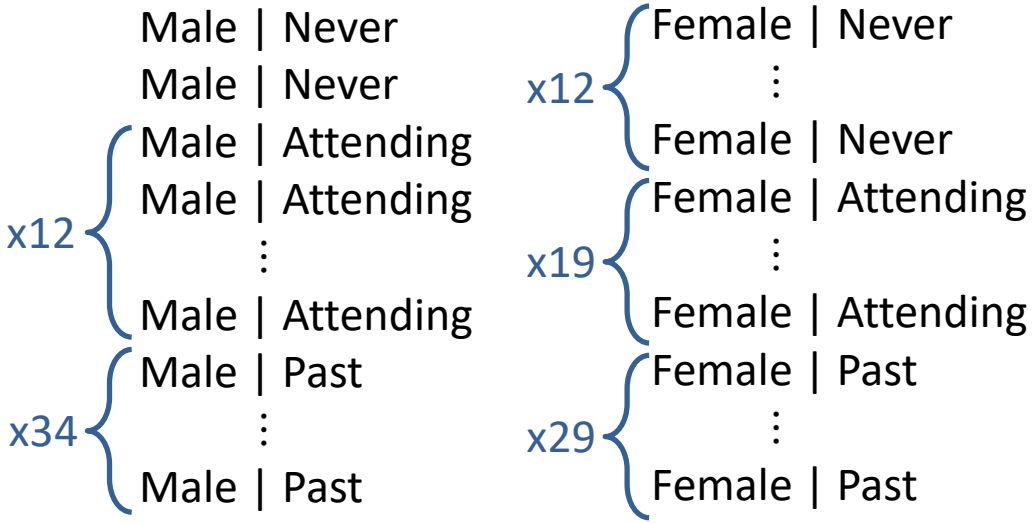


Add noise to cross-tab

	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

Construct synthetic microdata



DIFFERENTIAL PRIVACY AND CENSUS

Differential privacy and census

POLICY DECISIONS

Policy decisions

- Global privacy loss budget ($\boldsymbol{\varepsilon}$)
- Geographic levels
- Tables
- Invariants and constraints

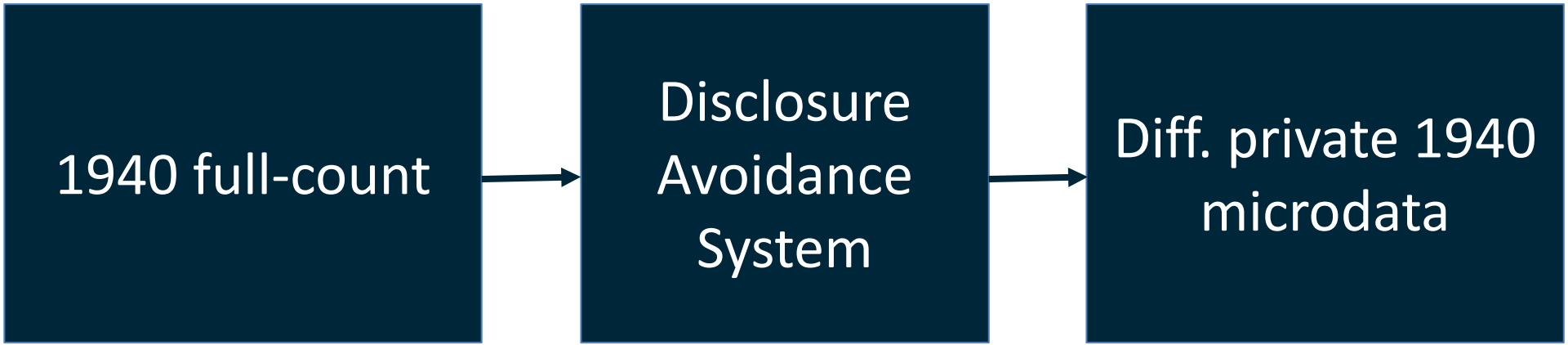
- 1940 Geographic levels
  - Nation
  - State
  - County
  - Enumeration district

- 1940 tables
  - Voting age [2] x Hispanic [2] x Race [6]
  - Households/group quarters type [8]
  - Detailed [192]
    - Voting age [2] x Hispanic [2] x Race [6] x GQ Type [8]

ANALYZING DIFFERENTIALLY PRIVATE 1940 CENSUS DATA

- Census Disclosure Avoidance System (DAS) source code published in April 2019
  - 2020 Census DAS Development Team, 2019

- Census Disclosure Avoidance System (DAS) source code published in April 2019
  - 2020 Census DAS Development Team, 2019
- Implements their TopDown algorithm
  - Abowd et al, 2019


1940 full-count
Disclosure Avoidance System
Diff. private 1940 microdata

Fixed parameters

- Four geographic levels
  - Nation, state, county, enumeration district

Fixed parameters

- Four geographic levels
  - Nation, state, county, enumeration district
- Three queries / tables
  - Voting age – Hispanic – Race
  - Houshold – group quarters
  - Detailed

Modifiable parameters

- Global privacy loss budget (ε)

Modifiable parameters

- Global privacy loss budget (ε)
- Fractional allocation to
  - Geographic levels
  - Tables

Modifiable parameters

- Global privacy loss budget (ε)
- Fractional allocation to
  - Geographic levels
  - Tables
- Number of runs

- Comparisons between ”true” data (IPUMS 1940 complete-count) and differentially private data

- Differences in total population for counties and enumeration districts
- County-level African American population
- ED-level proportion of total population who identify as African American
- Index of dissimilarity (D)
- Multigroup entropy (H)

Key takeaways

- Geographic units with smaller populations are less accurate
- Small sub-populations are less accurate
- Bias for segregation metrics concerning

Differentially private datasets

CENSUS DAS

Parameters

- Global privacy loss budget (ε)
  - 8 values: [0.25, 0.50, 0.75, 1.0, 2.0, 4.0, 6.0, 8.0]
- Four runs for each value of ε

Parameters

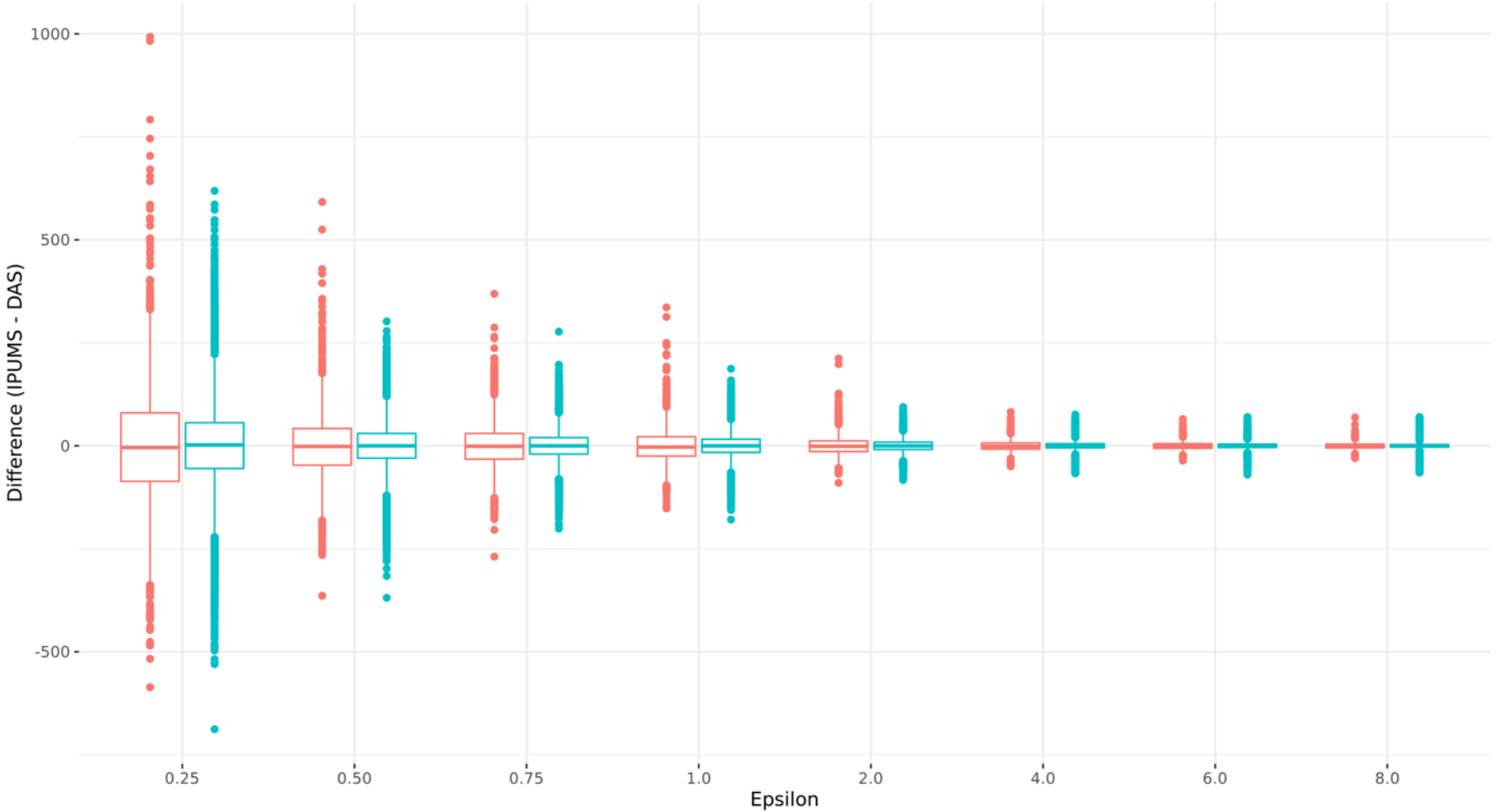
Geographic levels	Fraction
Nation	0.25
State	0.25
County	0.25
Enumeration district	0.25

