Differential privacy and the decennial census

David Van Riper

vanriper@umn.edu

Tracy Kugler

José Pacas

Jonathan Schroeder

IPUMS Differential Privacy Workshop
August 15, 2019

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

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Implications of Differential Privacy for Census Bureau Data and Scientific Research

Task Force on Differential Privacy for Census Data†
Institute for Social Research and Data Innovation (ISRDI)
University of Minnesota

December 2018
Version 5.1
Working Paper No. 2018-6
https://doi.org/10.18128/MPC2018-6

Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census

Fri Feb 15 2019

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

Disclosure Avoidance and the 2018 Census Test: Release of the Source Code

Thu Jun 06 2019

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY AND DR. SIMSON L. GARFINKEL, SENIOR COMPUTER SCIENTIST FOR CONFIDENTIALITY AND DATA ACCESS

Census Bureau Continues to Boost Data Safeguards

Tue Jul 30 2019

WRITTEN BY: RON JARMIN, DEPUTY DIRECTOR, US CENSUS BUREAU

SUBSCRIBE | RSS SMS EMAIL

2010 Test Products: Intro

- The DAS team is generating test products that demonstrate the computational capabilities of the DAS at present.
- The DAS is capable of processing the 2010 CEF and producing protected microdata adhering to (slightly simplified) 2020 MDF specifications.
- The test MDF can be used to tabulate about 70% of the tables in the proposed 2020 DHC data product.

https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=304239

https://twitter.com/john_abowd/status/1160213793370181632

Overview of the Proposed 2020 Census Data Products Plan

- We are currently proposing to produce the following products for 2020
  - Apportionment Product
  - Redistricting File (P.L. 94-171)
  - Demographic Profile
  - Demographic and Housing Characteristics File *(replaces Summary File 1)*
  - Congressional District Demographic and Housing Characteristics File *(replaces Congressional District Summary File)*

- We are currently unable to produce the following products, but we are internally researching ways we can produce them for 2020
  - Detailed race and Hispanic origin tables and some household tables from Summary File 1
  - Summary File 2
  - American Indian and Alaska Native Summary File
  - Public Use Microdata Sample (PUMS) File

- We are not proposing to produce the following products for 2020
  - Summary of Population and Housing Characteristics (CPH-1)
  - Population and Housing Unit Counts Report Series (CPH-2)

https://www2.census.gov/cac/nac/meetings/2019-05/devine-hollingsworth-proposed-2020-products.pdf?

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

$$\Pr[M(D) \in S\,] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S\,] + \delta$$

Differential privacy is…

- A guarantee "on the incremental disclosure risks of participating in *D* over whatever disclosure risks the data subjects face even if they do not participate in *D.*" (Reiter 2019)

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

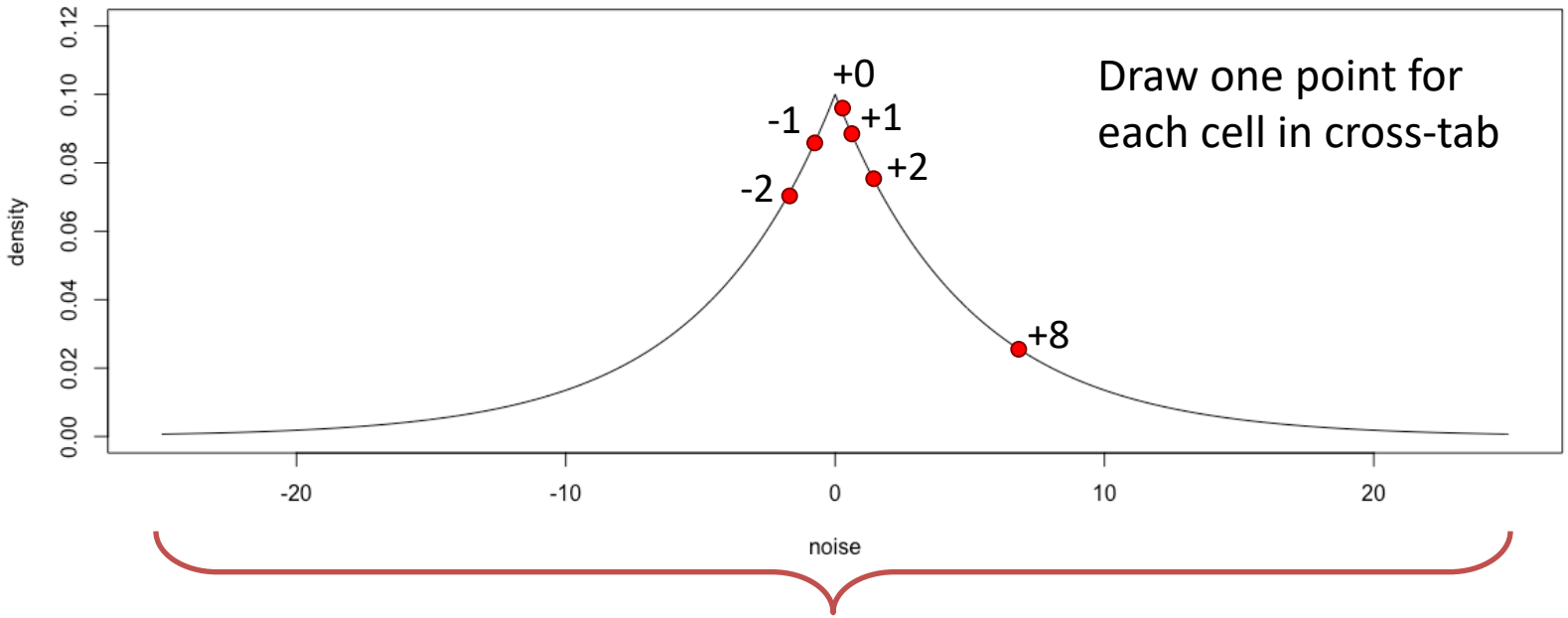
	Sex	School
x4	Female	Never
	⋮	
	Female	Never
x17	Female	Attending
	⋮	
	Female	Attending
x31	Female	Past
	⋮	
	Female	Past

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



spread is determined by **ε**

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

Male | Never
Male | Never
x12 { Male | Attending
Male | Attending
⋮
Male | Attending }
x34 { Male | Past
⋮
Male | Past }

x12 { Female | Never
⋮
Female | Never }
x19 { Female | Attending
⋮
Female | Attending }
x29 { Female | Past
⋮
Female | Past }

(True) vs. Noise infused table

	School Attendance		
	Never	Attending	Past
Male	(3) 2	(12) 12	(33) 34
Female	(4) 12	(17) 19	(31) 29

DIFFERENTIAL PRIVACY AND CENSUS

Differential privacy and census

POLICY DECISIONS

Policy decisions

- Global privacy loss budget (**ε**)

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels
  - Fraction of privacy budget allocated to each level

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels
  - Fraction of privacy budget allocated to each level
- Tables

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels
  - Fraction of privacy budget allocated to each level
- Tables
  - Fraction of privacy budget allocated to each table

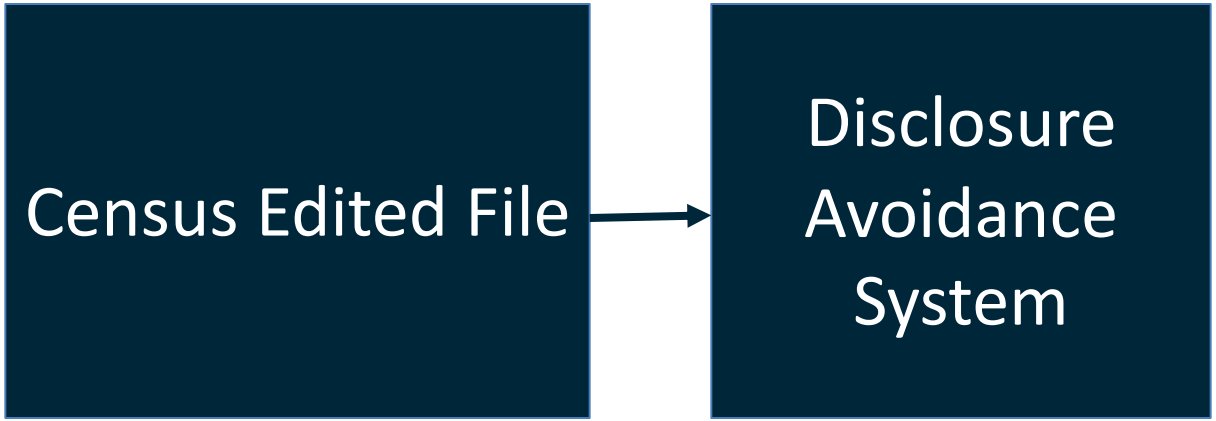
Policy decisions

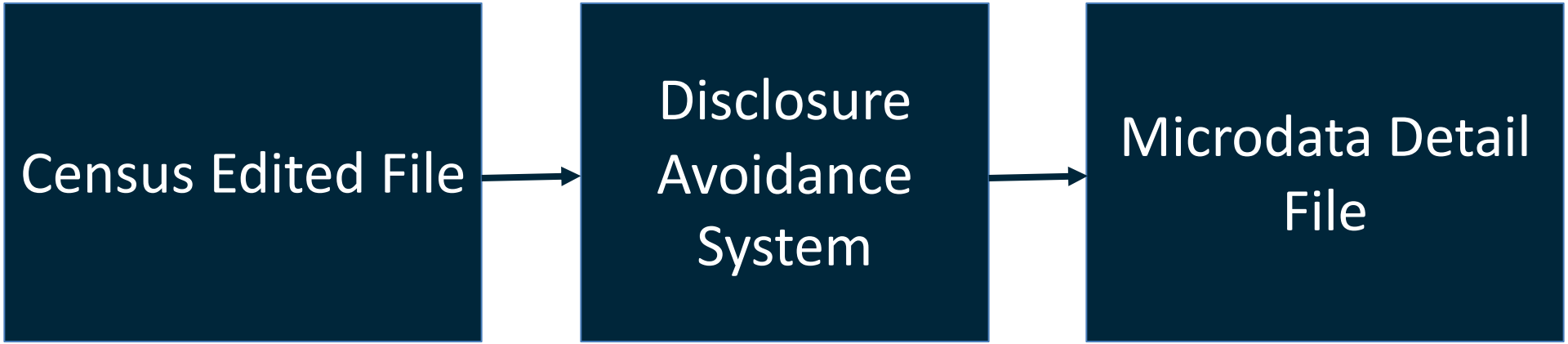
- Global privacy loss budget (**ε**)
- Geographic levels
  - Fraction of privacy budget allocated to each level
- Tables
  - Fraction of privacy budget allocated to each table
- Invariants and constraints

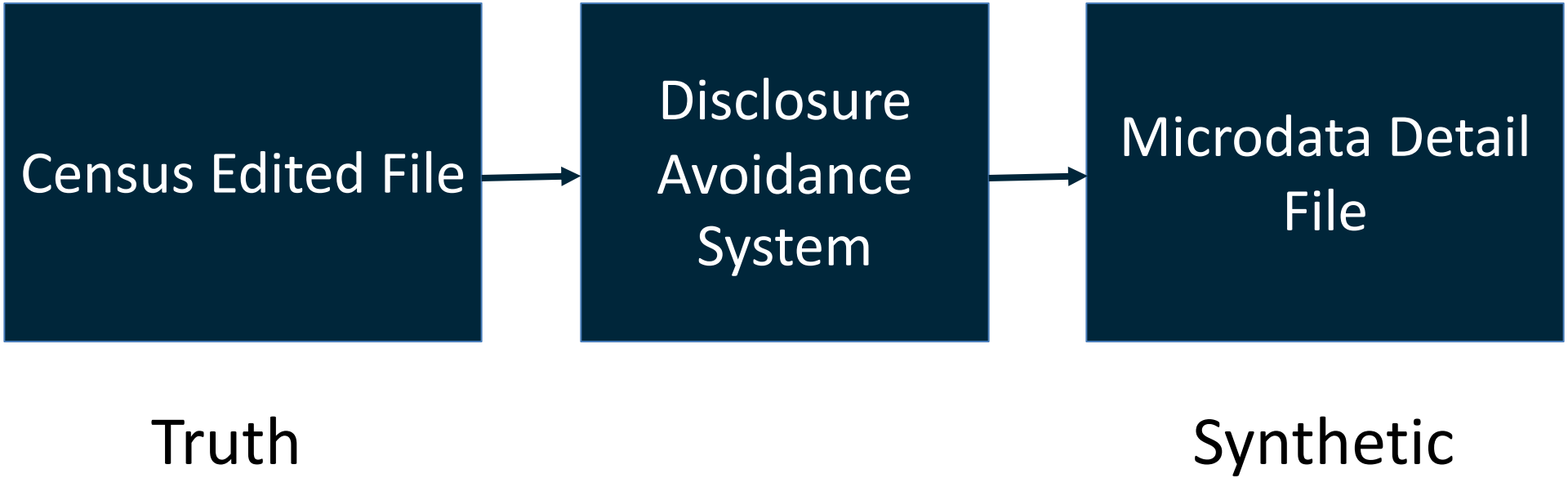
Differential privacy and census

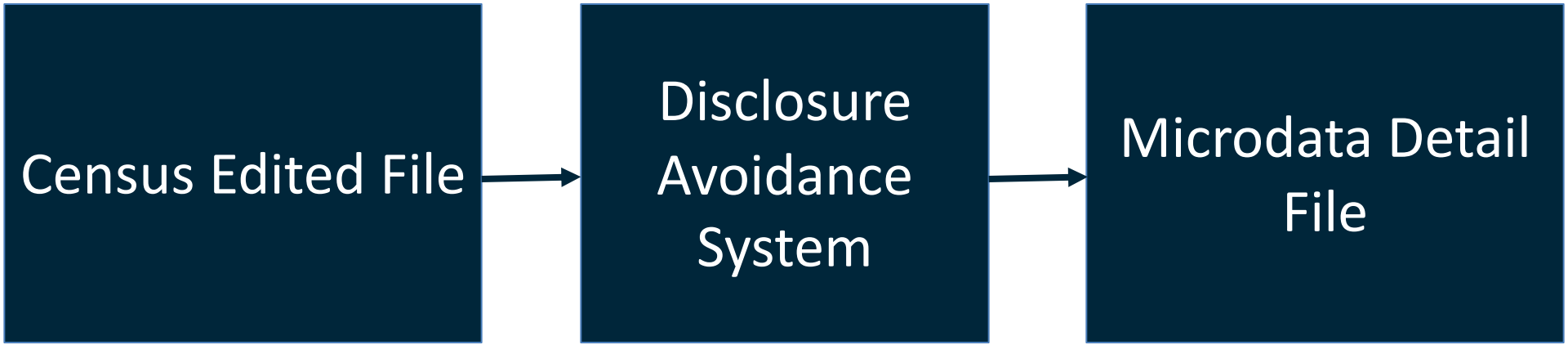
TECHNICAL IMPLEMENTATION

Census Edited File


Census Edited File
Disclosure Avoidance System

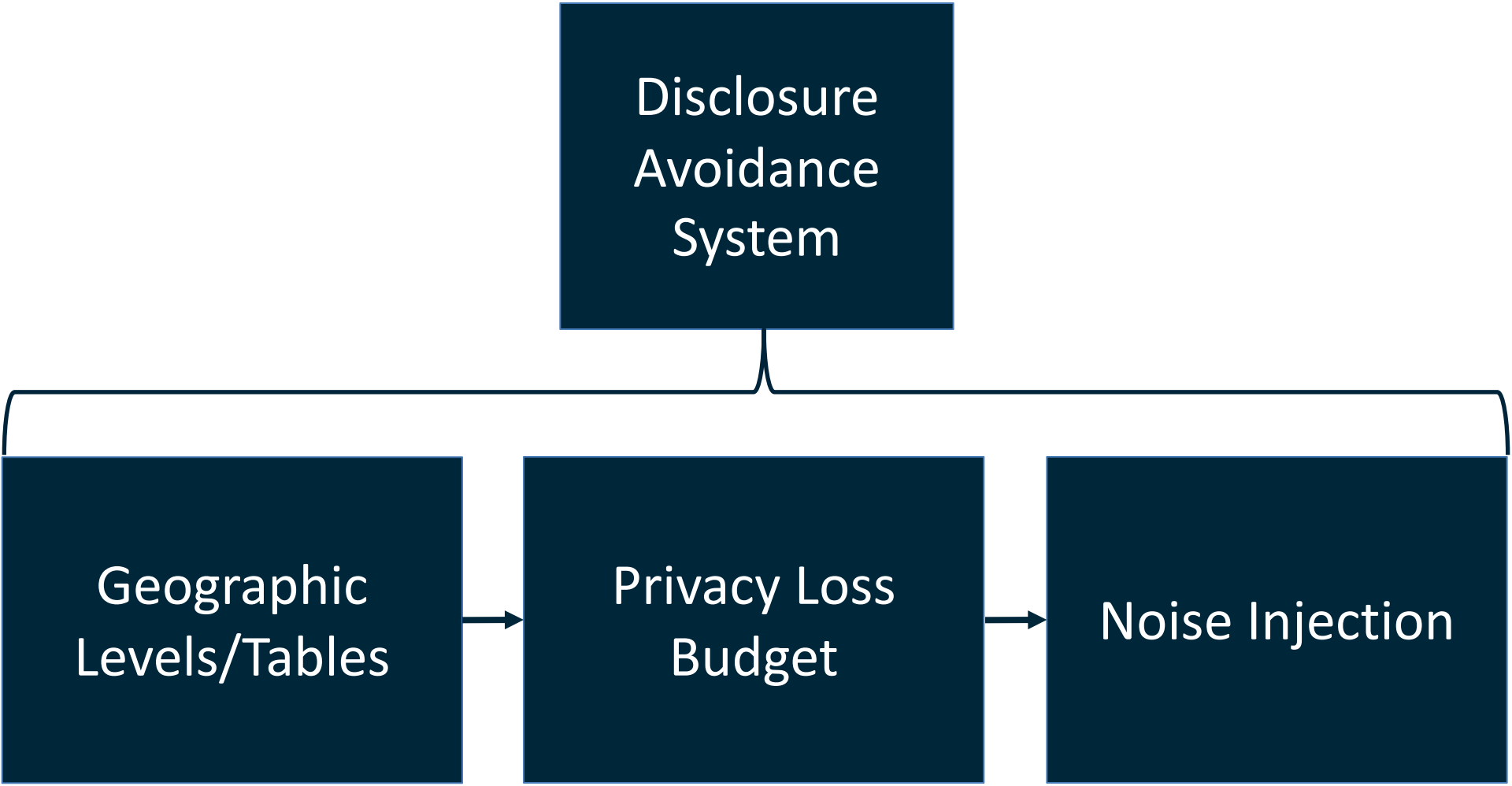

Census Edited File
Disclosure Avoidance System
Microdata Detail File


Census Edited File
Disclosure Avoidance System
Microdata Detail File
Truth
Synthetic