Parameters

Tables	**Fraction**
Voting age—Hispanic – Race	0.675
Household – Group quarters	0.225
Detailed	0.1

Difference between IPUMS and Census DAS total population counts

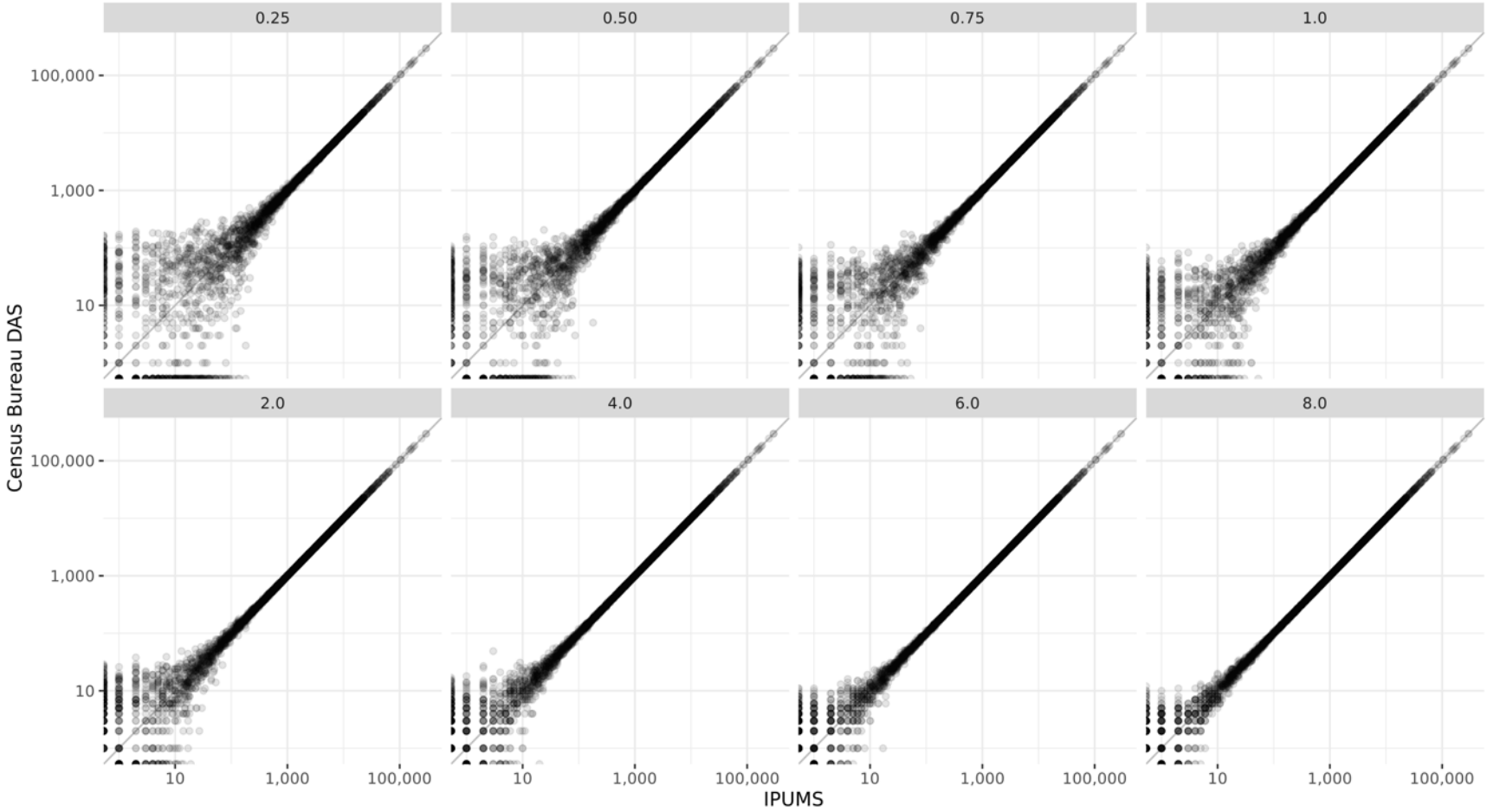
US counties (orange) and EDs (teal)



Source: Ruggles et al. (2018); US Census Bureau (2019)

African American population under different levels of noise injection

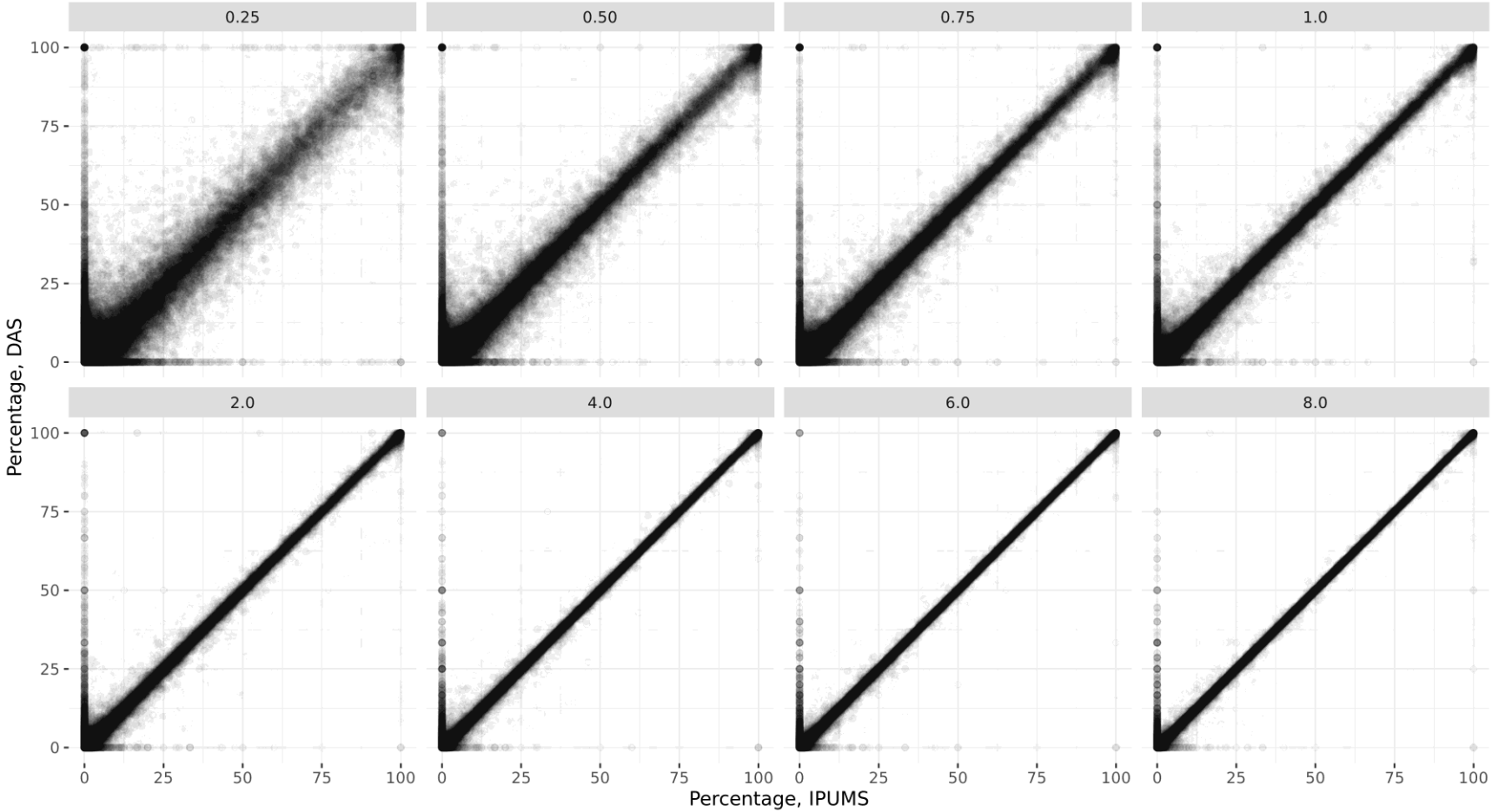
US counties



Source: Ruggles et al. (2018); US Census Bureau (2019)

Percentage of population who is African American

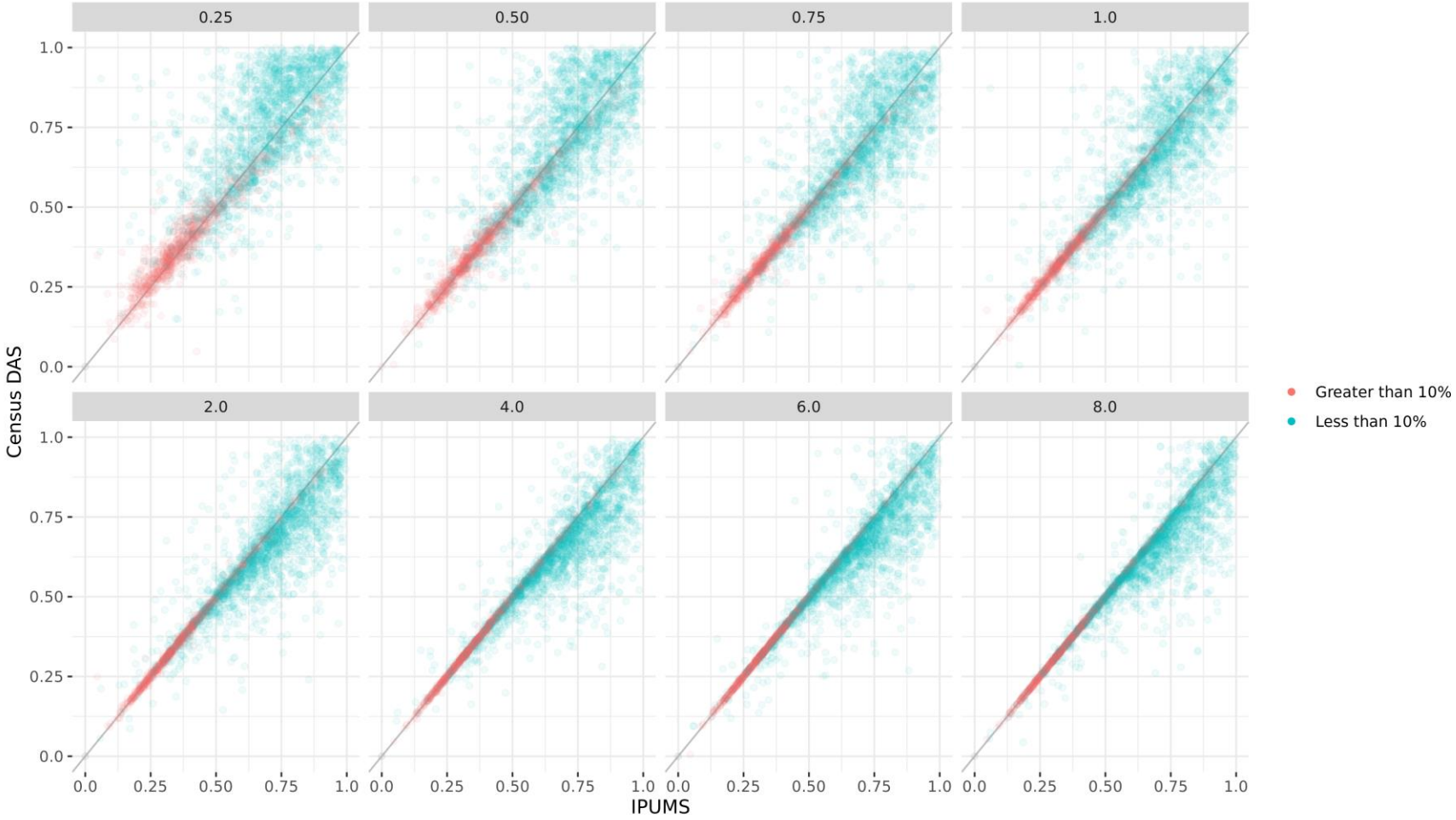
US enumeration districts



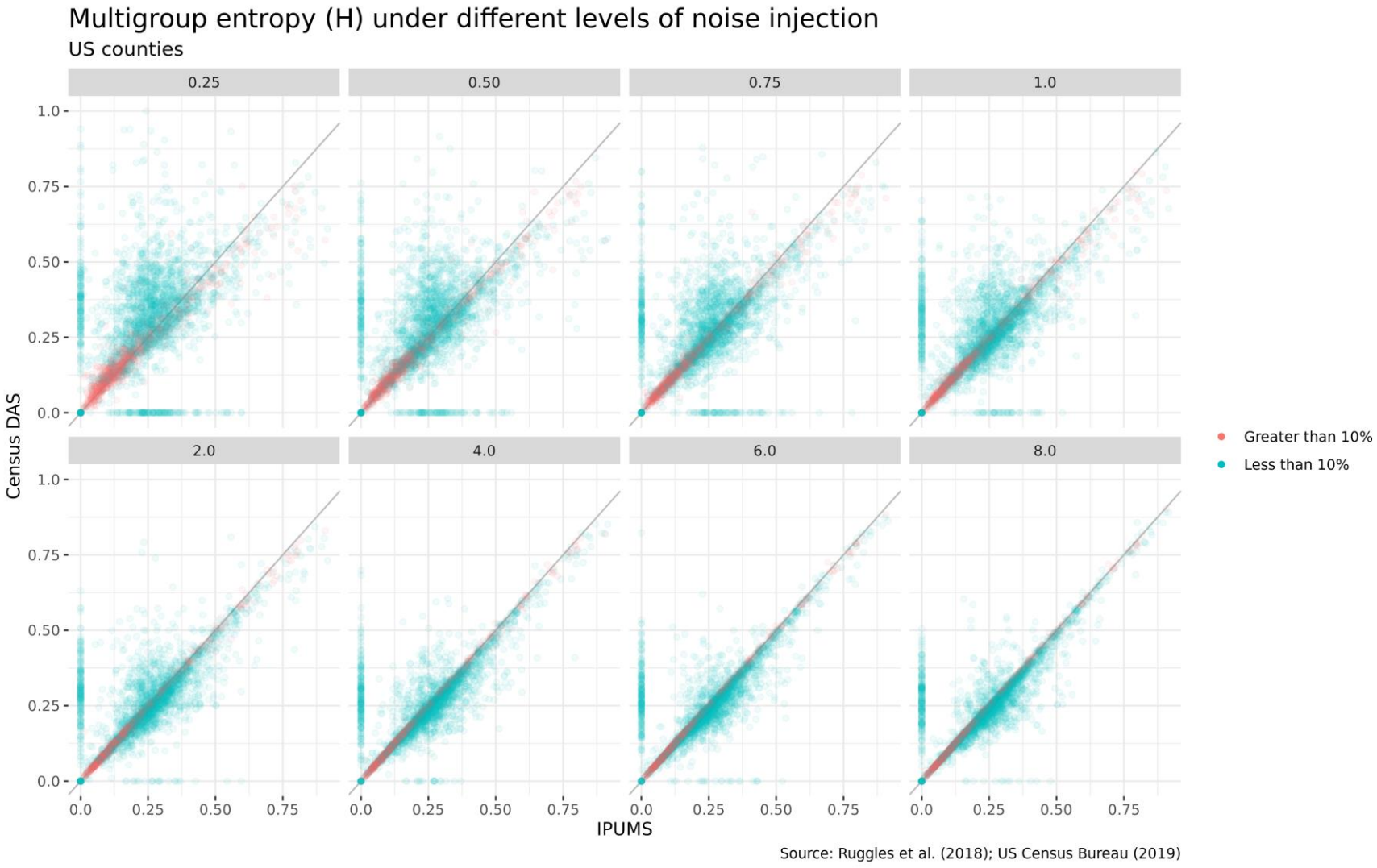
Source: Ruggles et al. (2018); US Census Bureau (2019)