1940
complete-count

Differentially private
1940 complete-count


Disclosure Avoidance System
Geographic Levels/Tables
Privacy Loss Budget
Noise Injection

Technical Implementation

GEOGRAPHIC LEVELS AND TABLES

One geographic level + One table

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

- Geographic levels

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

- Geographic levels
- Tables

- Tables
  - Cross-tabulations from microdata variables

- Tables
  - Cross-tabulations from microdata variables
- 1940 tables
  - Voting age x Hispanic x Race

- Tables
  - Cross-tabulations from microdata variables
- 1940 tables
  - Voting age [2] x Hispanic [2] x Race [6]
    - 2 * 2 * 6 = 24 cells

Voting Age by Hispanic by Race

Voting Age	Hispanic	Race	Voting Age	Hispanic	Race
Under 18	Non-Hispanic	White	18+	Non-Hispanic	White
Under 18	Non-Hispanic	Black	18+	Non-Hispanic	Black
Under 18	Non-Hispanic	American Indian/Alaska Native	18+	Non-Hispanic	American Indian/Alaska Native
Under 18	Non-Hispanic	Chinese	18+	Non-Hispanic	Chinese
Under 18	Non-Hispanic	Japanese	18+	Non-Hispanic	Japanese
Under 18	Non-Hispanic	Other Asian or Pacific Islander	18+	Non-Hispanic	Other Asian or Pacific Islander
Under 18	Hispanic	White	18+	Hispanic	White
Under 18	Hispanic	Black	18+	Hispanic	Black
Under 18	Hispanic	American Indian/Alaska Native	18+	Hispanic	American Indian/Alaska Native
Under 18	Hispanic	Chinese	18+	Hispanic	Chinese
Under 18	Hispanic	Japanese	18+	Hispanic	Japanese
Under 18	Hispanic	Other Asian or Pacific Islander	18+	Hispanic	Other Asian or Pacific Islander

- Tables
  - Cross-tabulations from microdata variables
- 1940 tables
  - Voting age [2] x Hispanic [2] x Race [6]
  - Households/group quarters type [8]

- Tables
  - Cross-tabulations from microdata variables
- 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]

Technical Implementation

FRACTIONS

Geog_level
Nation
State
County
Enum dist

Geog_level	$Fraction_{geog}$
Nation	0.25
State	0.25
County	0.25
Enum dist	0.25

Geog_level	$Fraction_{geog}$
Nation	0.25
State	0.25
County	0.25
Enum dist	0.25

Sum = 1.0

Table
VHR
HHGQ
Detailed

Table	$Fraction_{table}$
VHR	0.675
HHGQ	0.225
Detailed	0.1

Table	$Fraction_{table}$
VHR	0.675
HHGQ	0.225
Detailed	0.1

Sum = 1.0

Geog_level
Nation
State
County
Enum dist

X

Table
VHR
HHGQ
Detailed

Geog_level	Table
Nation	VHR
Nation	HHGQ
Nation	Detailed
State	VHR
State	HHGQ
State	Detailed
County	VHR
County	HHGQ
County	Detailed
Enum dist	VHR
Enum dist	HHGQ
Enum dist	Detailed

Geog_level	$Fraction_{geog}$	Table	$Fraction_{table}$
Nation	0.25	VHR	0.675
Nation	0.25	HHGQ	0.225
Nation	0.25	Detailed	0.1
State	0.25	VHR	0.675
State	0.25	HHGQ	0.225
State	0.25	Detailed	0.1
County	0.25	VHR	0.675
County	0.25	HHGQ	0.225
County	0.25	Detailed	0.1
Enum dist	0.25	VHR	0.675
Enum dist	0.25	HHGQ	0.225
Enum dist	0.25	Detailed	0.1

Geog_level	$Fraction_{geog}$	Table	$Fraction_{table}$	$Frac_{geog}$ * $Frac_{table}$
Nation	0.25	VHR	0.675	
Nation	0.25	HHGQ	0.225	
Nation	0.25	Detailed	0.1	
State	0.25	VHR	0.675	
State	0.25	HHGQ	0.225	
State	0.25	Detailed	0.1	
County	0.25	VHR	0.675	
County	0.25	HHGQ	0.225	
County	0.25	Detailed	0.1	
Enum dist	0.25	VHR	0.675	
Enum dist	0.25	HHGQ	0.225	
Enum dist	0.25	Detailed	0.1	

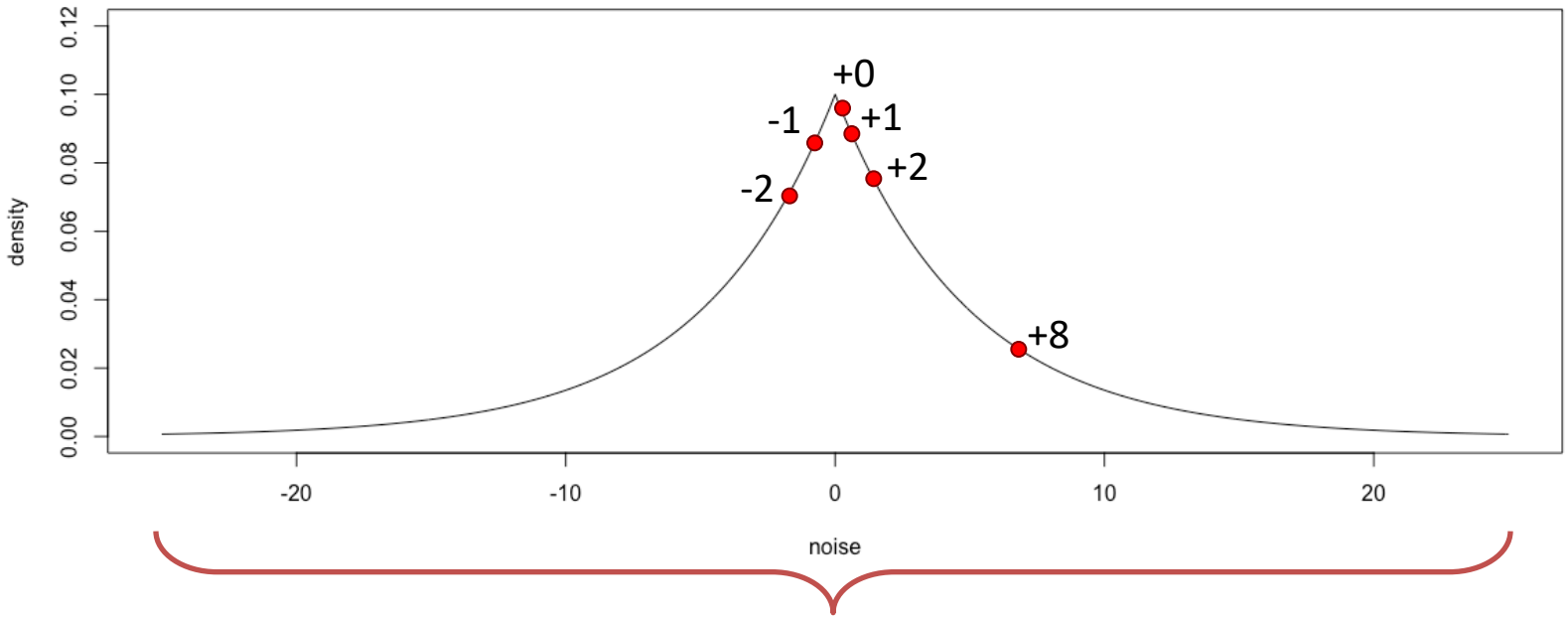
Geog_level	$Fraction_{geog}$	Table	$Fraction_{table}$	$Frac_{geog}$ * $Frac_{table}$
Nation	0.25	VHR	0.675	0.16875
Nation	0.25	HHGQ	0.225	0.05625
Nation	0.25	Detailed	0.1	0.025
State	0.25	VHR	0.675	0.16875
State	0.25	HHGQ	0.225	0.05625
State	0.25	Detailed	0.1	0.025
County	0.25	VHR	0.675	0.16875
County	0.25	HHGQ	0.225	0.05625
County	0.25	Detailed	0.1	0.025
Enum dist	0.25	VHR	0.675	0.16875
Enum dist	0.25	HHGQ	0.225	0.05625
Enum dist	0.25	Detailed	0.1	0.025

Geog_level	$Fraction_{geog}$	Table	$Fraction_{table}$	$Frac_{geog}$ * $Frac_{table}$
Nation	0.25	VHR	0.675	**0.16875**
Nation	0.25	HHGQ	0.225	**0.05625**
Nation	0.25	Detailed	0.1	**0.025**
State	0.25	VHR	0.675	**0.16875**
State	0.		5	**0.05625**
State	0.			**0.025**
County	0.25	VHR	0.675	**0.16875**
County	0.25	HHGQ	0.225	**0.05625**
County	0.25	Detailed	0.1	**0.025**
Enum dist	0.25	VHR	0.675	**0.16875**
Enum dist	0.25	HHGQ	0.225	**0.05625**
Enum dist	0.25	Detailed	0.1	**0.025**

Sum = 1.0

Technical Implementation

GLOBAL PRIVACY LOSS BUDGET (ε)



spread is determined by **ε**

$$\varepsilon_{total} = \sum_{g}^{n} \sum_{t}^{m} \varepsilon_{gt}$$

Where *g* = geographic level and *t = table*

$$\varepsilon_{gt} = \varepsilon_{total} * fraction_{gt}$$

Where ε_{total} is the global privacy budget, and $fraction_{gt}$ is the fraction of the privacy budget allocated to a particular *geographic level – table* combination

Geog_level	Table	$Frac_{geog} * Frac_{table}$
Nation	VHR	0.16875
Nation	HHGQ	0.05625
Nation	Detailed	0.025
State	VHR	0.16875
State	HHGQ	0.05625
State	Detailed	0.025
County	VHR	0.16875
County	HHGQ	0.05625
County	Detailed	0.025
Enum dist	VHR	0.16875
Enum dist	HHGQ	0.05625
Enum dist	Detailed	0.025

Geog_level	Table	$Frac_{geog} * Frac_{table}$	ε_{total}
Nation	VHR	0.16875	0.25
Nation	HHGQ	0.05625	0.25
Nation	Detailed	0.025	0.25
State	VHR	0.16875	0.25
State	HHGQ	0.05625	0.25
State	Detailed	0.025	0.25
County	VHR	0.16875	0.25
County	HHGQ	0.05625	0.25
County	Detailed	0.025	0.25
Enum dist	VHR	0.16875	0.25
Enum dist	HHGQ	0.05625	0.25
Enum dist	Detailed	0.025	0.25

Geog_level	Table	$Frac_{geog} * Frac_{table}$	ε_{total}	$\varepsilon_{geog*table}$
Nation	VHR	0.16875	0.25	0.0421875
Nation	HHGQ	0.05625	0.25	0.0140625
Nation	Detailed	0.025	0.25	0.00625
State	VHR	0.16875	0.25	0.0421875
State	HHGQ	0.05625	0.25	0.0140625
State	Detailed	0.025	0.25	0.00625
County	VHR	0.16875	0.25	0.0421875
County	HHGQ	0.05625	0.25	0.0140625
County	Detailed	0.025	0.25	0.00625
Enum dist	VHR	0.16875	0.25	0.0421875
Enum dist	HHGQ	0.05625	0.25	0.0140625
Enum dist	Detailed	0.025	0.25	0.00625

Geog_level	Table	$Frac_{geog} * Frac_{table}$	ε_{total}	$\varepsilon_{geog*table}$
Nation	VHR	0.16875	0.25	**0.0421875**
Nation	HHGQ	0.05625	0.25	**0.0140625**
Nation	Detailed	0.025	0.25	**0.00625**
State	VHR	0.16875	0.25	**0.0421875**
State			0.25	**0.0140625**
State			0.25	**0.00625**
County	VHR	0.16875	0.25	**0.0421875**
County	HHGQ	0.05625	0.25	**0.0140625**
County	Detailed	0.025	0.25	**0.00625**
Enum dist	VHR	0.16875	0.25	**0.0421875**
Enum dist	HHGQ	0.05625	0.25	**0.0140625**
Enum dist	Detailed	0.025	0.25	**0.00625**