Index of dissimilarity (D) under different levels of noise injection

US counties



Source: Ruggles et al. (2018); US Census Bureau (2019)


Multigroup entropy (H) under different levels of noise injection
US counties
0.25
0.50
0.75
1.0
2.0
4.0
6.0
8.0
Census DAS
IPUMS
0.0
0.25
0.50
0.75
1.0
Greater than 10%
Less than 10%
Source: Ruggles et al. (2018); US Census Bureau (2019)

Differentially private datasets

GEOGRAPHIC LEVELS

Parameters

- Global privacy loss budget (ε)
  - One value: 1.0
- One run

Parameters

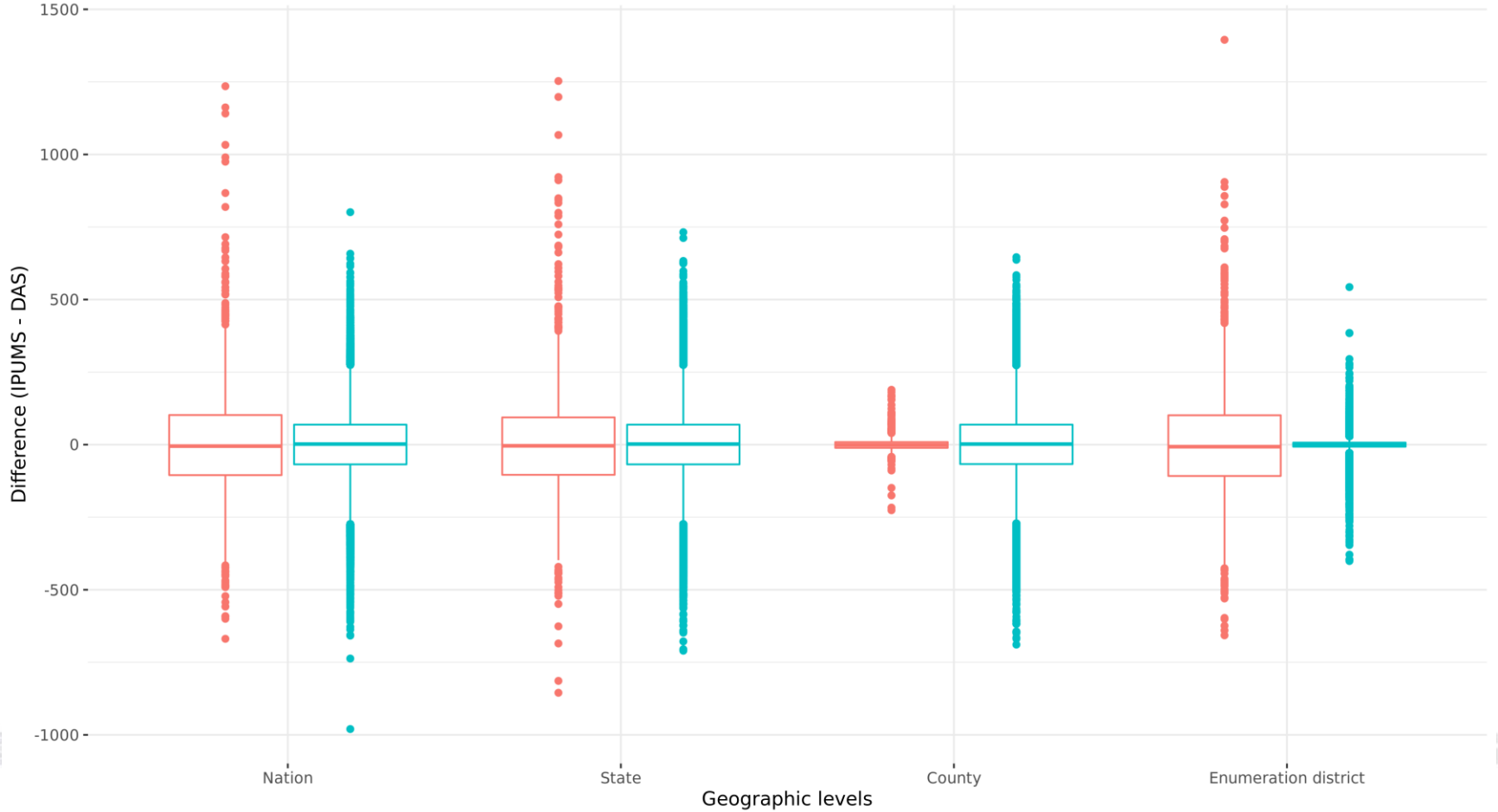
Geographic levels	**Fraction***
Nation	0.85
State	0.05
County	0.05
Enumeration district	0.05

Parameters

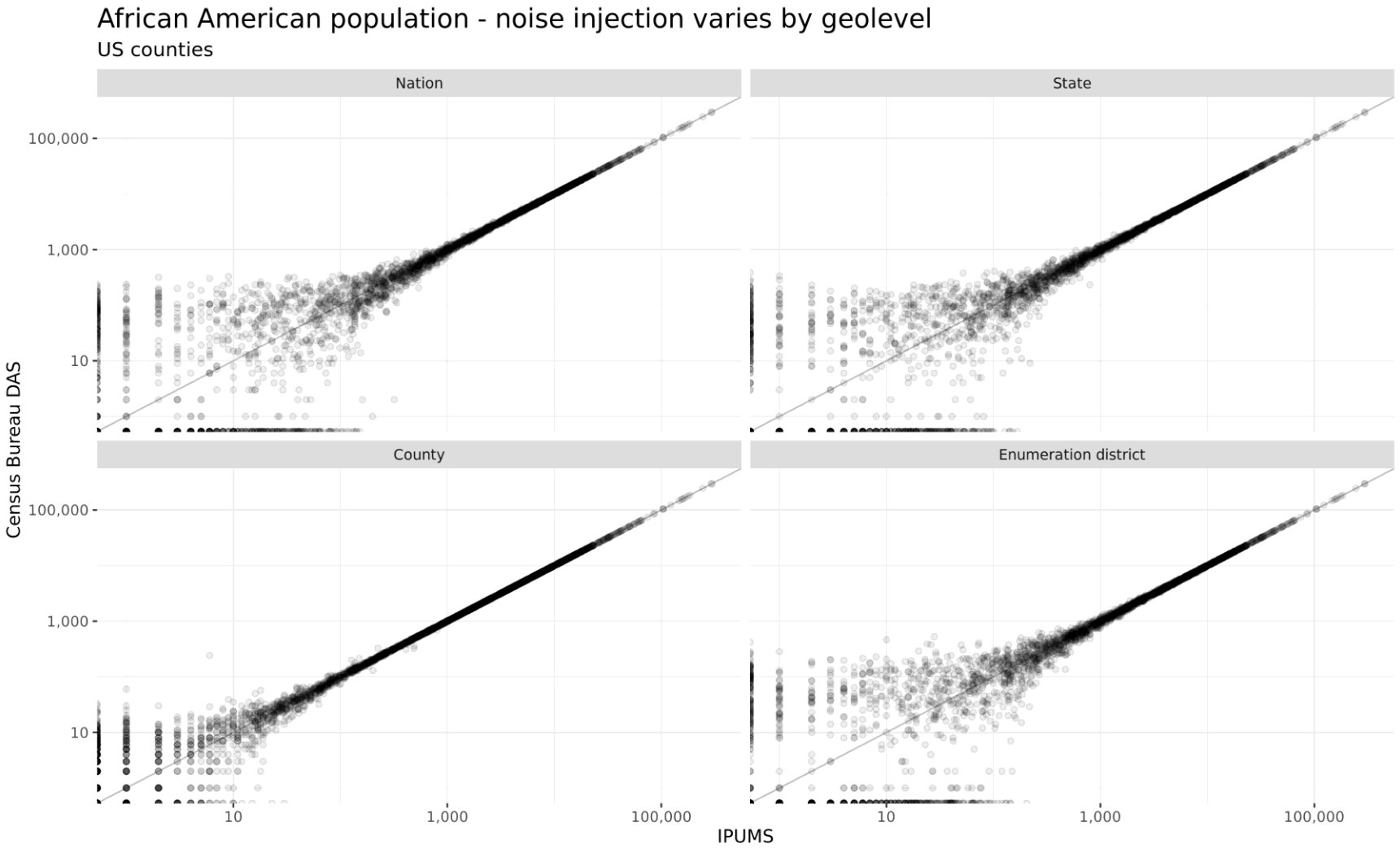
Tables	**Fraction**
Voting age – Hispanic – Race	0.675
Household – Group quarters	0.225
Detailed	0.1

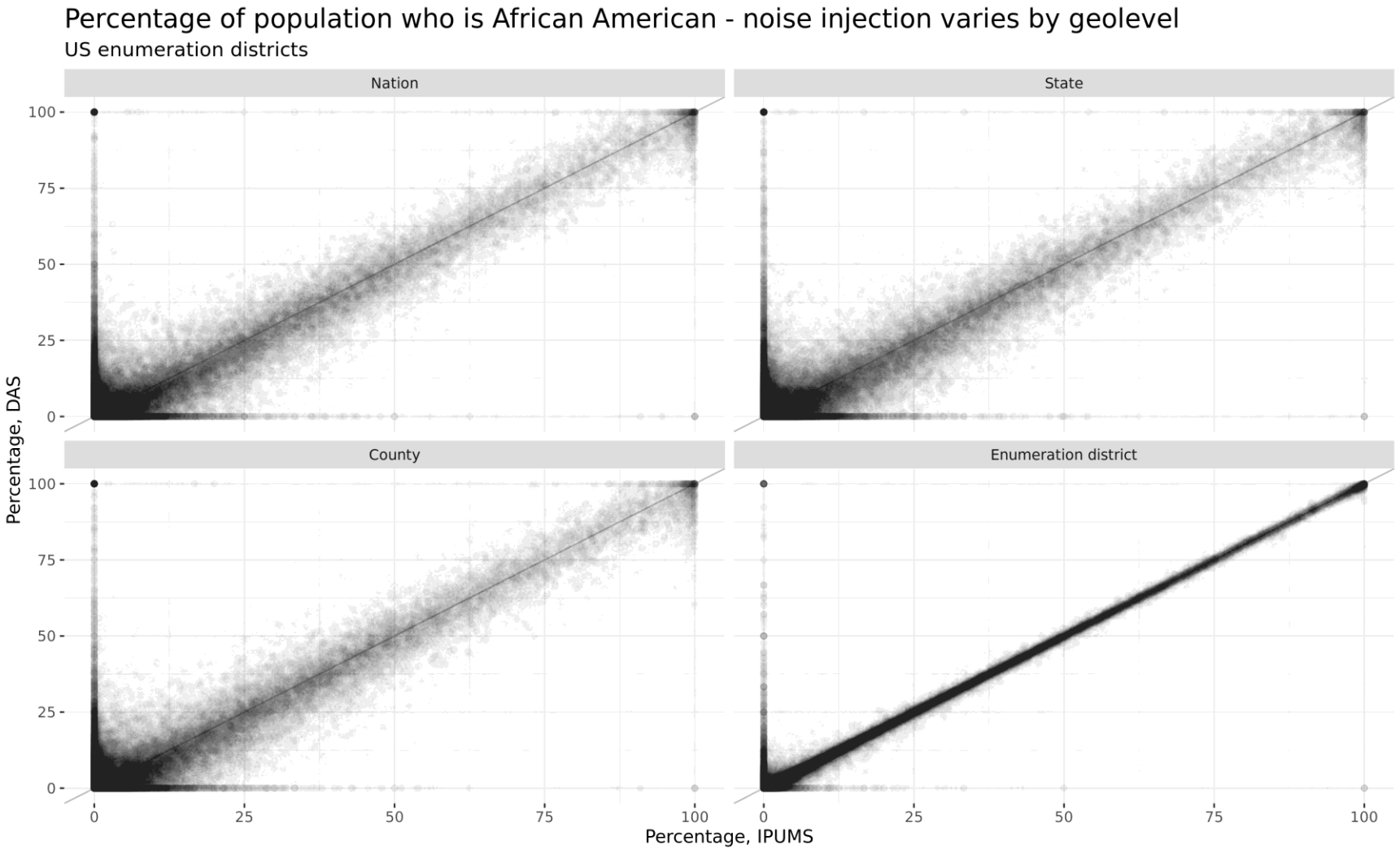
Difference between IPUMS and Census DAS total population counts

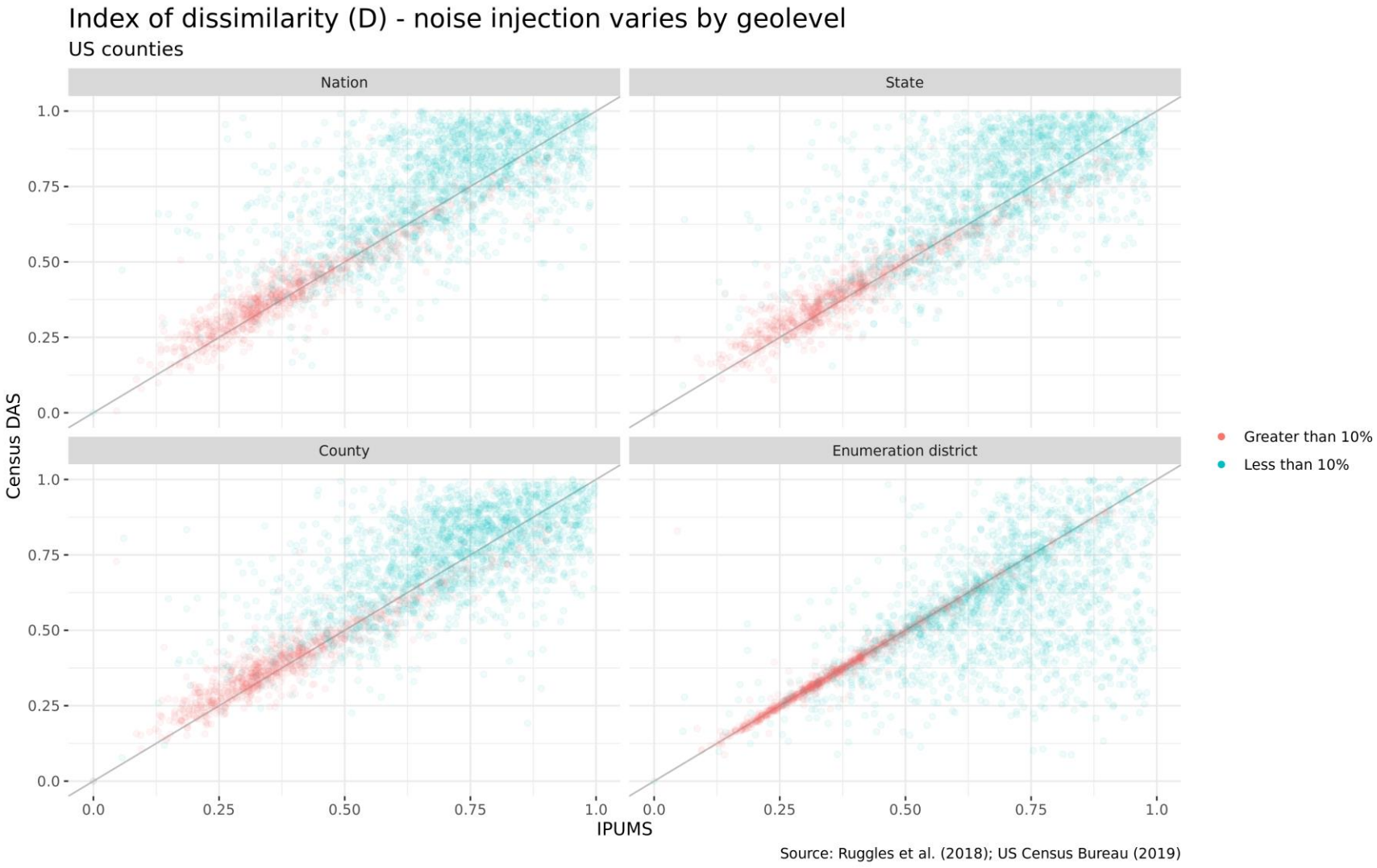
US counties (orange) and EDs (teal)