Sum = 0.25

Technical Implementation

NOISE INJECTION

	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**

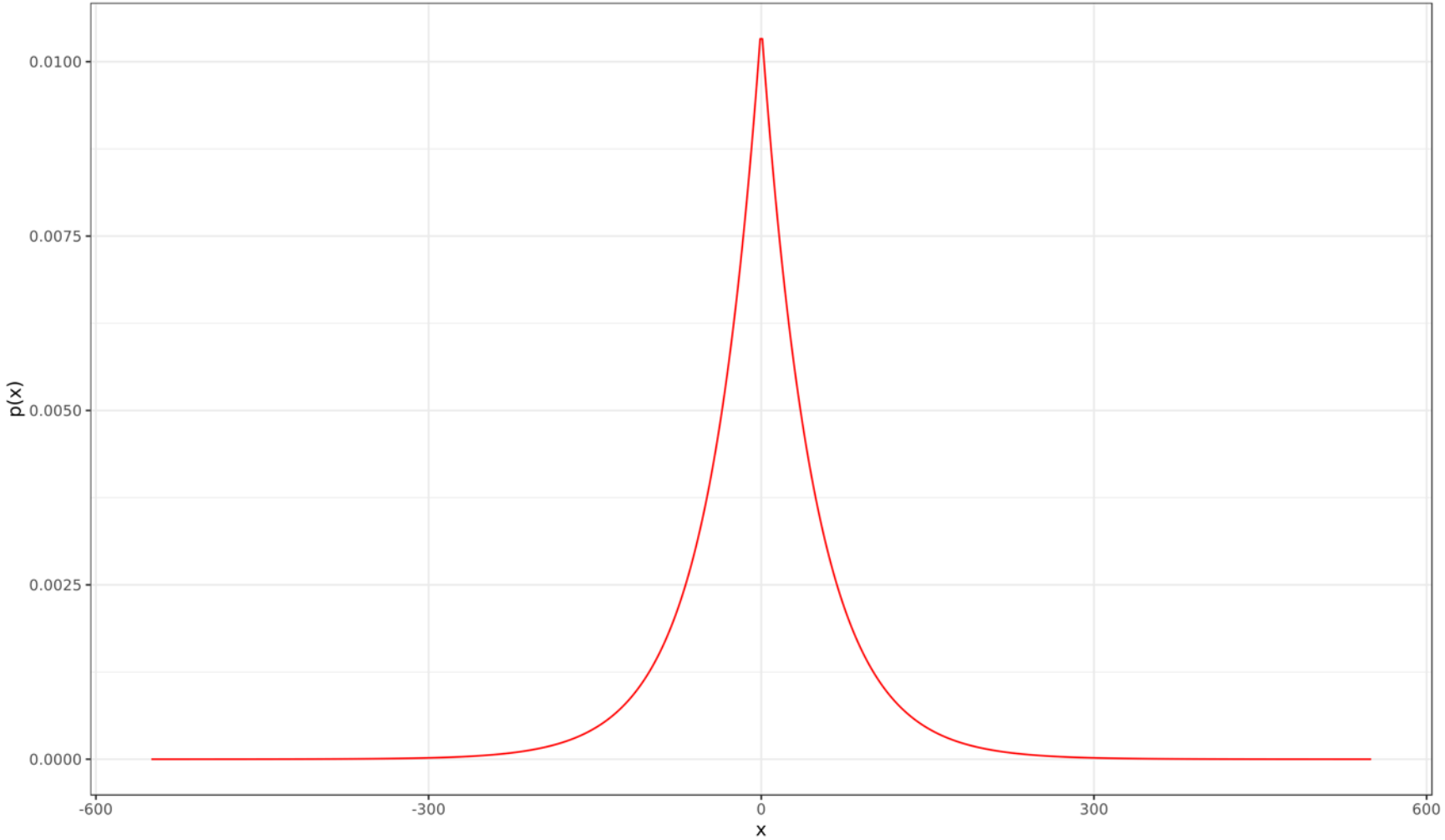
Laplace distribution*

- Centered on a mean
  - Usually zero

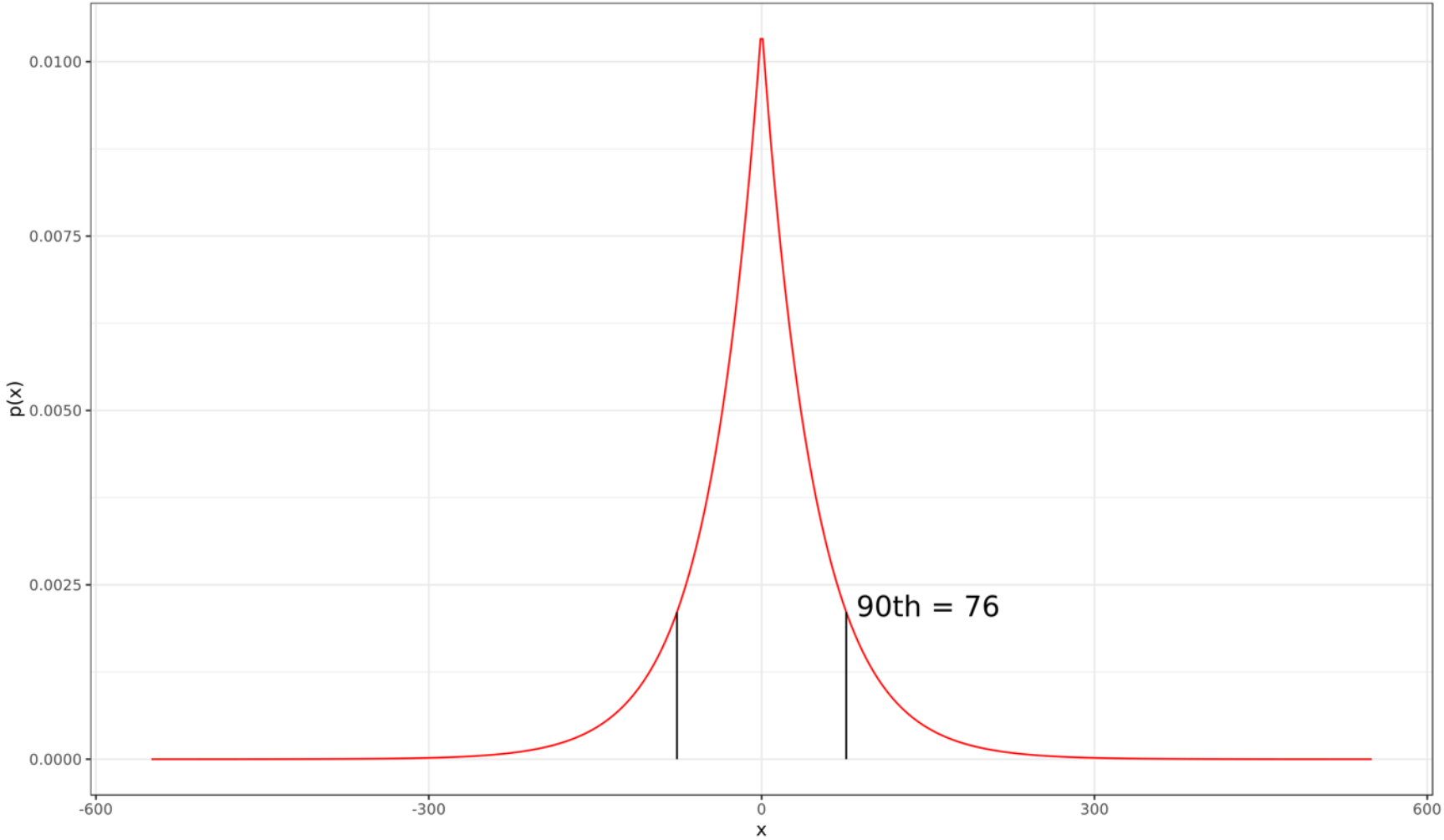
Laplace distribution

- Centered on a mean
  - Usually zero
- Spread determined by a dispersion parameter whose value is controlled by $\varepsilon_{geog*table}$

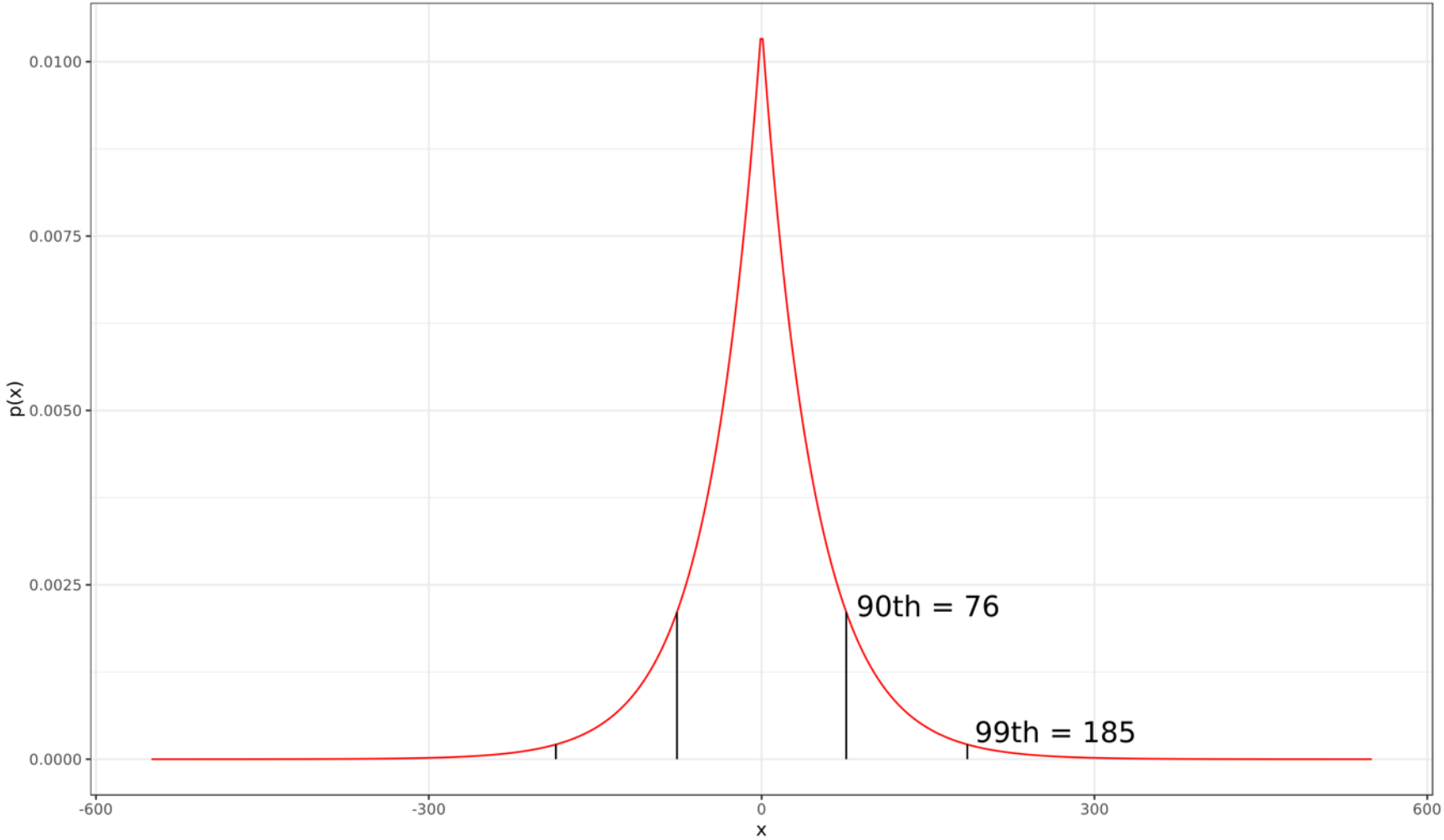
Laplace distribution



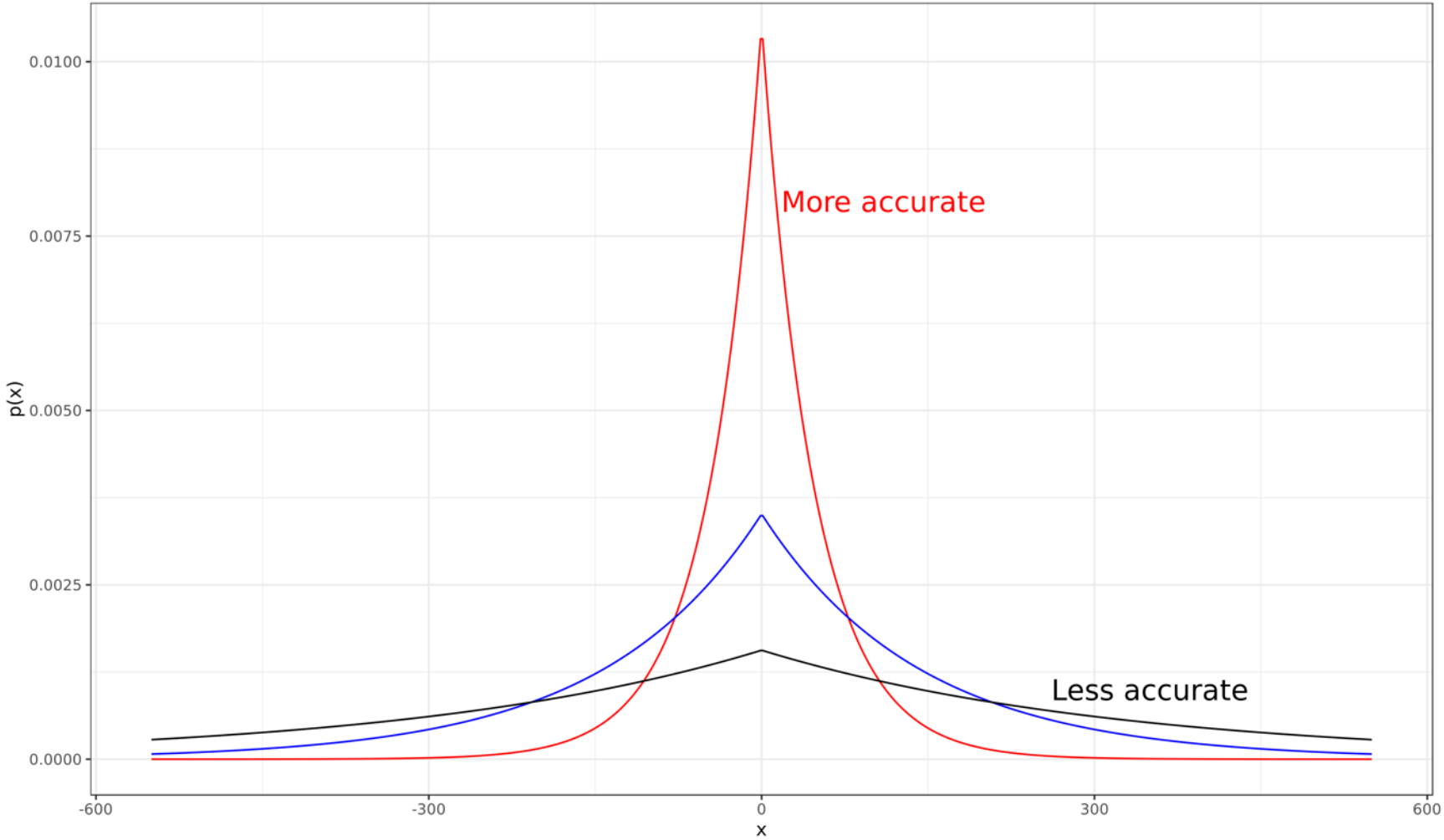
Percentiles for Laplace distribution

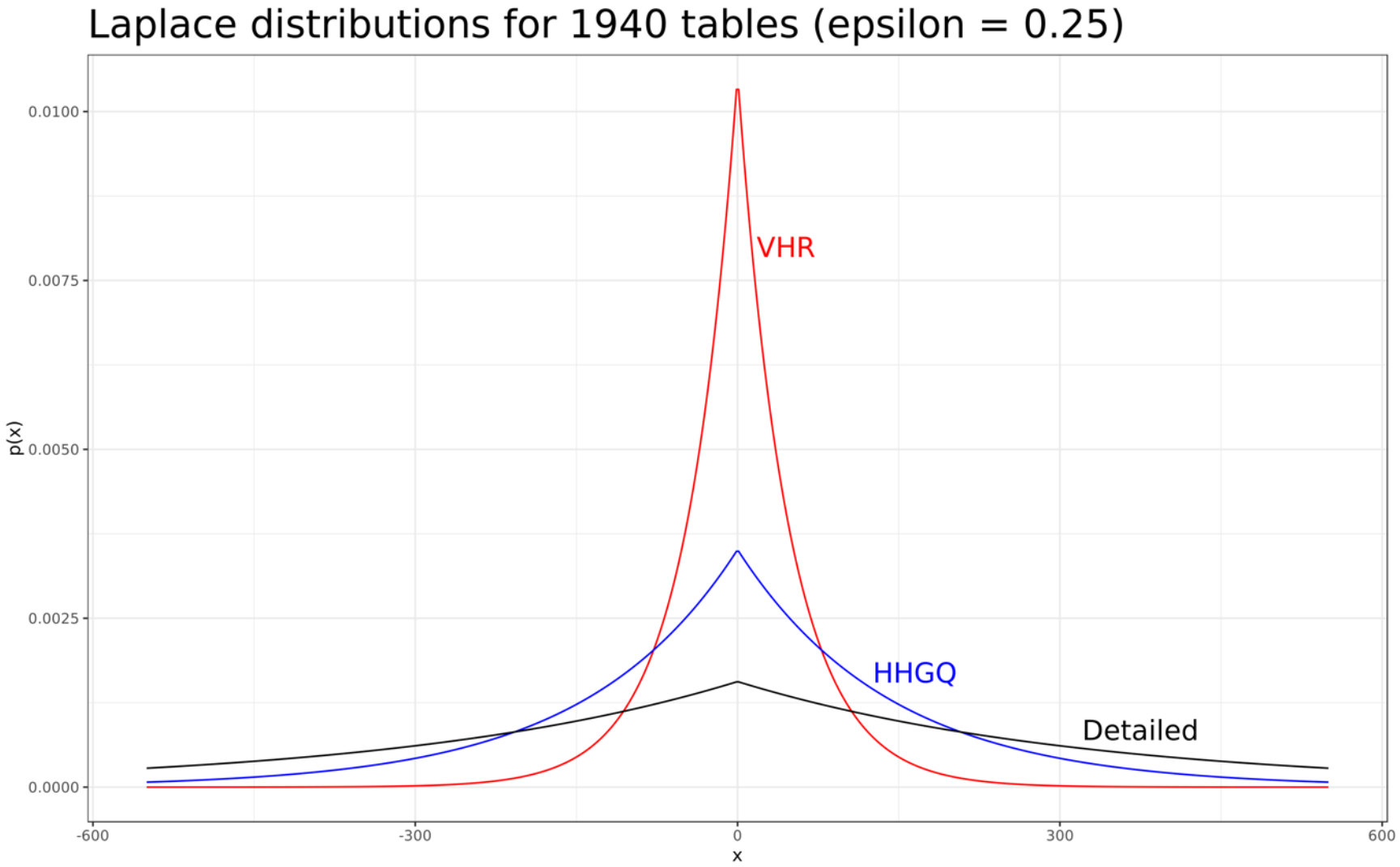


Percentiles for Laplace distribution



Laplace distributions for different epsilons




Laplace distributions for 1940 tables (epsilon = 0.25)
0.0100
0.0075
0.0050
0.0025
0.0000
p(x)
VHR
HHGQ
Detailed
-600
-300
0
300
600
x

FULL IMPLEMENTATION

1. Generate microdata without geographic identifiers
2. Assign geographic identifiers to each microdata record

Step 1

1. Create national table from “true” data

Step 1

1. Create national table from “true” data
2. For each cell in table, infuse noise drawn from Laplace distribution

Step 1

1. Create national table from “true” data
2. For each cell in table, infuse noise drawn from Laplace distribution
3. Generate microdata with no geographic identifiers from (2) via database reconstruction

Step 2

1. Create state table from “true” data

Step 2

1. Create state table from “true” data
2. For each cell in table, infuse noise drawn from Laplace distribution

Step 2

1. Create state table from “true” data
2. For each cell in table, infuse noise drawn from Laplace distribution
3. Use linear optimization to fit Step 1 microdata to “noisy” state cells

Step 2

1. Create state table from “true” data
2. For each cell in table, infuse noise drawn from Laplace distribution
3. Use linear optimization to fit Step 1 microdata to “noisy” state cells
4. Assign state identifier to each Step 1 microdata record

Step 2

1. Create state table from “true” data
2. For each cell in table, infuse noise drawn from Laplace distribution
3. Use linear optimization to fit Step 1 microdata to “noisy” state cells
4. Assign state identifier to each Step 1 microdata record
5. Repeat (1) – (4) for remaining geographic levels (counties and enumeration districts)

Output

- Microdata records with state, county, and enumeration district identifiers

Output

- Microdata records with state, county, and enumeration district identifiers
- Process does not change characteristics of individual microdata records generated in Step 1
  - But it assigns geographic identifiers so final microdata are “best fitting” to the ”noisy” cross tabs for enumeration districts

Unknowns (things I don’t understand)

Unknowns (things I don’t understand)

- Invariants

Unknowns (things I don’t understand)

- Invariants

1

EFFECTIVE PRIVACY AFTER ADJUSTING FOR INVARIANTS WITH APPLICATIONS TO THE 2020 CENSUS

ROBERT ASHMEAD, DANIEL KIFER, PHILIP LECLERC, ASHWIN MACHANAVAJJHALA, AND WILLIAM SEXTON

ABSTRACT. Differential privacy is a mathematical tool for protecting the confidentiality of records belonging to individuals. One of the key premises of differential privacy is that *any* measurement based on the confidential data must be altered with carefully chosen random noise before publication to mask the possible effect of one person on the data. However, the deployment of differentially private disclosure limitation technologies by official statistical agencies may not always occur under ideal conditions. For instance, internal decisions or external requirements (e.g., legal or contractual obligations) may stipulate that certain statistics must be published *exactly*. Additionally, overlapping datasets may have already been published. In this paper, we explain (1) the semantics of algorithms that satisfy differential privacy, (2) how the semantics are affected by release of exact statistics (computed directly from the confidential data), (3) how to attribute responsibility for any resulting information leakage, (4) how to provide privacy semantics for the combined information leakage.

https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0941_Effective_Privacy_after_Adjusting_for_Constraints__With_applications_to_the_2020_Census.pdf