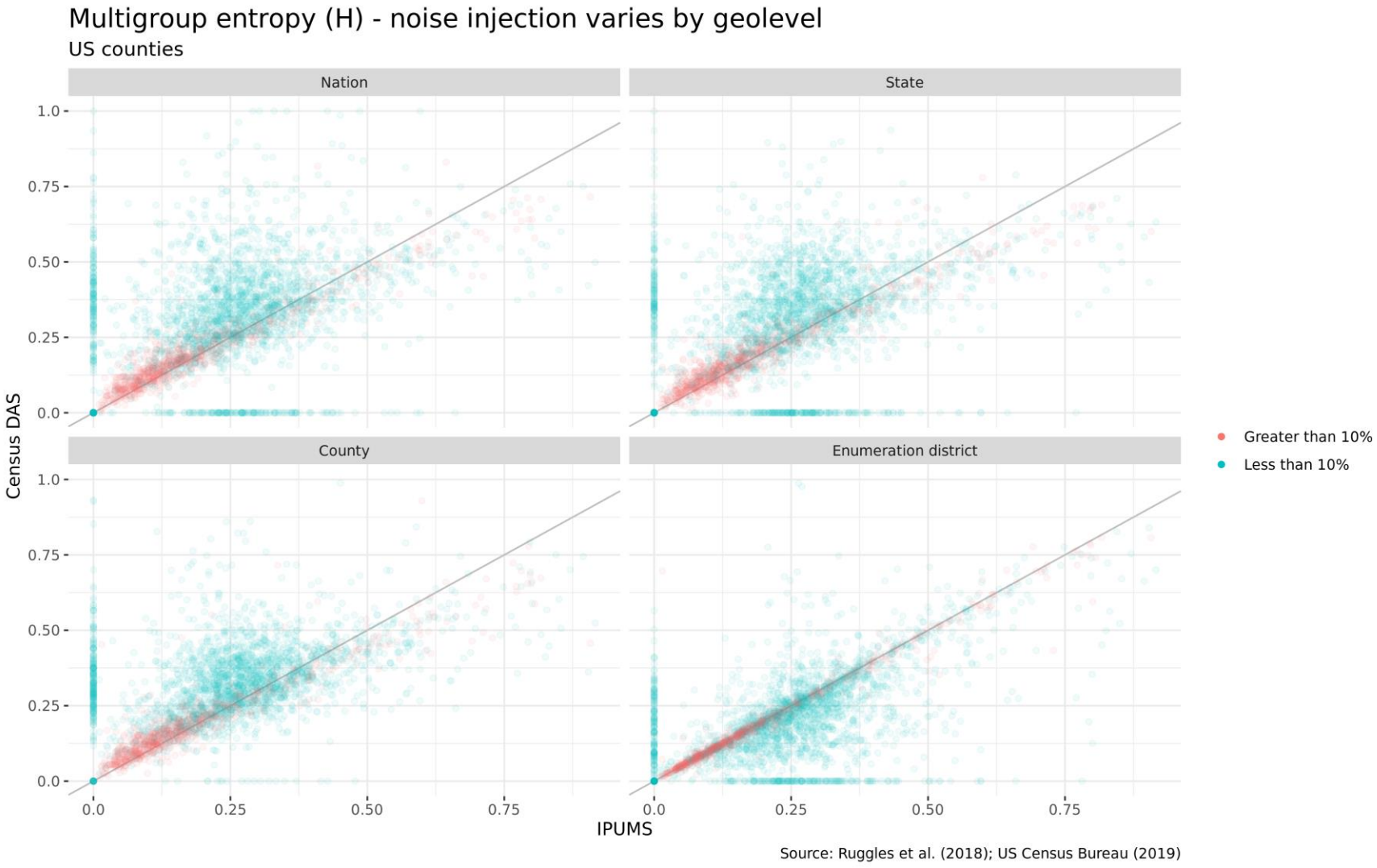


Source: Ruggles et al. (2018); US Census Bureau (2019)


African American population - noise injection varies by geolevel
US counties
Nation
State
County
Enumeration district
Census Bureau DAS
IPUMS
100,000
1,000
10
10
1,000
100,000
Source: Ruggles et al. (2018); US Census Bureau (2019)


Percentage of population who is African American - noise injection varies by geolevel
US enumeration districts
Nation
State
County
Enumeration district
Percentage, DAS
Percentage, IPUMS
0
25
50
75
100
Source: Ruggles et al. (2018); US Census Bureau (2019)


Index of dissimilarity (D) - noise injection varies by geolevel
US counties
Nation
State
County
Enumeration district
Census DAS
IPUMS
0.0
0.25
0.50
0.75
1.0
Greater than 10%
Less than 10%
Source: Ruggles et al. (2018); US Census Bureau (2019)


Multigroup entropy (H) - noise injection varies by geolevel
US counties
Nation
State
County
Enumeration district
Census DAS
IPUMS
0.0
0.25
0.50
0.75
1.0
Greater than 10%
Less than 10%
Source: Ruggles et al. (2018); US Census Bureau (2019)

Differentially private datasets

TABLES

Parameters

- Global privacy loss budget (ε)
  - One value: 1.0
- One run

Parameters

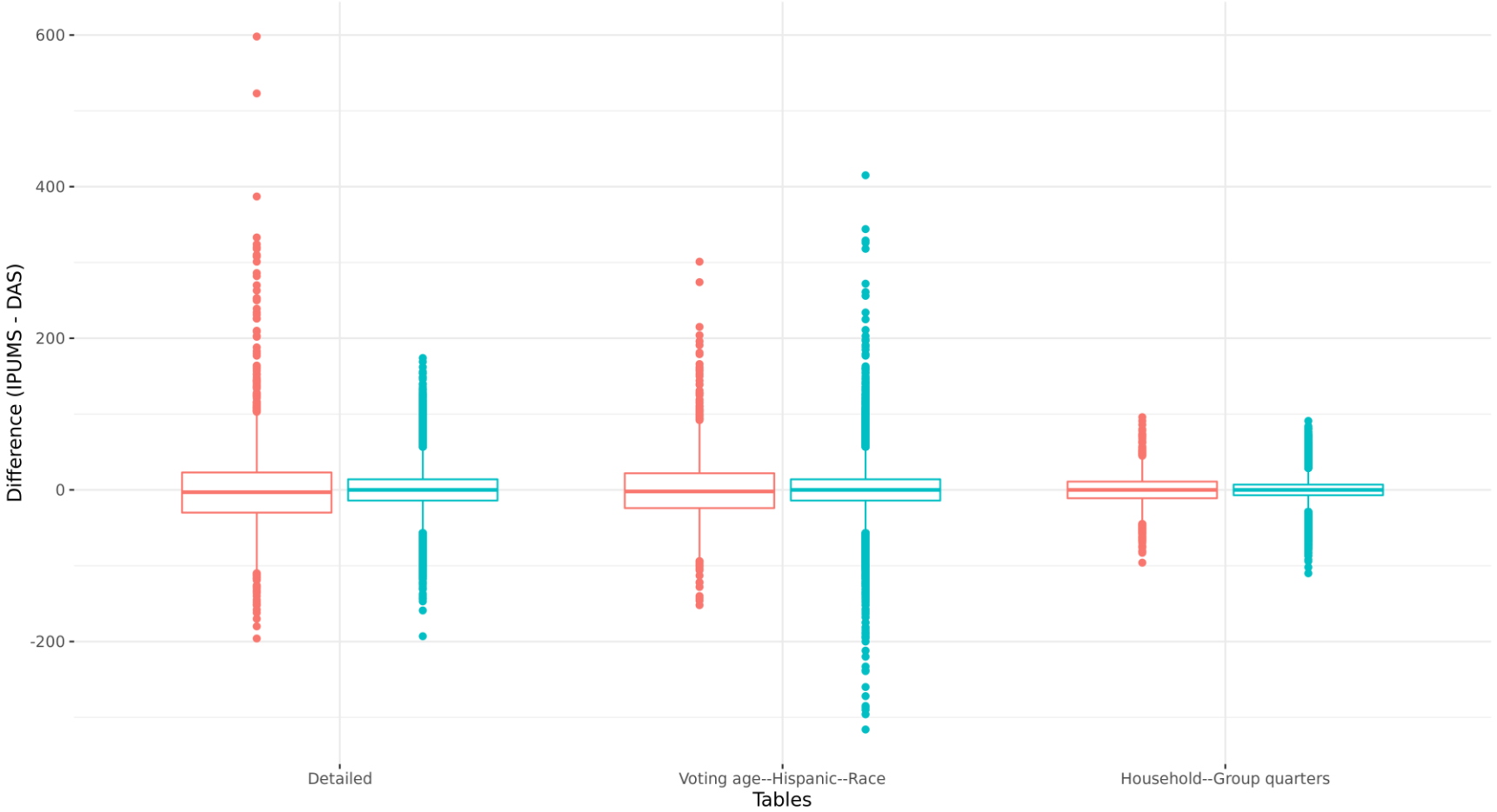
Geographic levels	Fraction
Nation	0.25
State	0.25
County	0.25
Enumeration district	0.25

Parameters

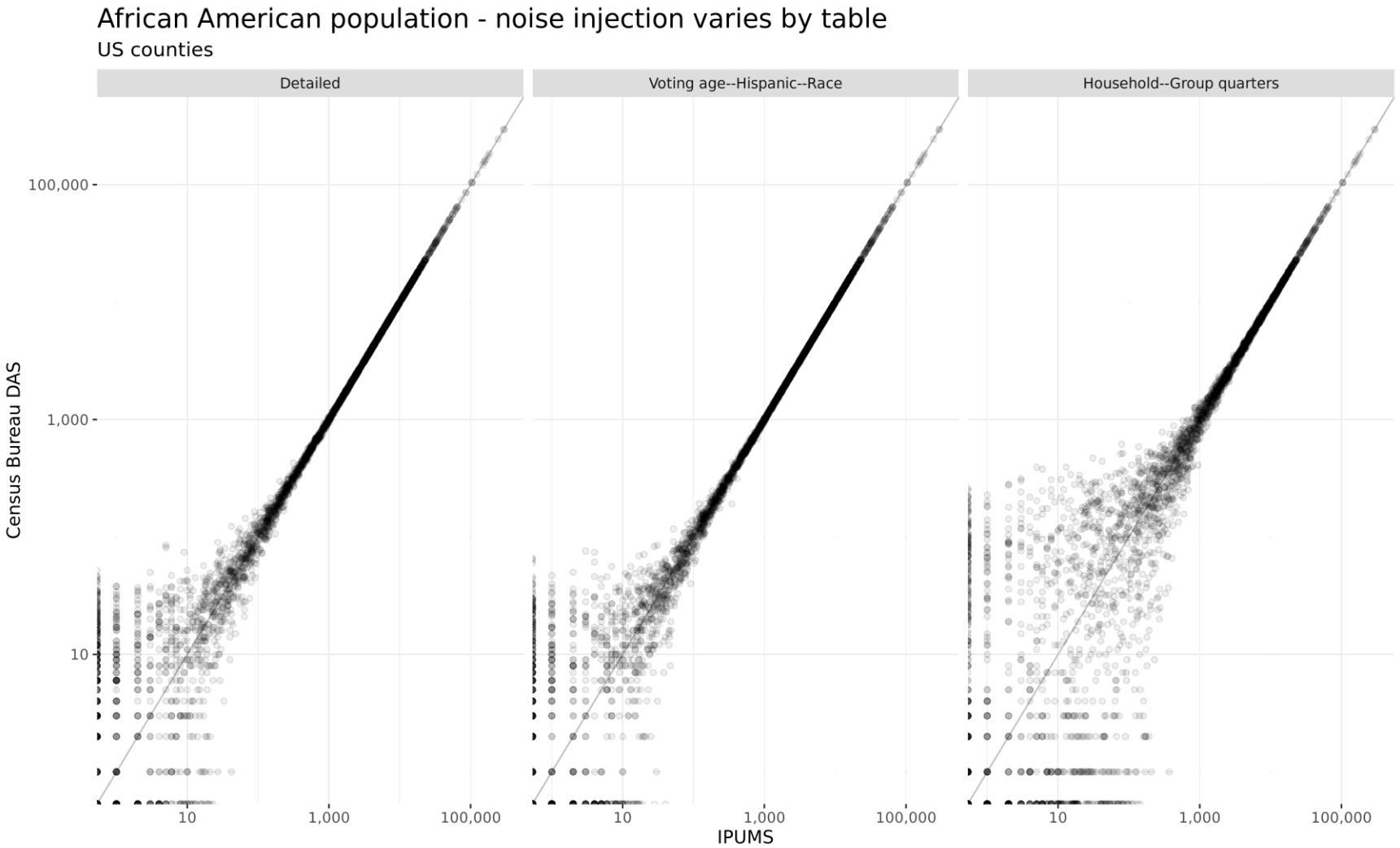
Tables	**Fraction***
Voting age – Hispanic – Race	0.9
Household – Group quarters	0.05
Detailed	0.05

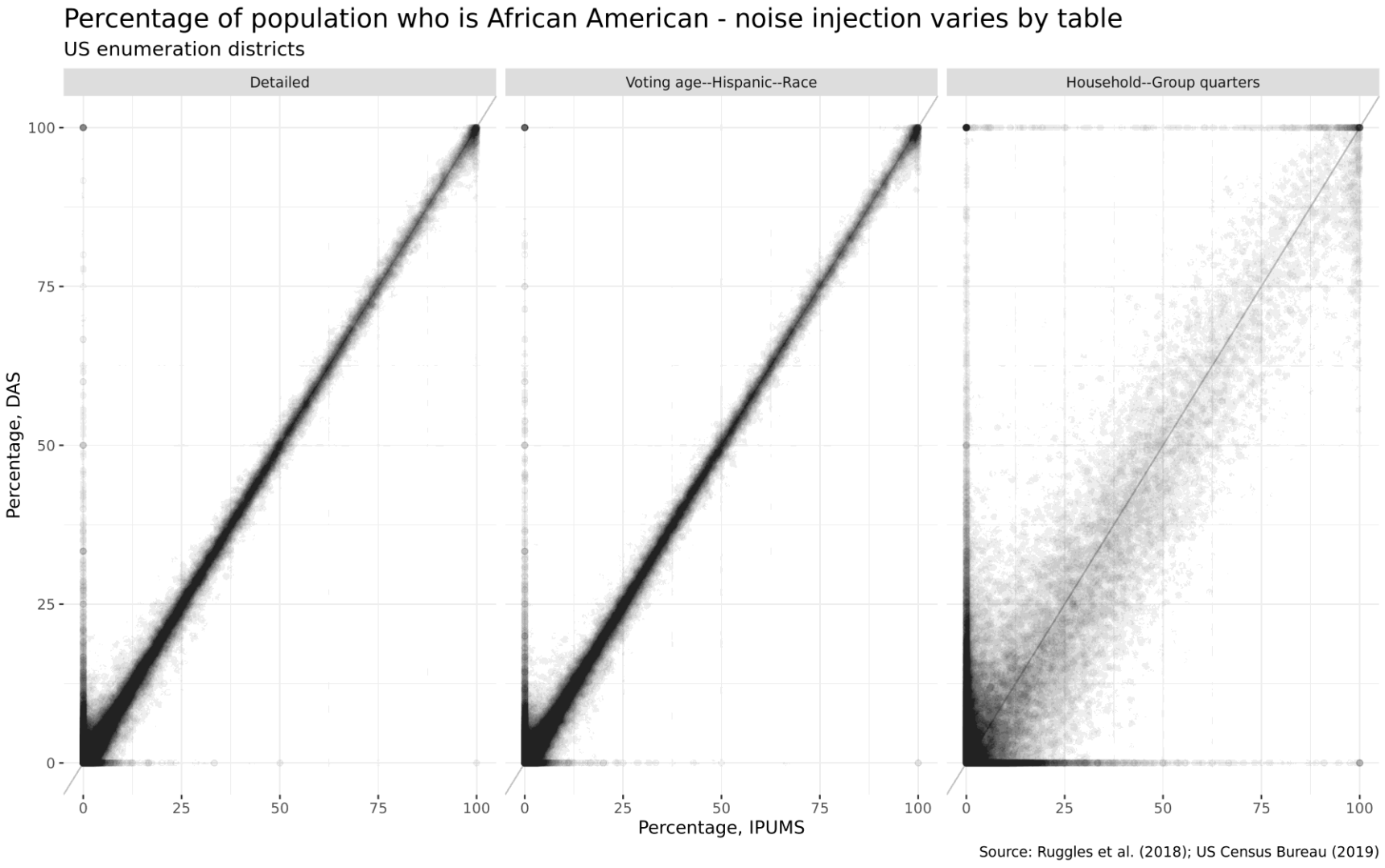
Difference between IPUMS and Census DAS total population counts