Unknowns (things I don’t understand)

- Invariants
  - 2000 and 2010 invariants (block level)
    - Total population
    - Voting age population
    - Housing units
    - Occupied housing units
    - Group quarters

Unknowns (things I don’t understand)

- Household – person join tables

Unknowns (things I don't understand)

- Household – person join tables

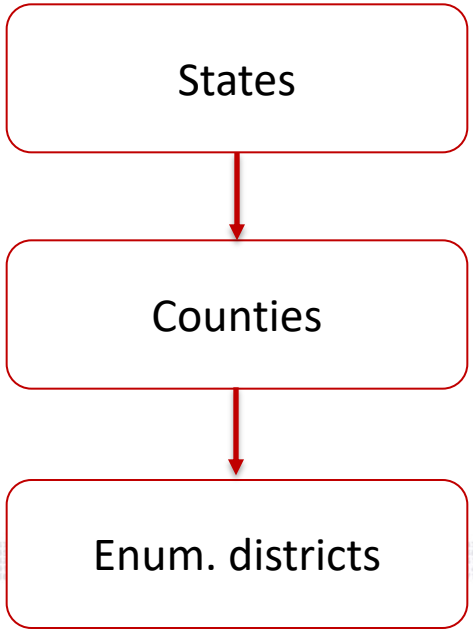
Redesign of Data Products

- Microdata supported products (PL94, DHC-Persons and DHC-Households, Demo Profile).
- Other Products (Detailed Race/AIAN, Person-Household Joins) - Out of scope for TopDown.
- Table classification:
  - by geography.
  - by universe: what is the item being counted (person or household)?

https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=304239

Unknowns (things I don’t understand)

- Inheritance of accuracy

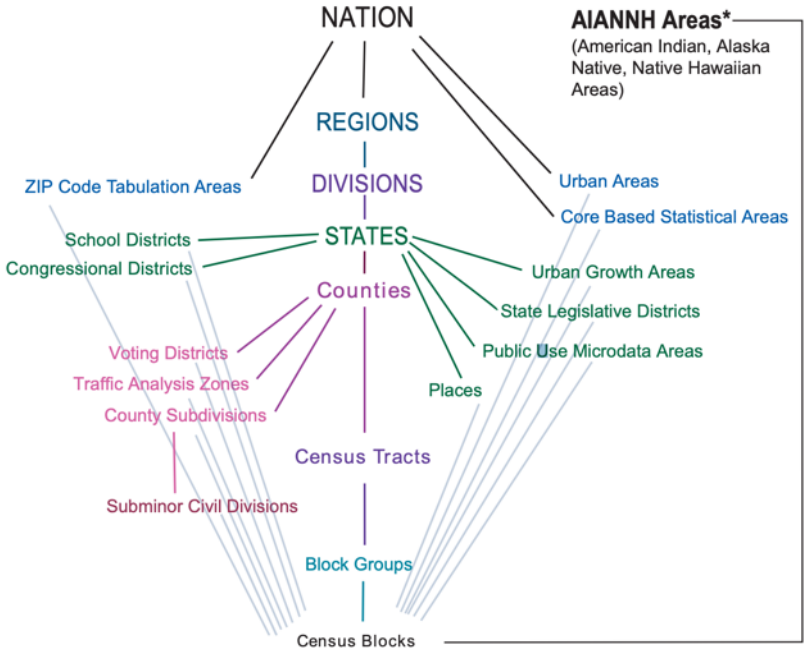


Unknowns (things I don’t understand)

- Off-spine geographic levels

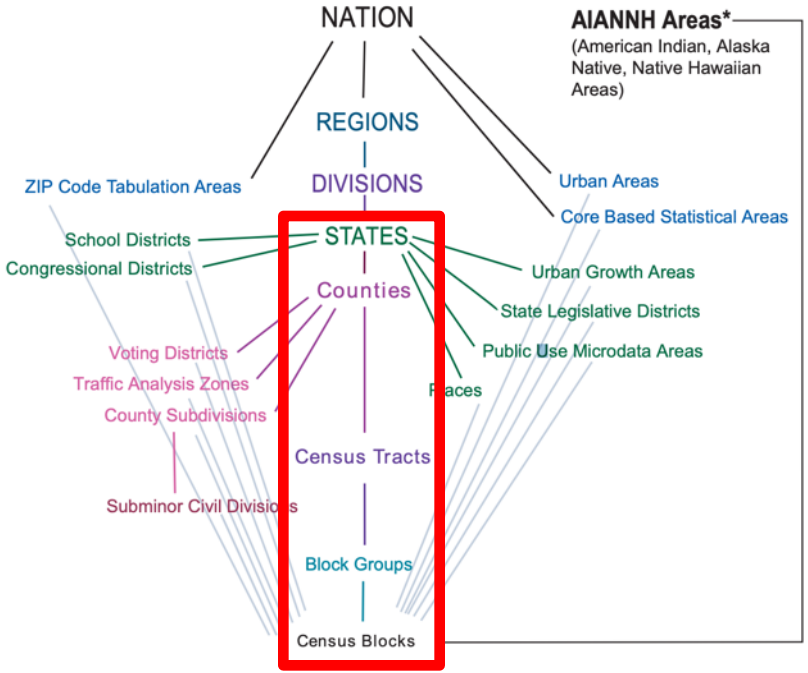
Unknowns (things I don’t understand)

- Off-spine geographic levels



Unknowns (things I don’t understand)

- Off-spine geographic levels



ANALYZING DIFFERENTIALLY PRIVATE 1940 CENSUS DATA

- Census Bureau released differentially private 1940 census data on June 6, 2019

- Census Bureau released differentially private 1940 census data on June 6, 2019
- Eight different ε values

1940 Privacy Budgets (ε)

0.25	2.0
0.50	4.0
0.75	6.0
1.0	8.0

- Census Bureau released differentially private 1940 census data on June 6, 2019
- Eight different ε values
- Four runs for each ε value
  - 32 total data files

```
MPD1940|3.1.4|39|0890|0050|938427|1|5|99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|4|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|1|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|1|5|99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
```

```
MPD1940|3.1.4 39|0890|0050 938427|1|5|99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0050 938427|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0050 938427|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0050 938427|4|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0080 939107|1|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0080 939107|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0080 939107|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0090 939819|1|5|99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0090 939819|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4 39|0890|0090 939819|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
```

```
MPD1940|3.1.4|39|0890|0050 938427|1|5 99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050 938427|2|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050 938427|3|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050 938427|4|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080 939107|1|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080 939107|2|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080 939107|3|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090 939819|1|5 99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090 939819|2|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090 939819|3|5 99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
```

```
MPD1940|3.1.4|39|0890|0050|938427|1|5|99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|4|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|1|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|1|5|99|9|17|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|2|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|3|5|99|9|18|1|1|9999|9999|9999|9999|9999|9999|9999|9999|9999|9
```

```
MPD1940|3.1.4|39|0890|0050|938427|1|5 99|9 17|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|2|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|3|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0050|938427|4|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|1|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|2|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0080|939107|3|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|1|5 99|9 17|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|2|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
MPD1940|3.1.4|39|0890|0090|939819|3|5 99|9 18|1|1 9999|9999|9999|9999|9999|9999|9999|9999|9999|9
```

- State, county, enumeration district
- Voting age
- Hispanic
- Race
- Group quarters type

- Focused on:
  - Total population
  - Voting age
  - African Americans
  - Segregation
- For counties and enumeration districts

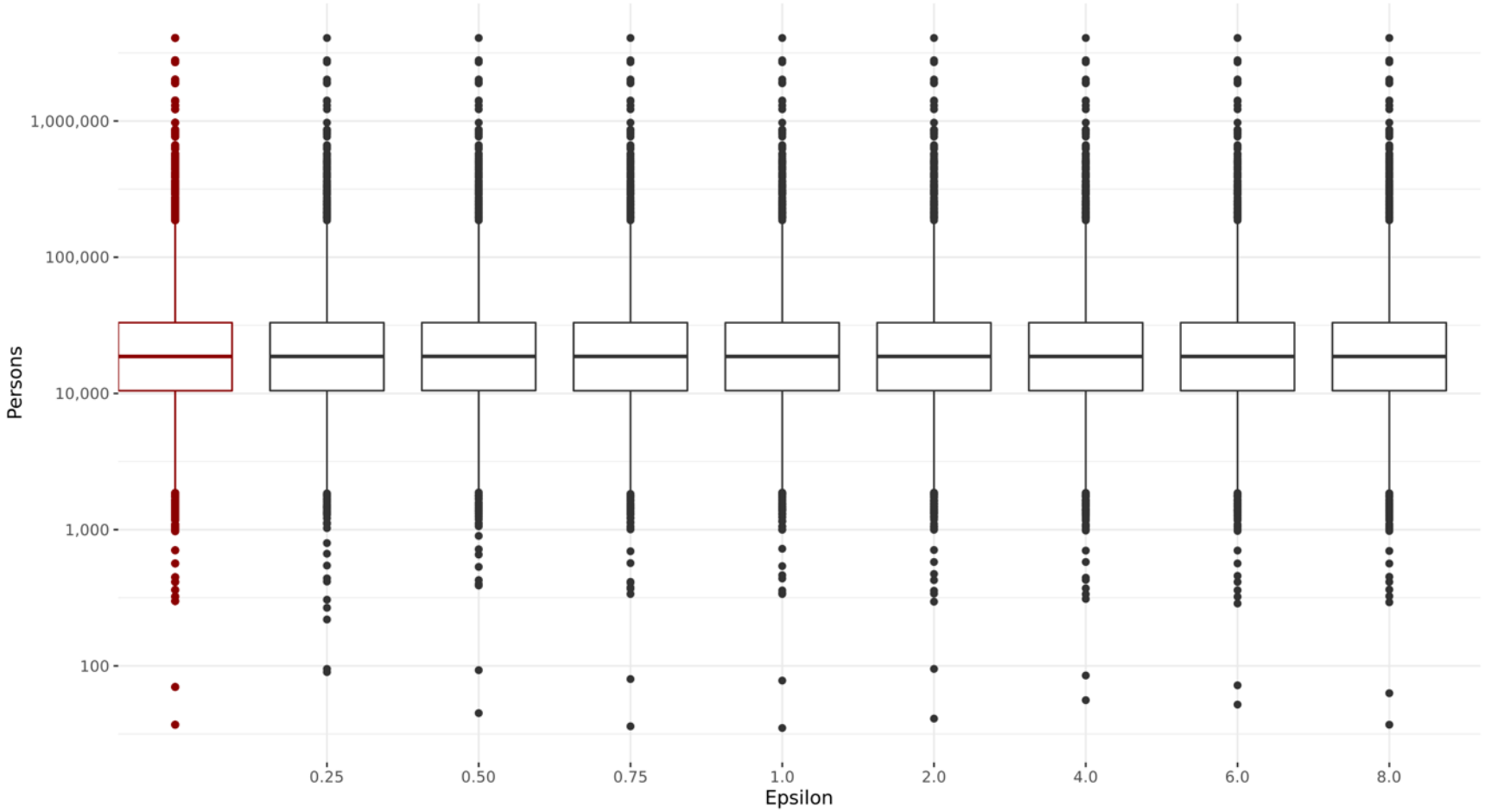
- Comparisons between ”true” data (IPUMS 1940 complete-count) and differentially private data
- Examine impact noise injection has on counts

Key takeaways

- Geographic units with smaller populations are less accurate
- Small sub-populations are less accurate

Distribution of total population under different levels of noise injection

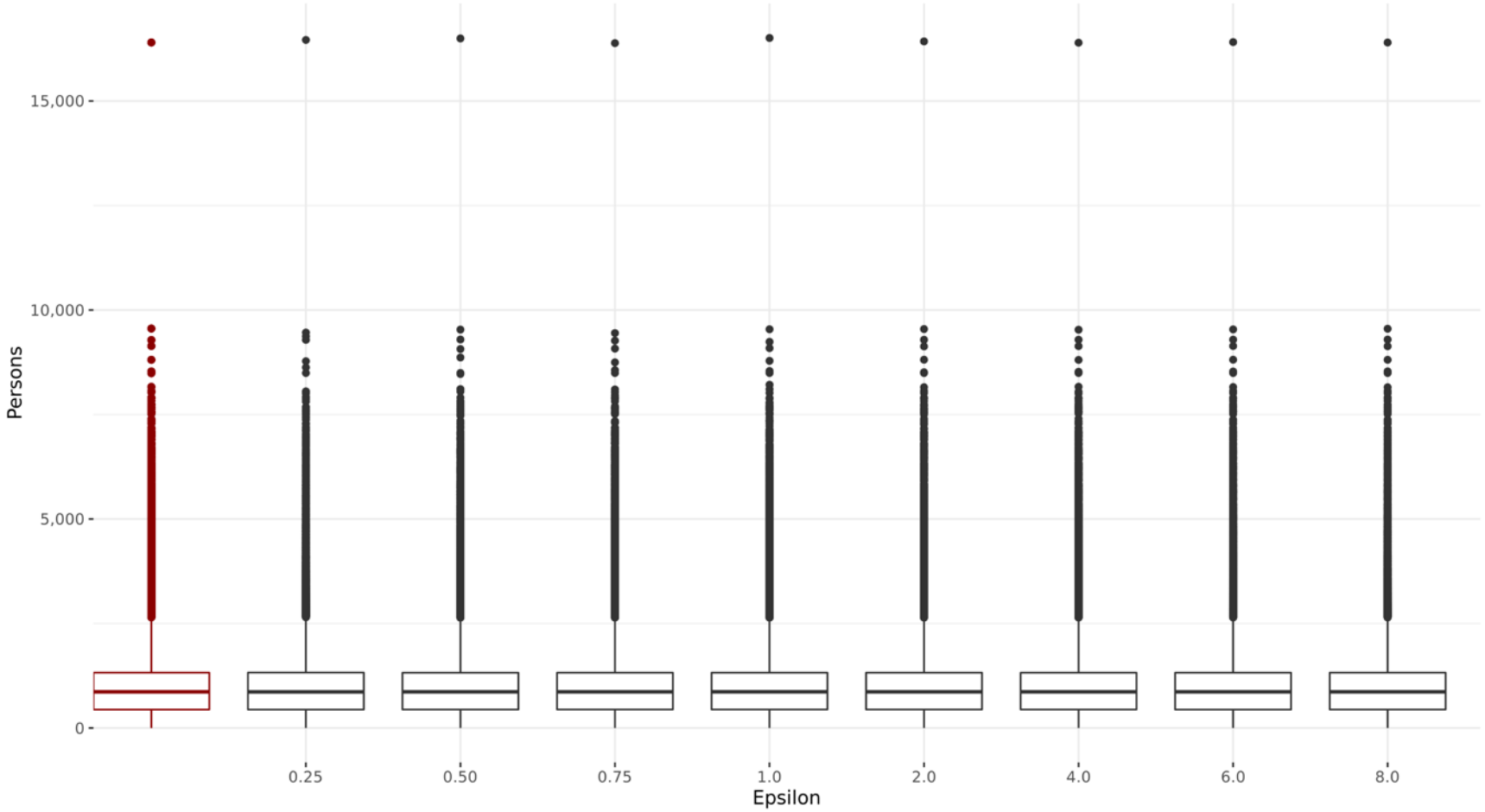
US counties



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

Distribution of total population under different levels of noise injection

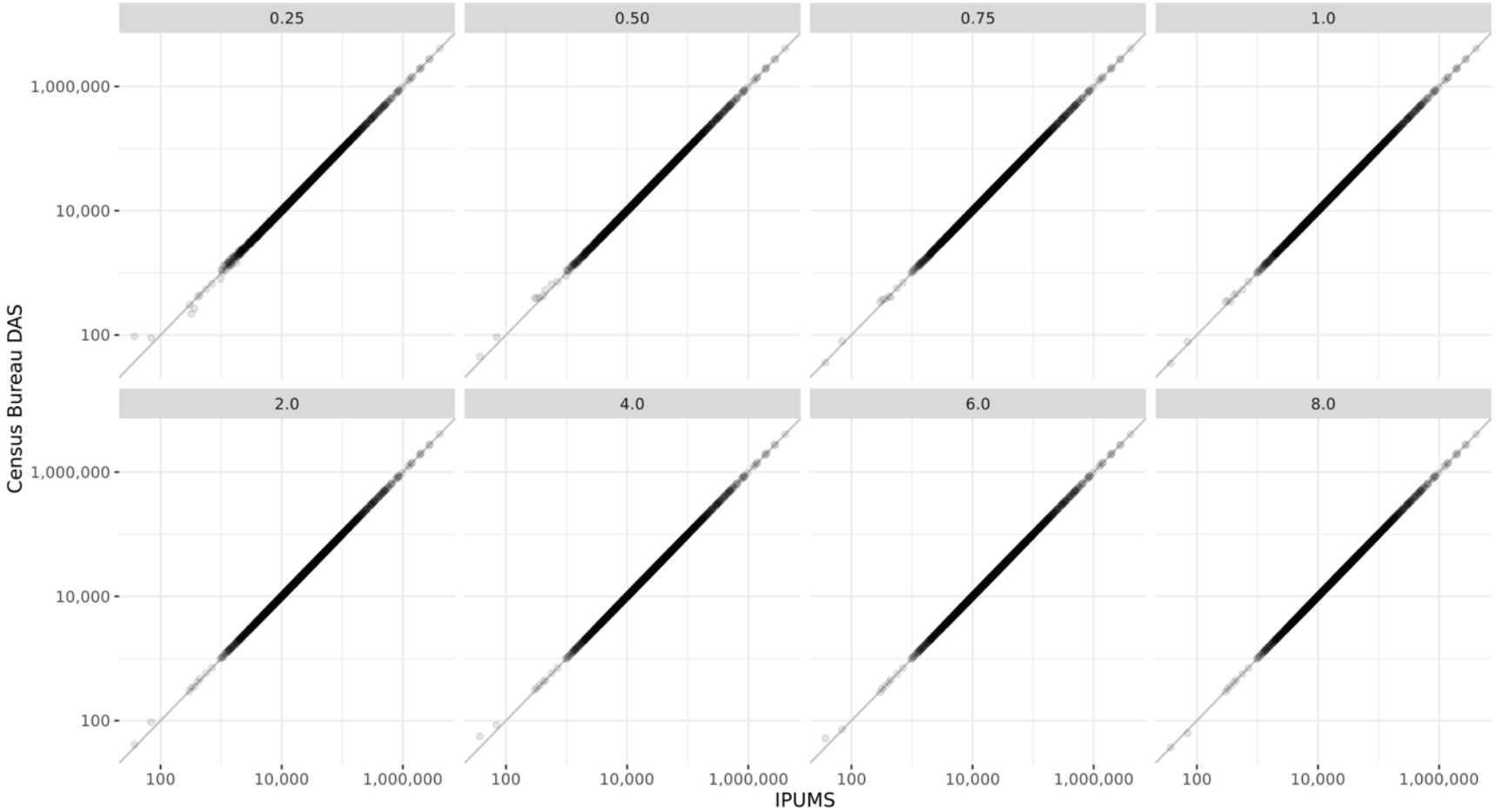
US enumeration districts



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

Total population under different levels of noise injection

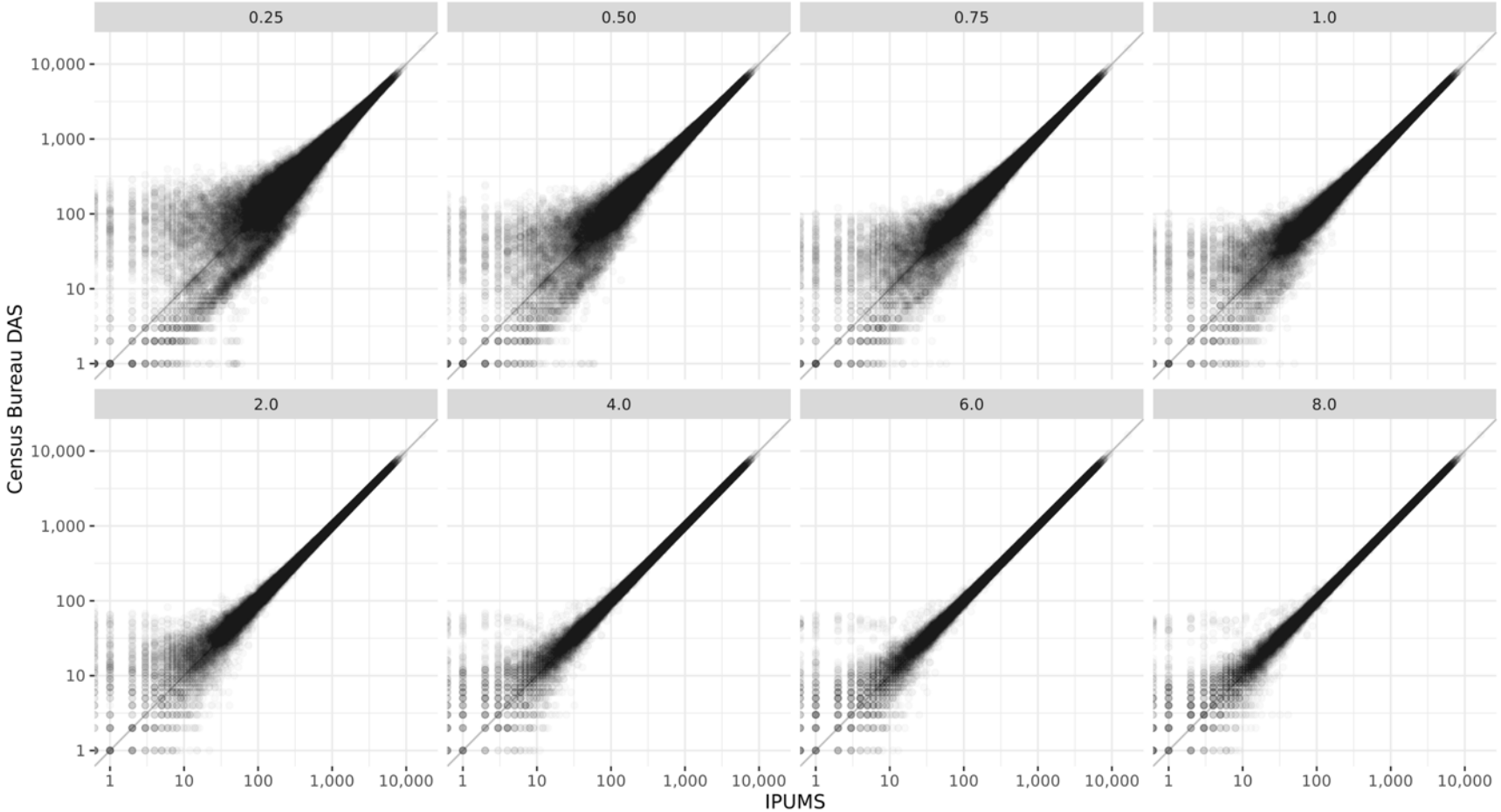
US counties



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

Total population under different levels of noise injection

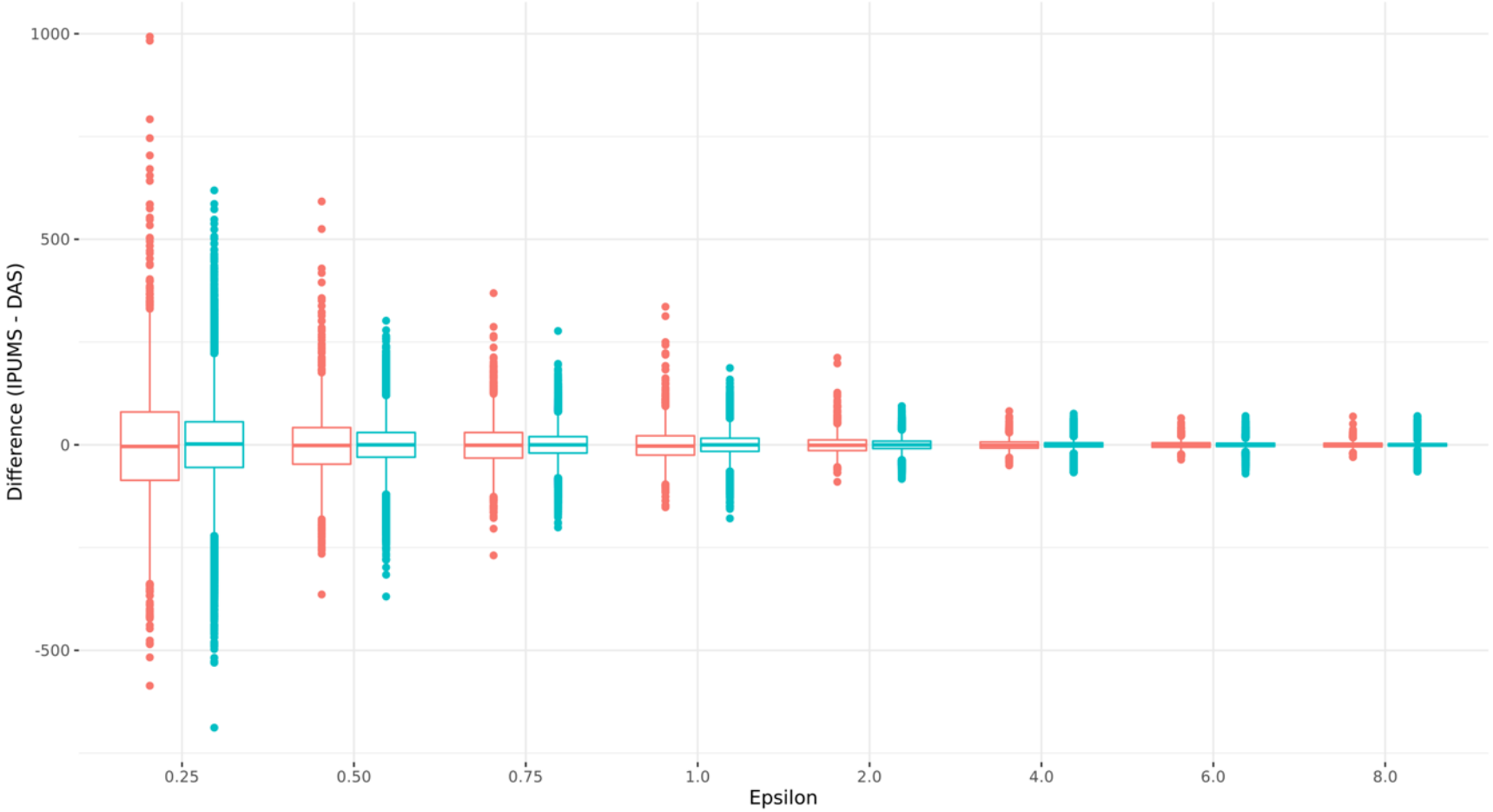
US enumeration districts



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

Difference between IPUMS and Census DAS total population counts

US counties (orange) and EDs (teal)



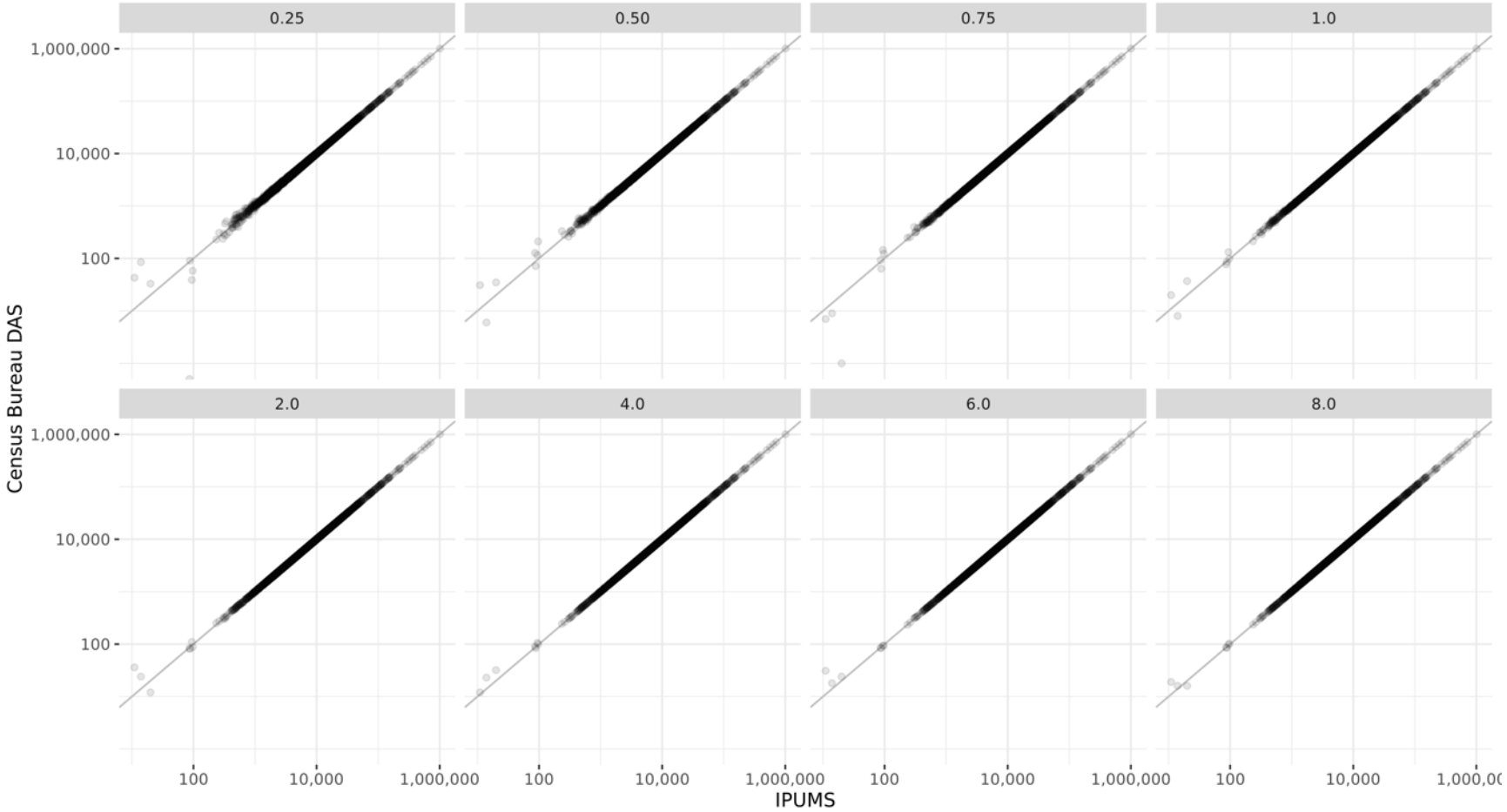
Source: Ruggles et al. (2018); US Census Bureau (2019)

TABULAR SUMMARIES

NON-VOTING AGE POPULATION

Population under age 18 for different levels of noise injection

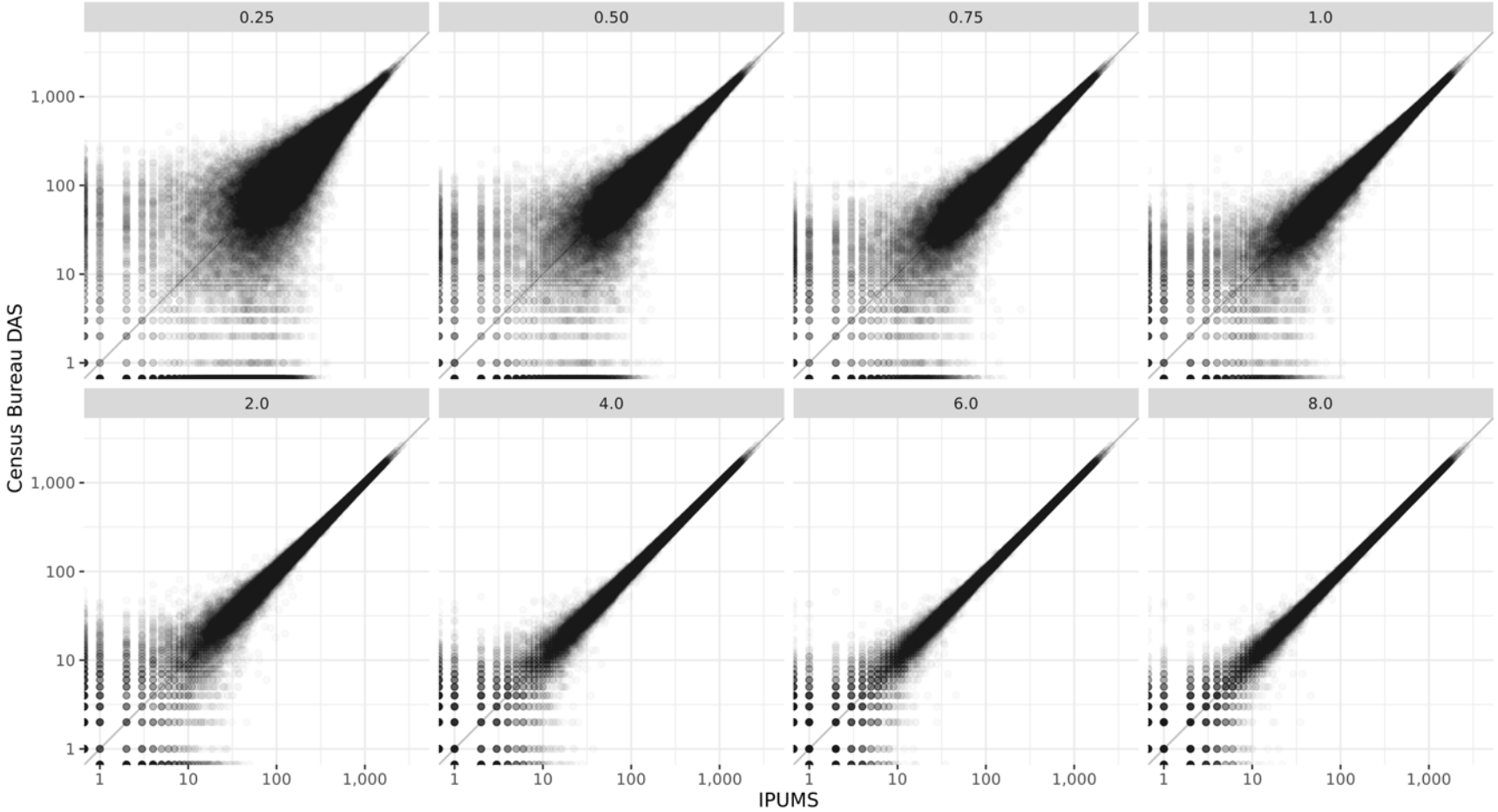
US counties



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

Population under age 18 for different levels of noise injection

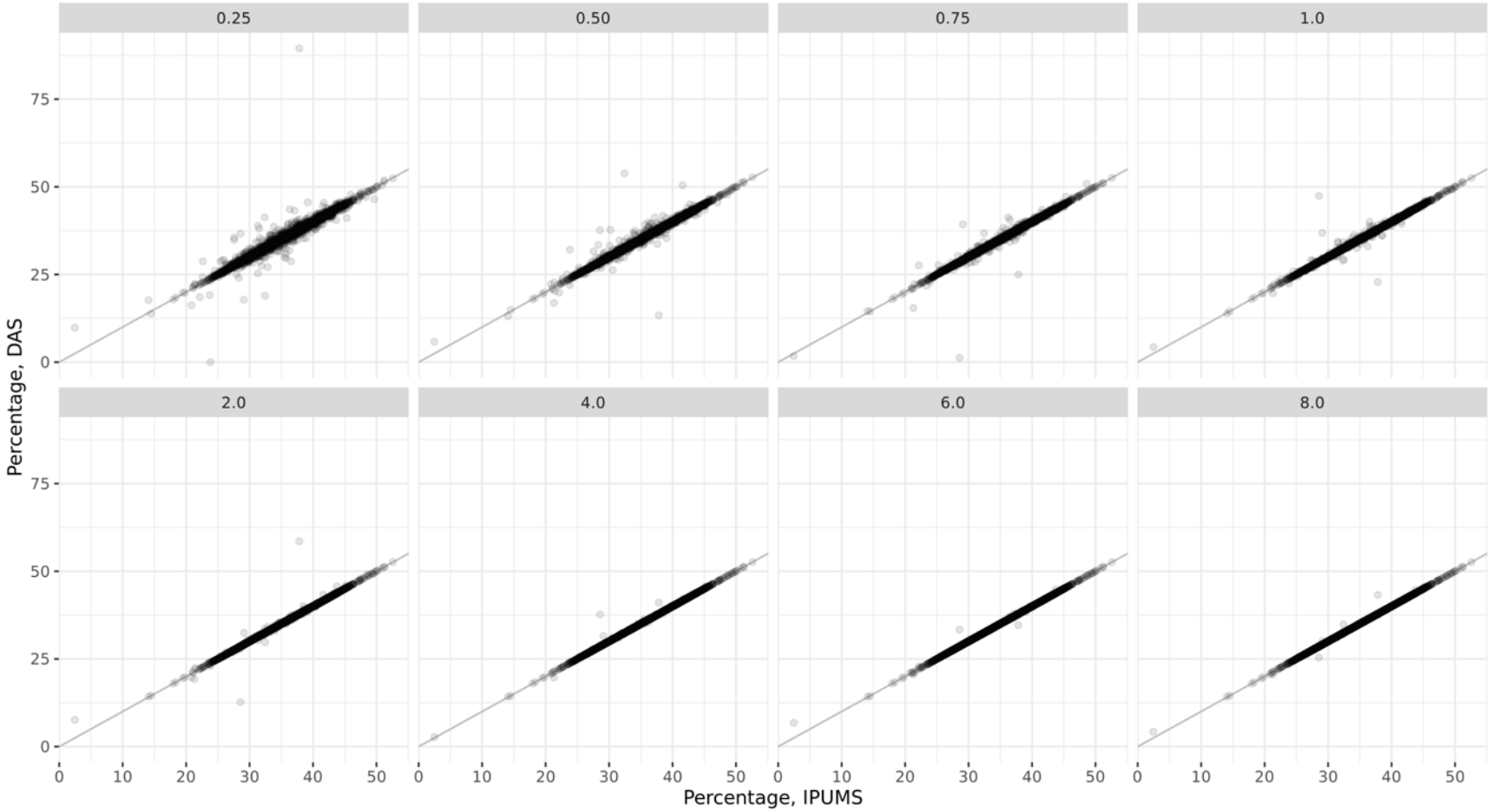
US enumeration districts



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

Percentage of population under age 18

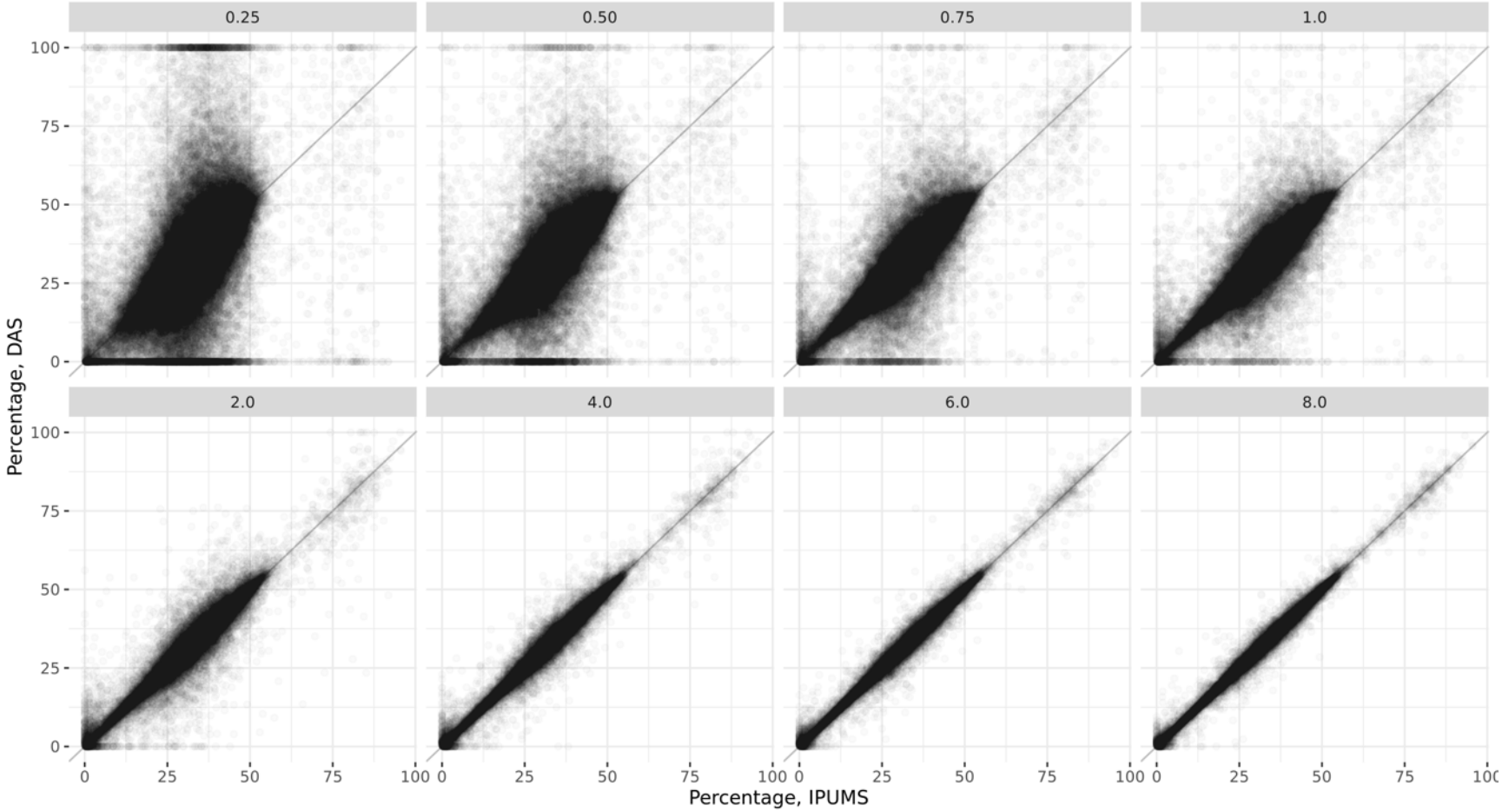
US counties



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

Percentage of population under age 18

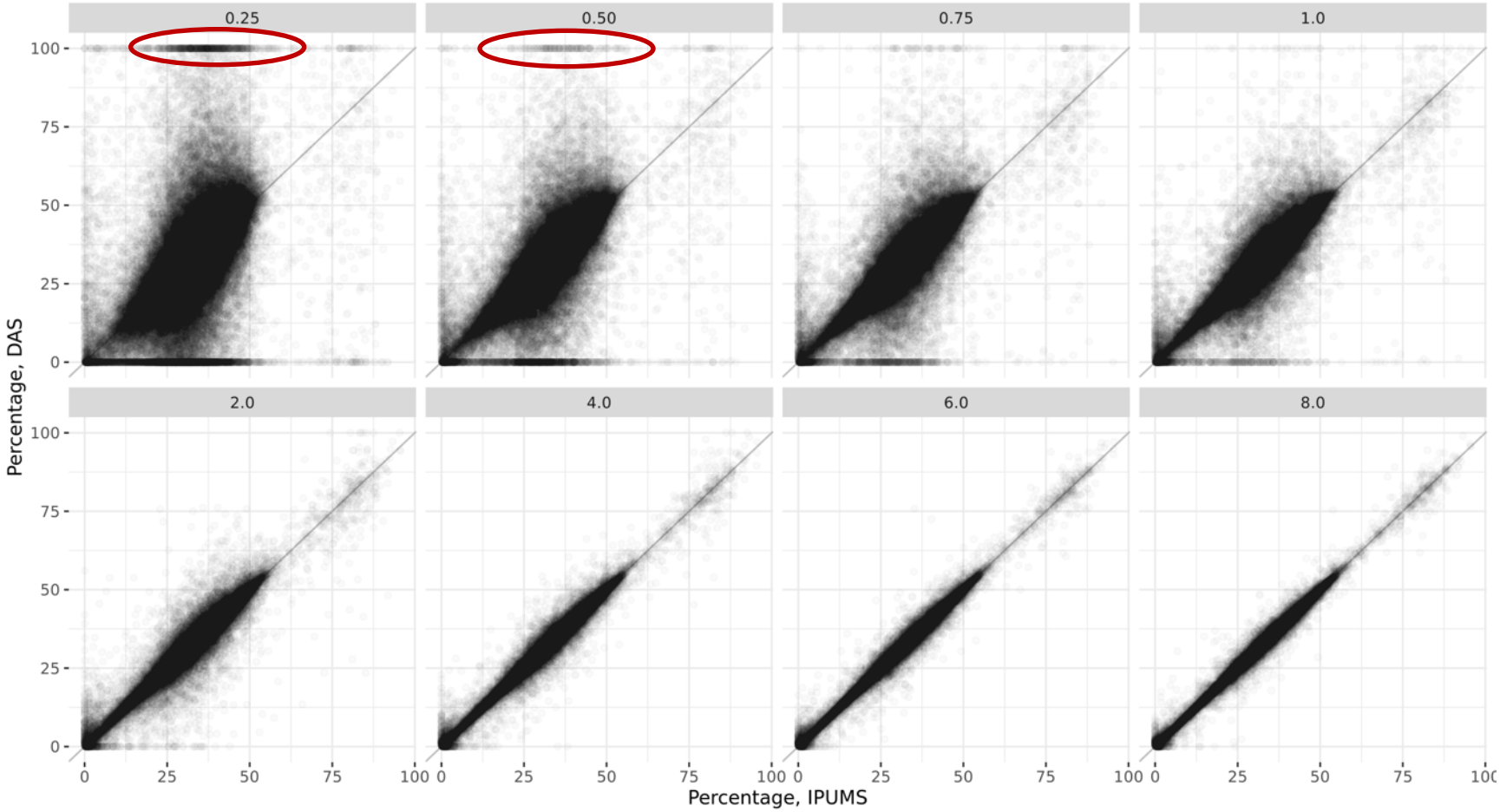
EDs containing 50 or more persons



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

Percentage of population under age 18

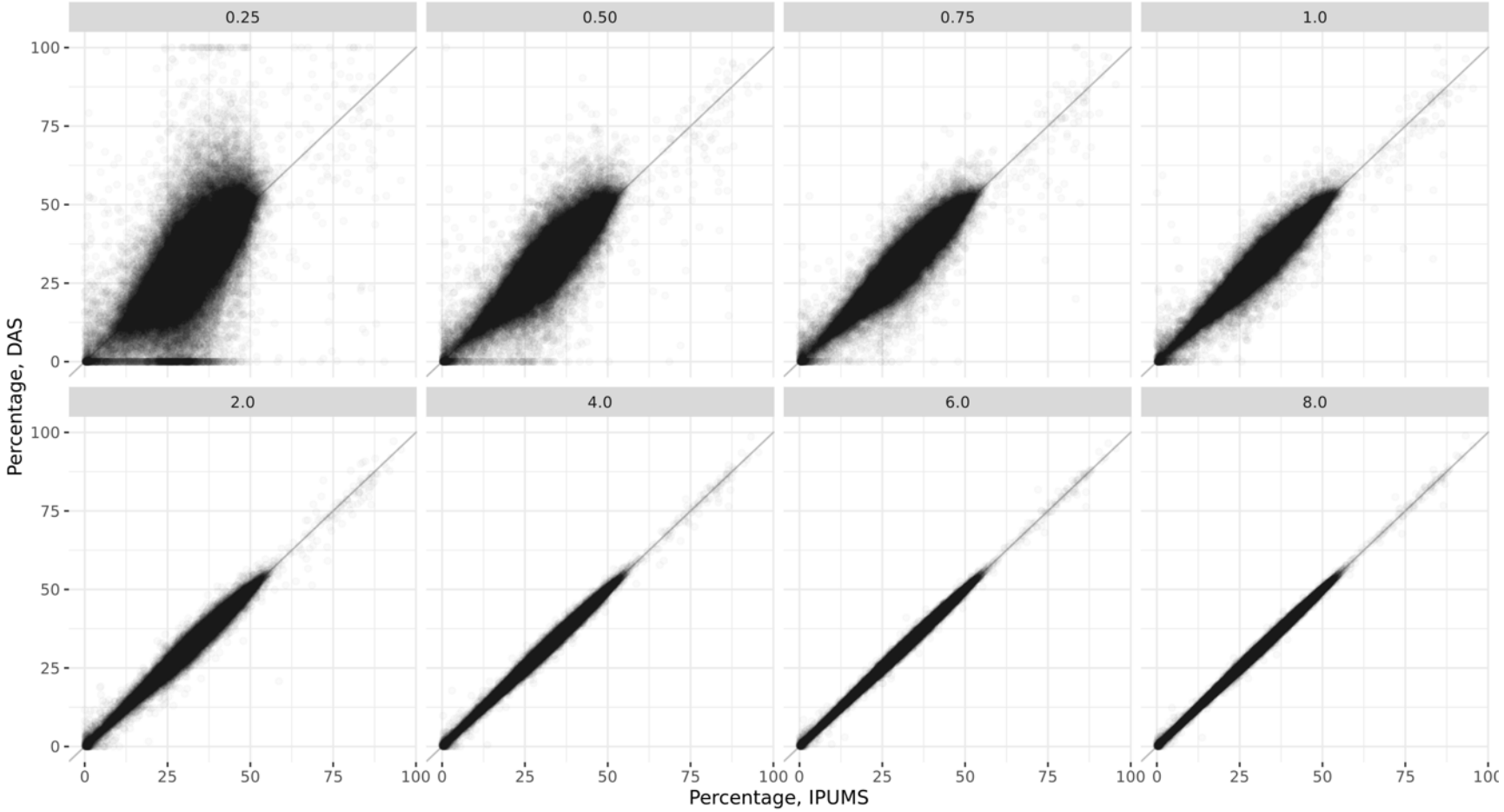
EDs containing 50 or more persons



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

Percentage of population under age 18

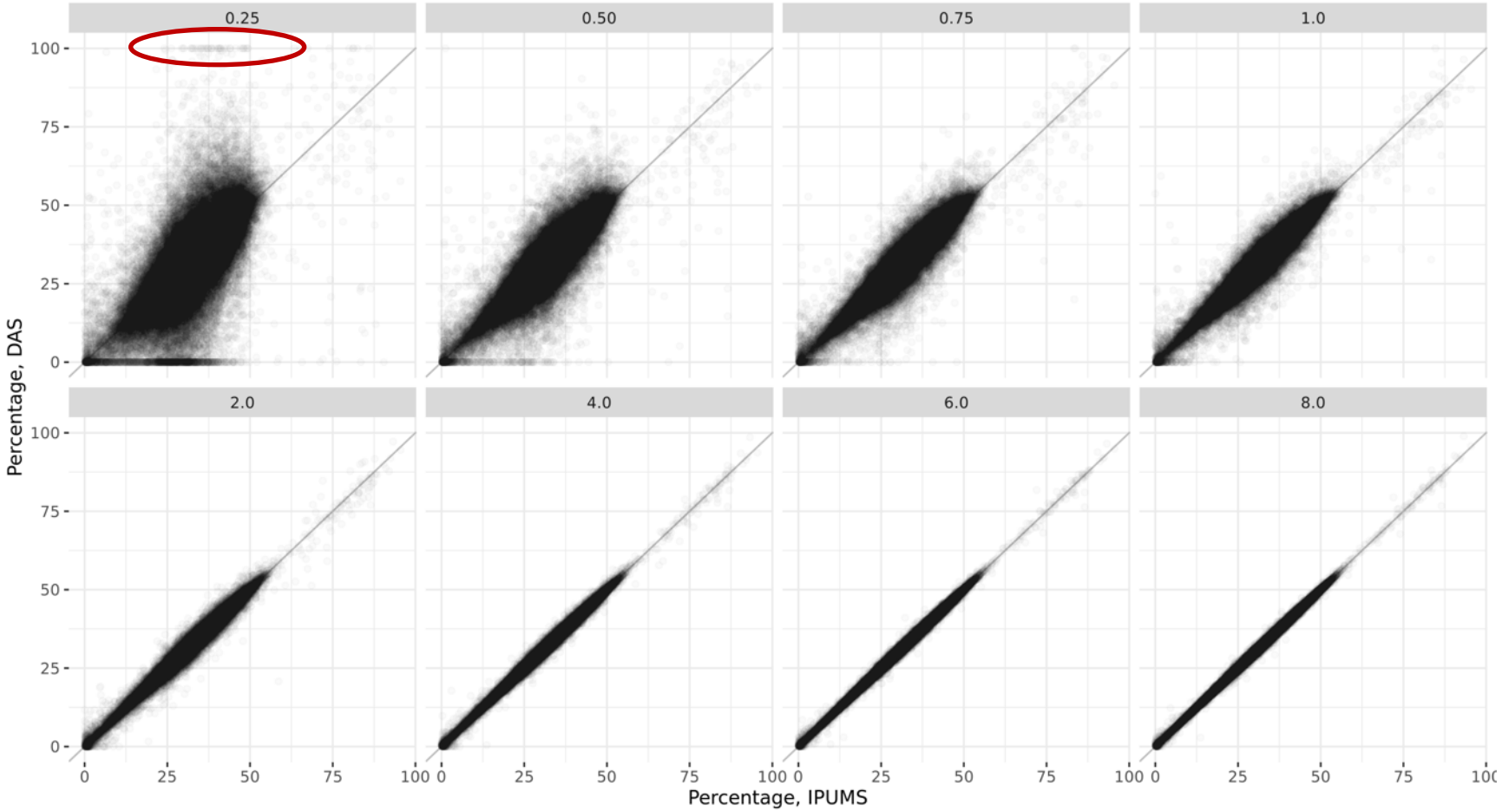
EDs containing 250 or more persons



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

Percentage of population under age 18

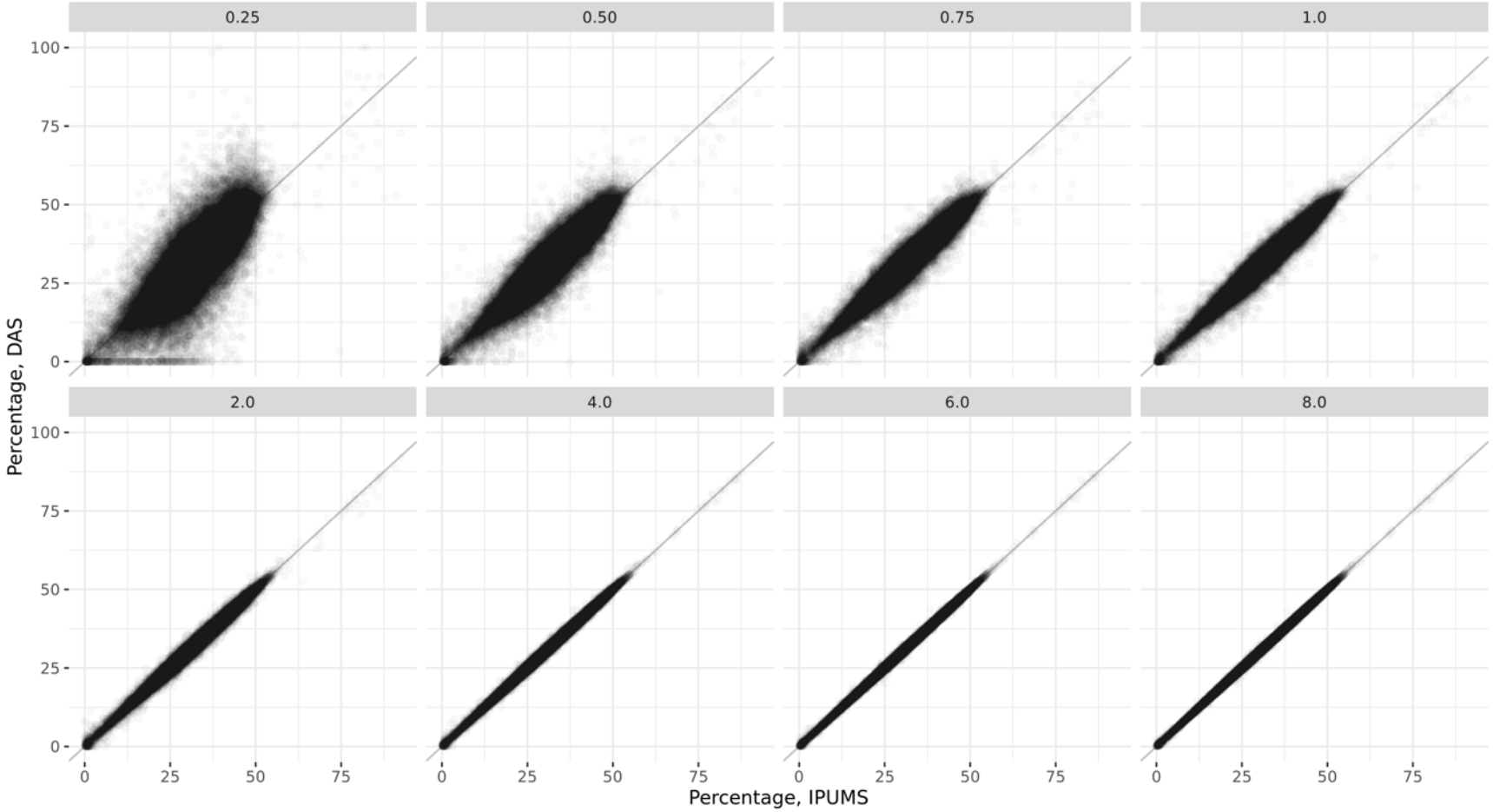
EDs containing 250 or more persons



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

Percentage of population under age 18

EDs containing 500 or more persons

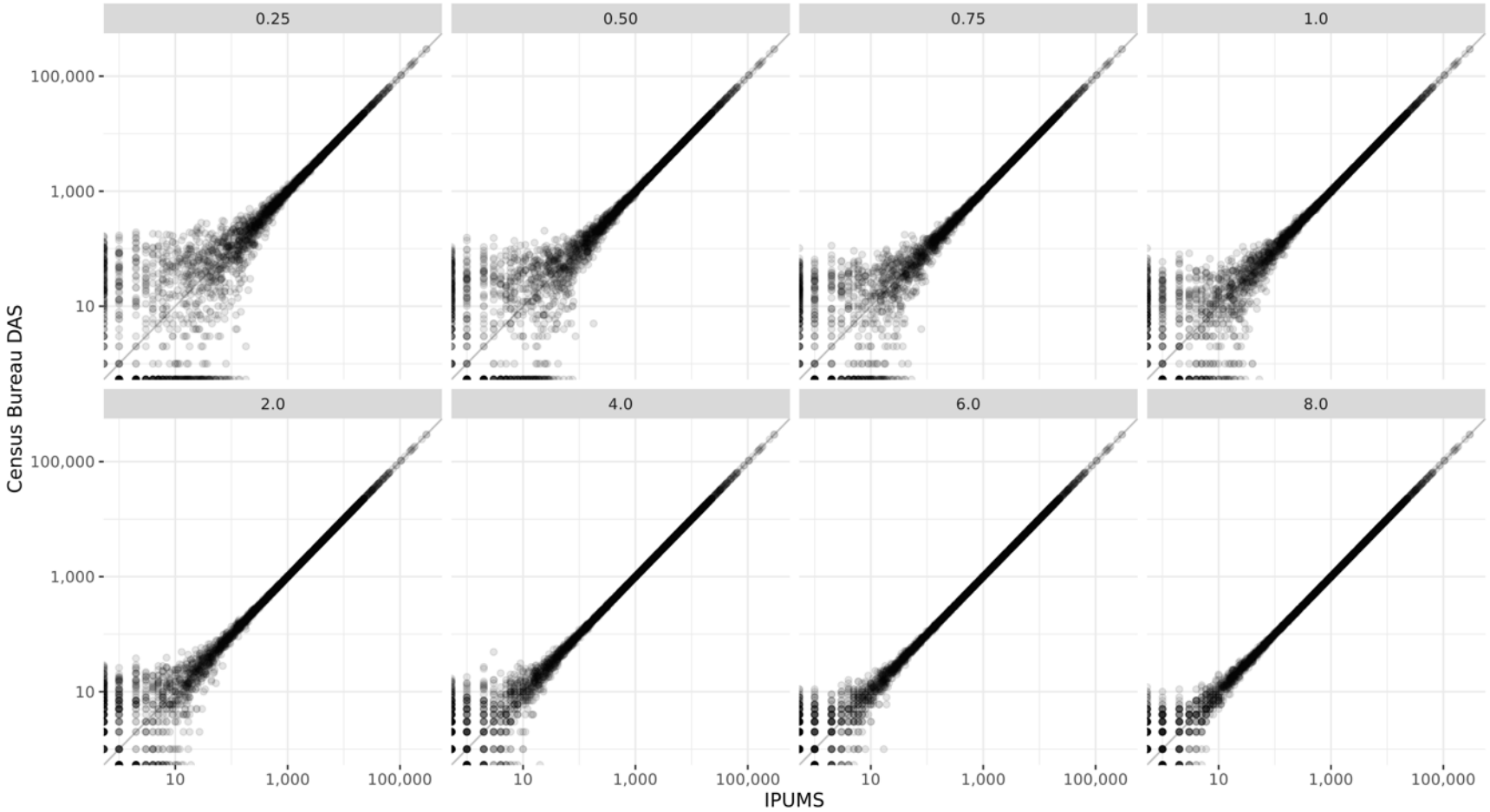


Source: Ruggles et al. (2018); US Census Bureau (2019)

AFRICAN AMERICAN POPULATION

African American population under different levels of noise injection

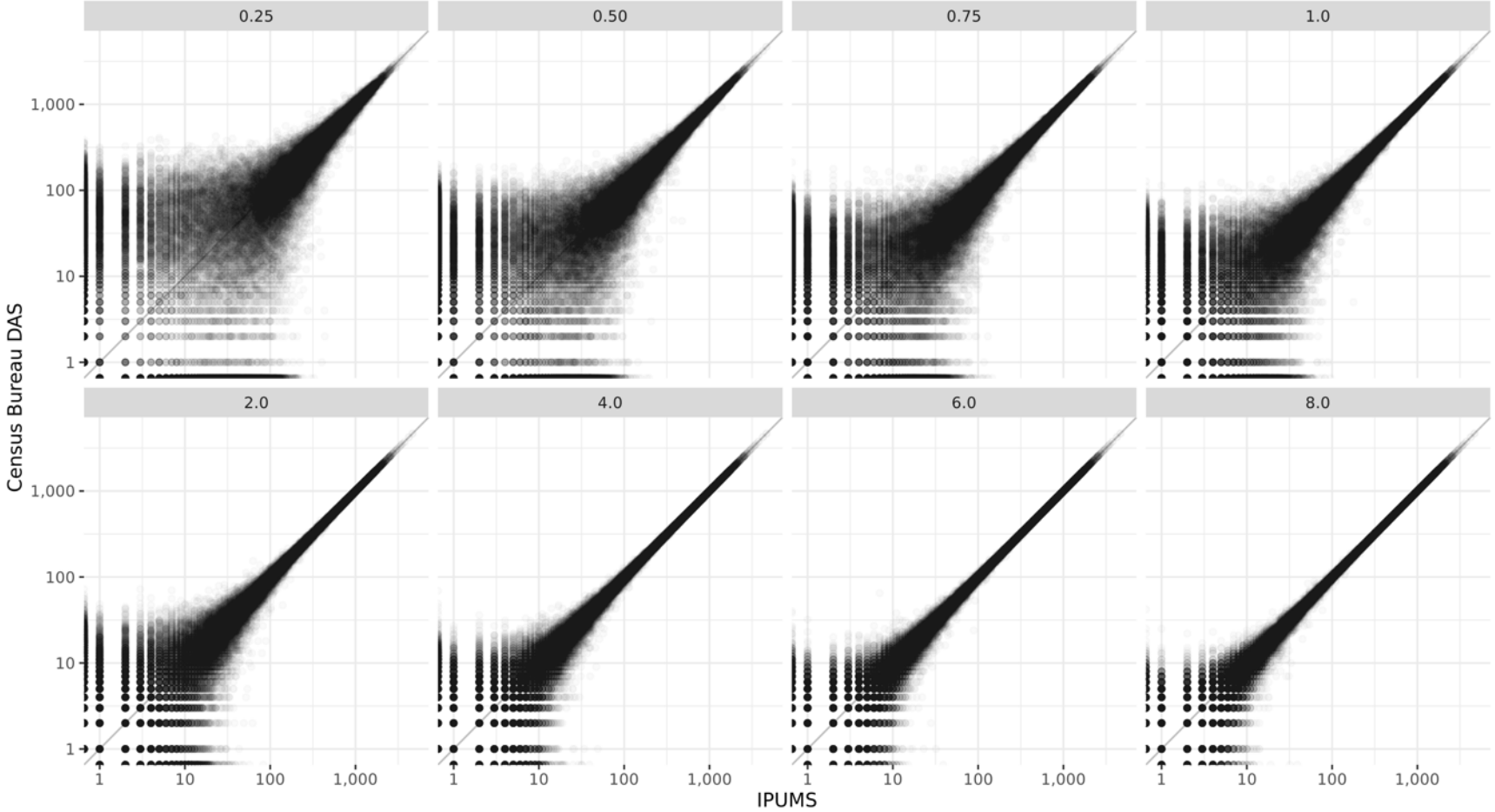
US counties



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

African American population under different levels of noise injection

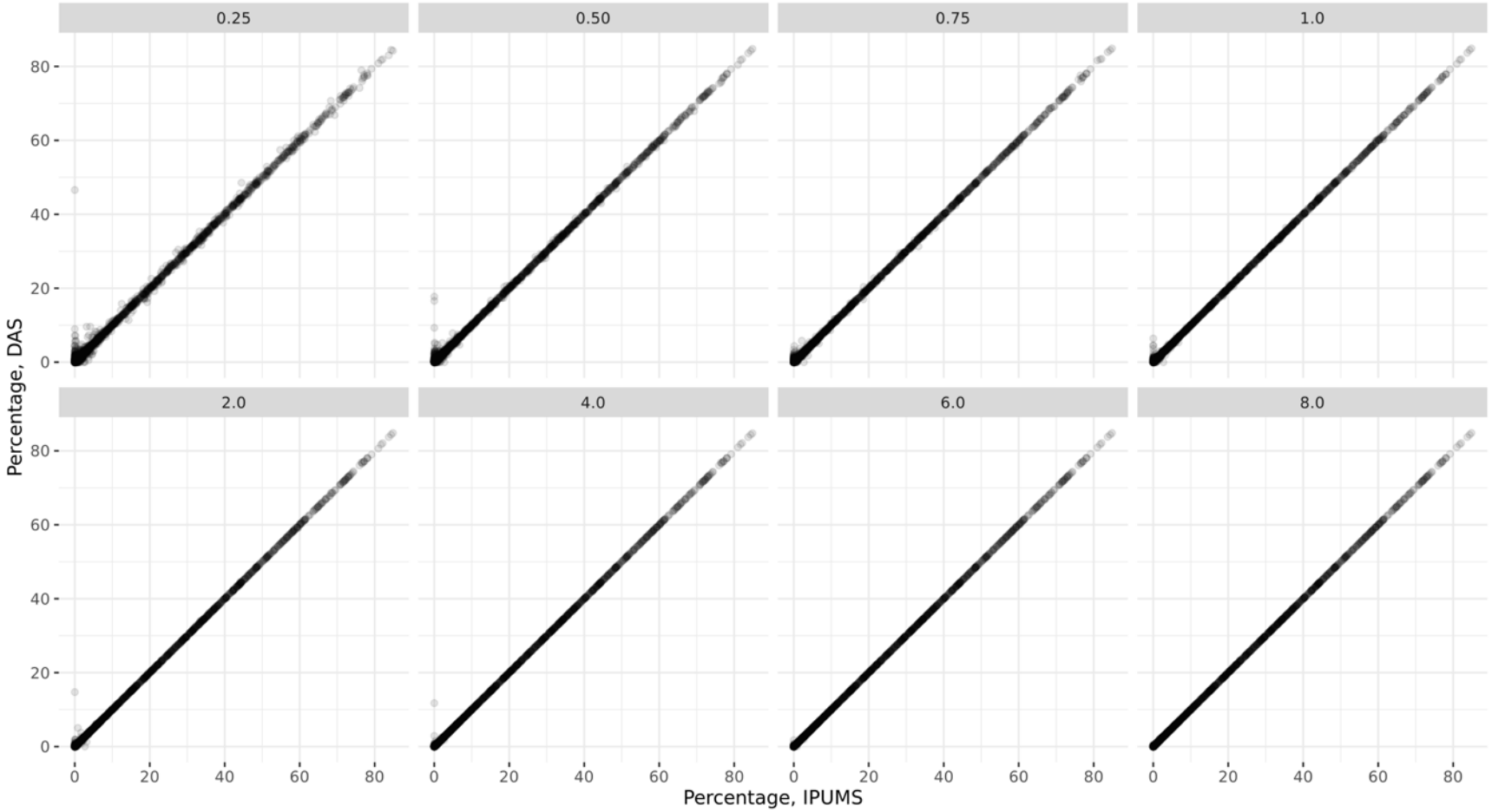
US enumeration districts



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

Percentage of population who is African American

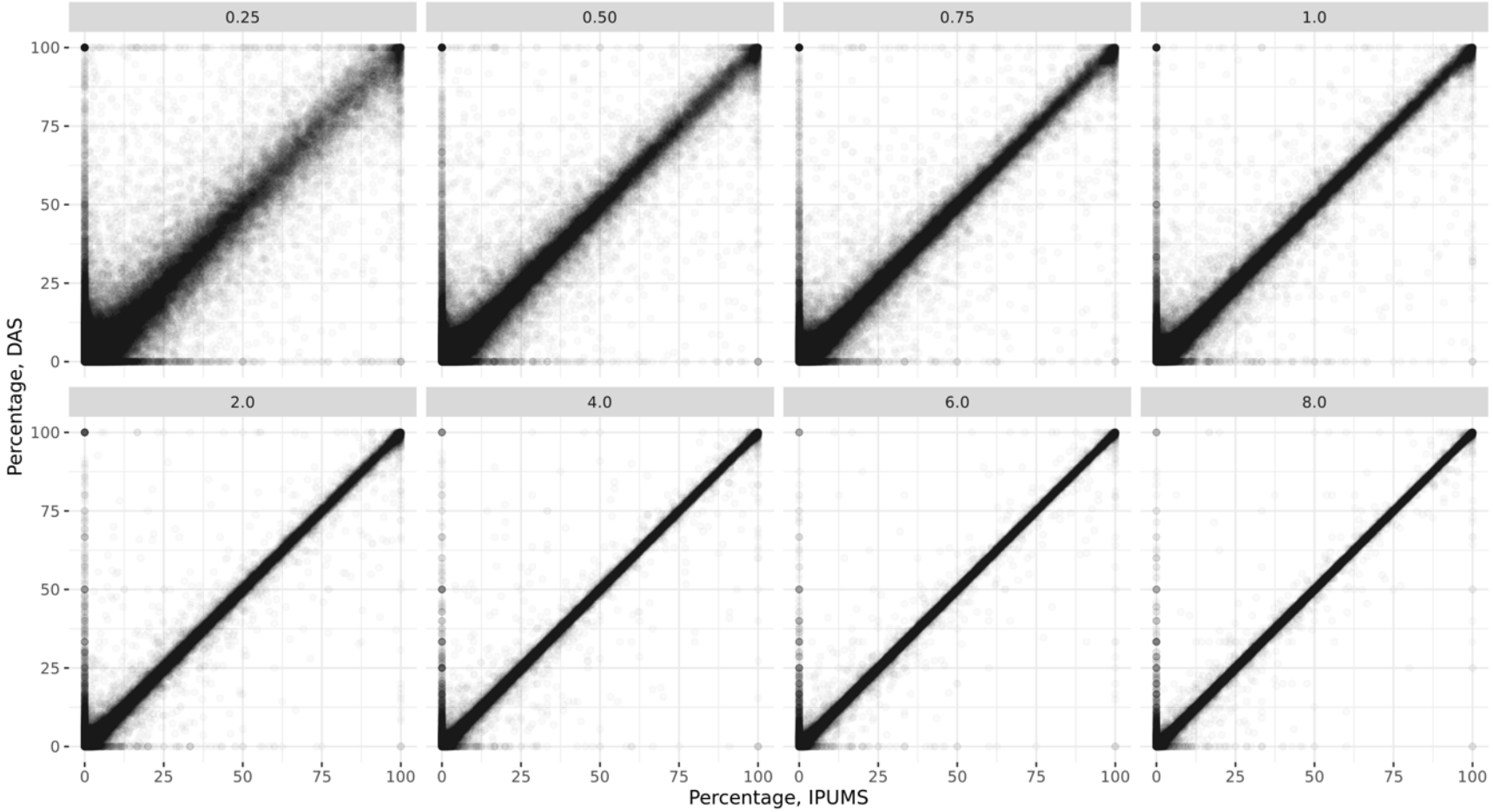
US counties



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

Percentage of population who is African American

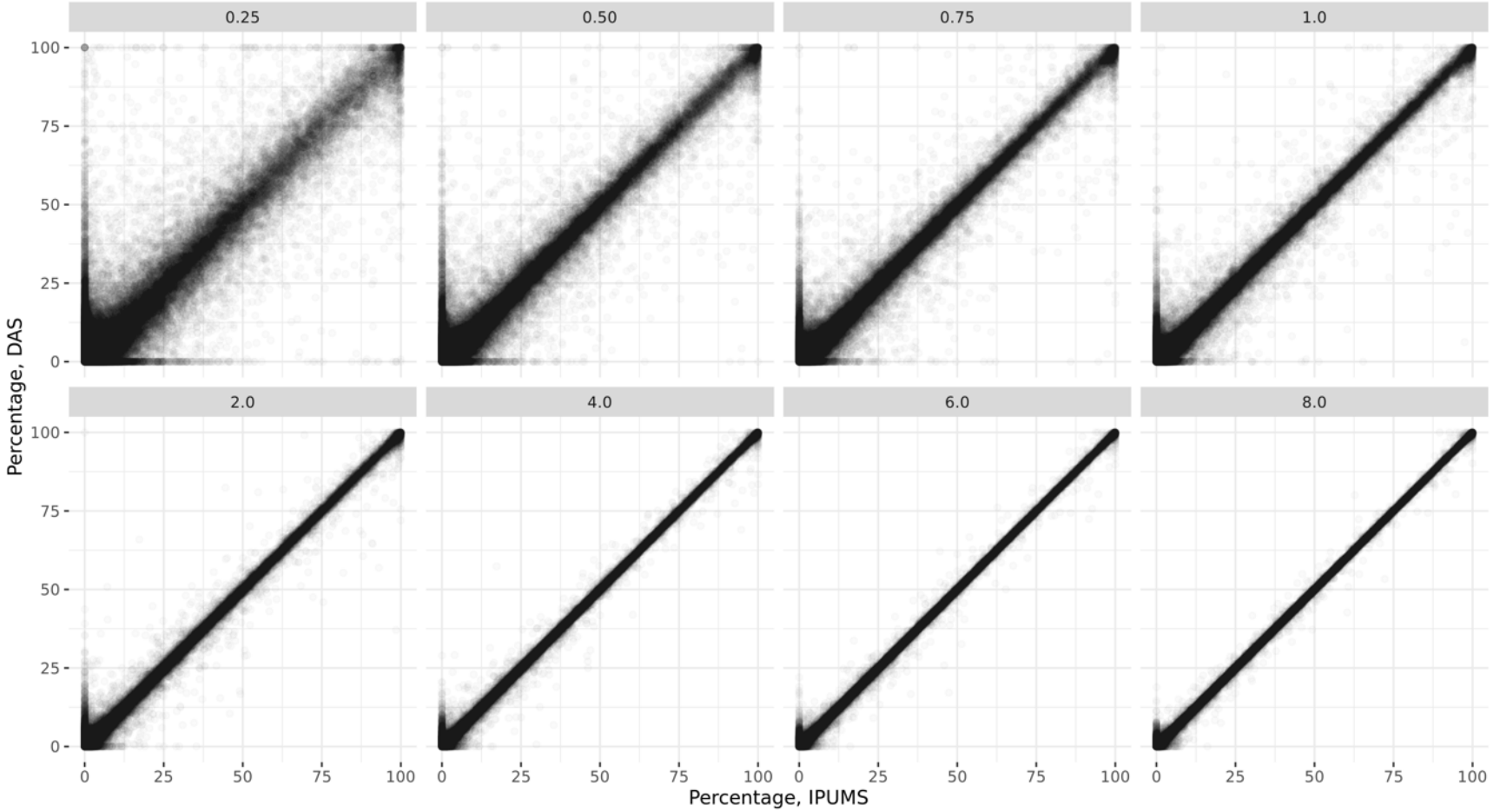
US enumeration districts



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

Percentage of population who is African American

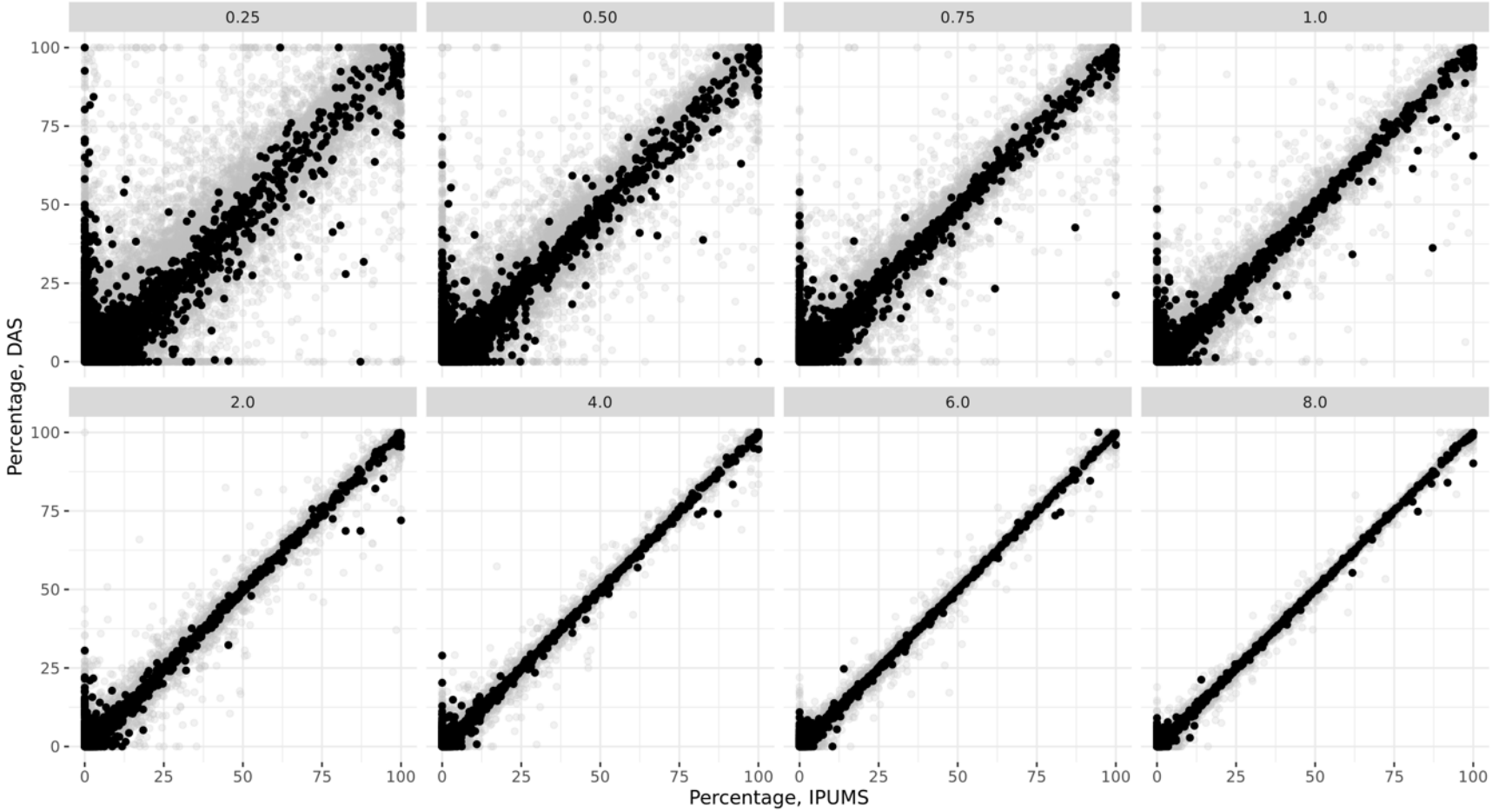
EDs containing 50 or more persons



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

Percentage of population who is African American

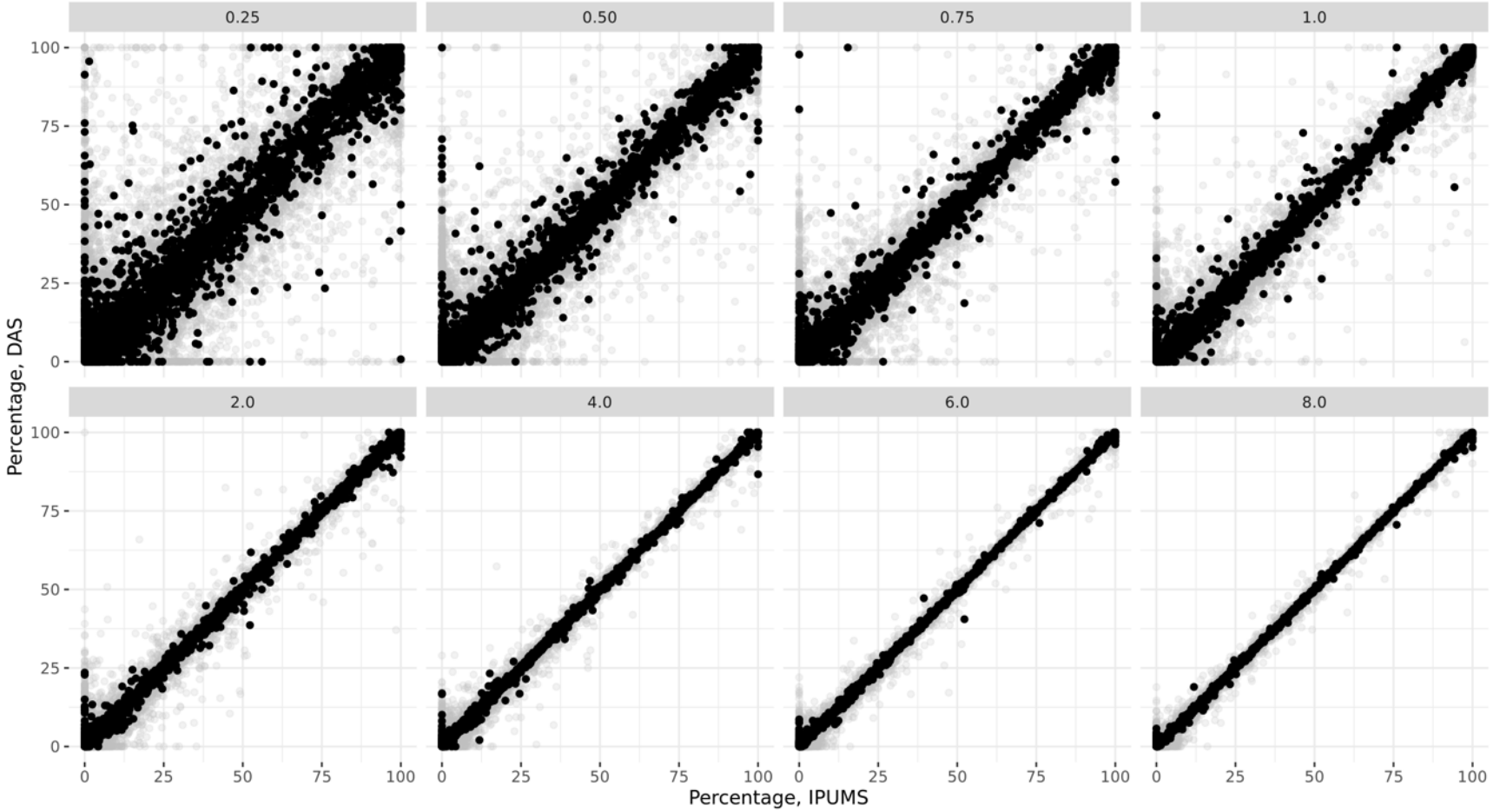
New York EDs containing 50 or more persons



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

Percentage of population who is African American

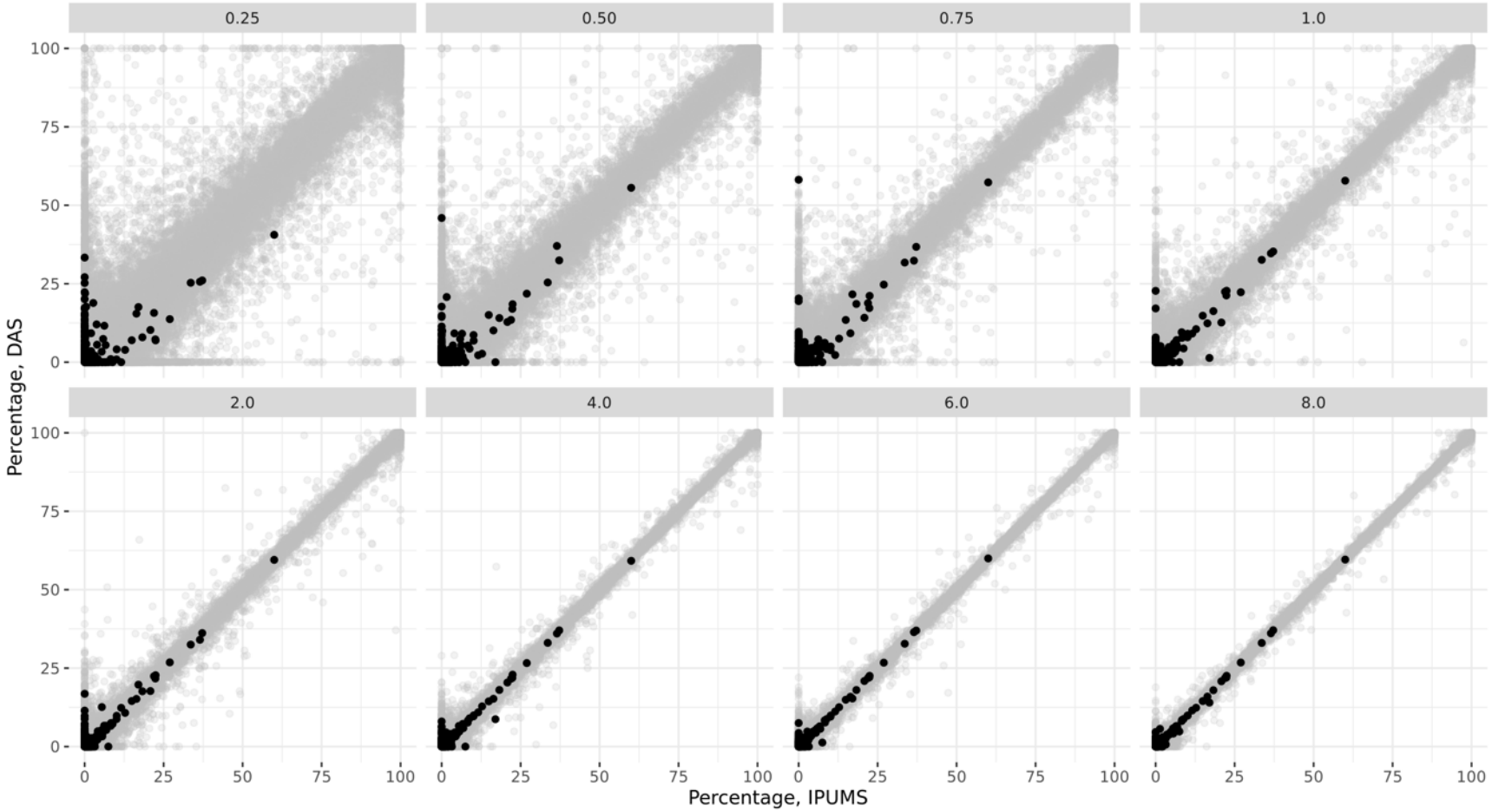
Alabama EDs containing 50 or more persons



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

Percentage of population who is African American

Minnesota EDs containing 50 or more persons



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