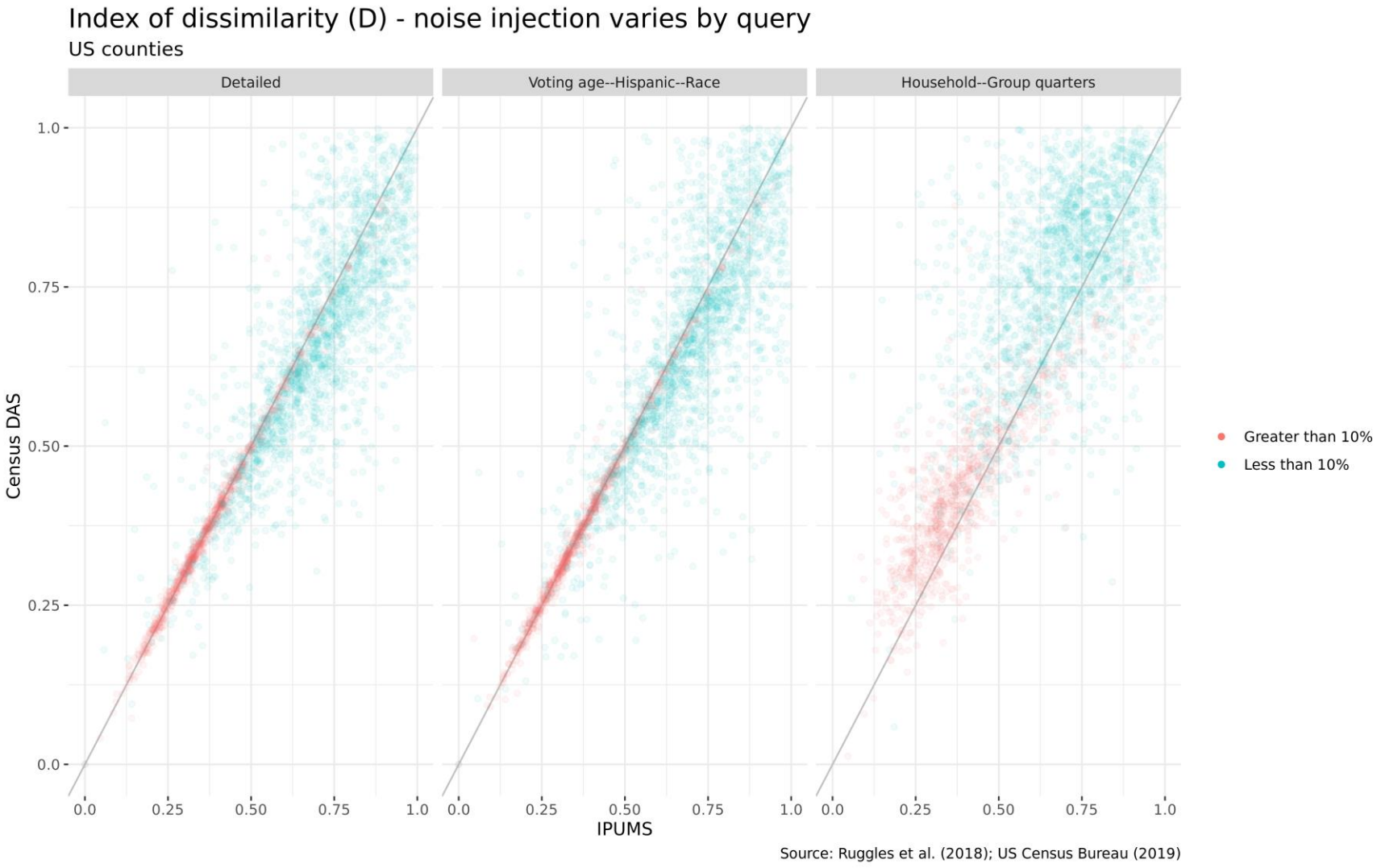
US counties (orange) and EDs (teal)

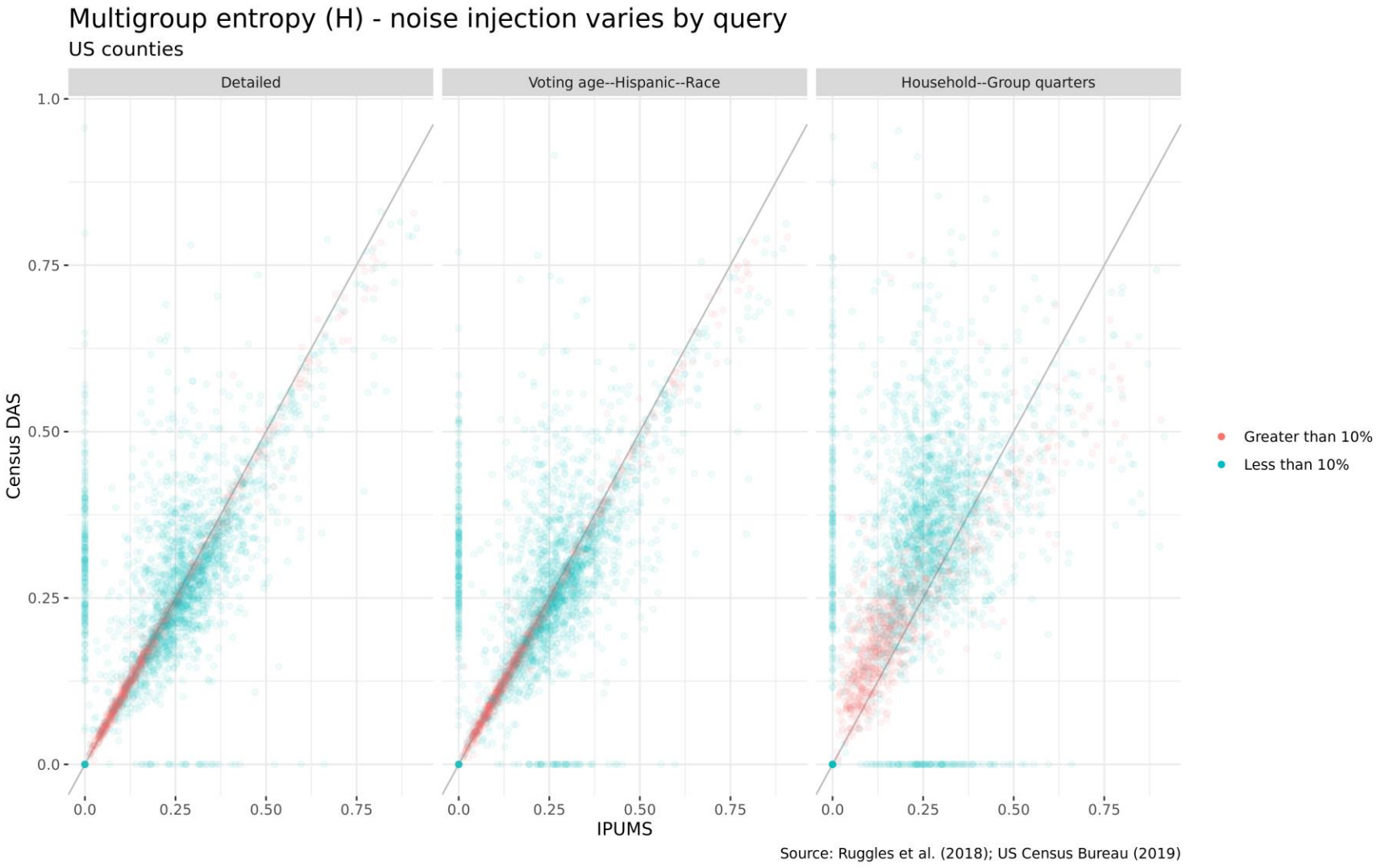


Source: Ruggles et al. (2018); US Census Bureau (2019)


African American population - noise injection varies by table
US counties
Detailed
Voting age--Hispanic--Race
Household--Group quarters
Census Bureau DAS
IPUMS
100,000
1,000
10
10
1,000
100,000
Source: Ruggles et al. (2018); US Census Bureau (2019)


Percentage of population who is African American - noise injection varies by table
US enumeration districts
Detailed
Voting age--Hispanic--Race
Household--Group quarters
Percentage, DAS
Percentage, IPUMS
0
25
50
75
100
Source: Ruggles et al. (2018); US Census Bureau (2019)


Index of dissimilarity (D) - noise injection varies by query
US counties
Detailed
Voting age--Hispanic--Race
Household--Group quarters
Census DAS
IPUMS
0.0
0.25
0.50
0.75
1.0
Greater than 10%
Less than 10%
Source: Ruggles et al. (2018); US Census Bureau (2019)


Multigroup entropy (H) - noise injection varies by query
US counties
Detailed
Voting age--Hispanic--Race
Household--Group quarters
Census DAS
IPUMS
1.0
0.75
0.50
0.25
0.0
Greater than 10%
Less than 10%
Source: Ruggles et al. (2018); US Census Bureau (2019)

Conclusions

- Diff. privacy less complicated than expected

Conclusions

- Diff. privacy less complicated than expected
- Fundamental importance of policy decisions

Conclusions

- Diff. privacy less complicated than expected
- Fundamental importance of policy decisions
- Largest impact on accuracy of small areas and small sub-populations

Conclusions

- Diff. privacy less complicated than expected
- Fundamental importance of policy decisions
- Largest impact on accuracy of small areas and small sub-populations
- Bias for segregation metrics concerning

References

2020 Census DAS Development Team. (2019) 2019. *Disclosure Avoidance System for the 2020 Census, End-to-End Release: Uscensusbureau/Census2020-Das-E2e*. Python. US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e.

Abowd, John, Daniel Kifer, Brett Moran, Robert Ashmead, Philip Leclerc, William Sexton, Simson Garfinkel, and Ashwin Machanavajjhala. 2019. "Census TopDown: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge." US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf.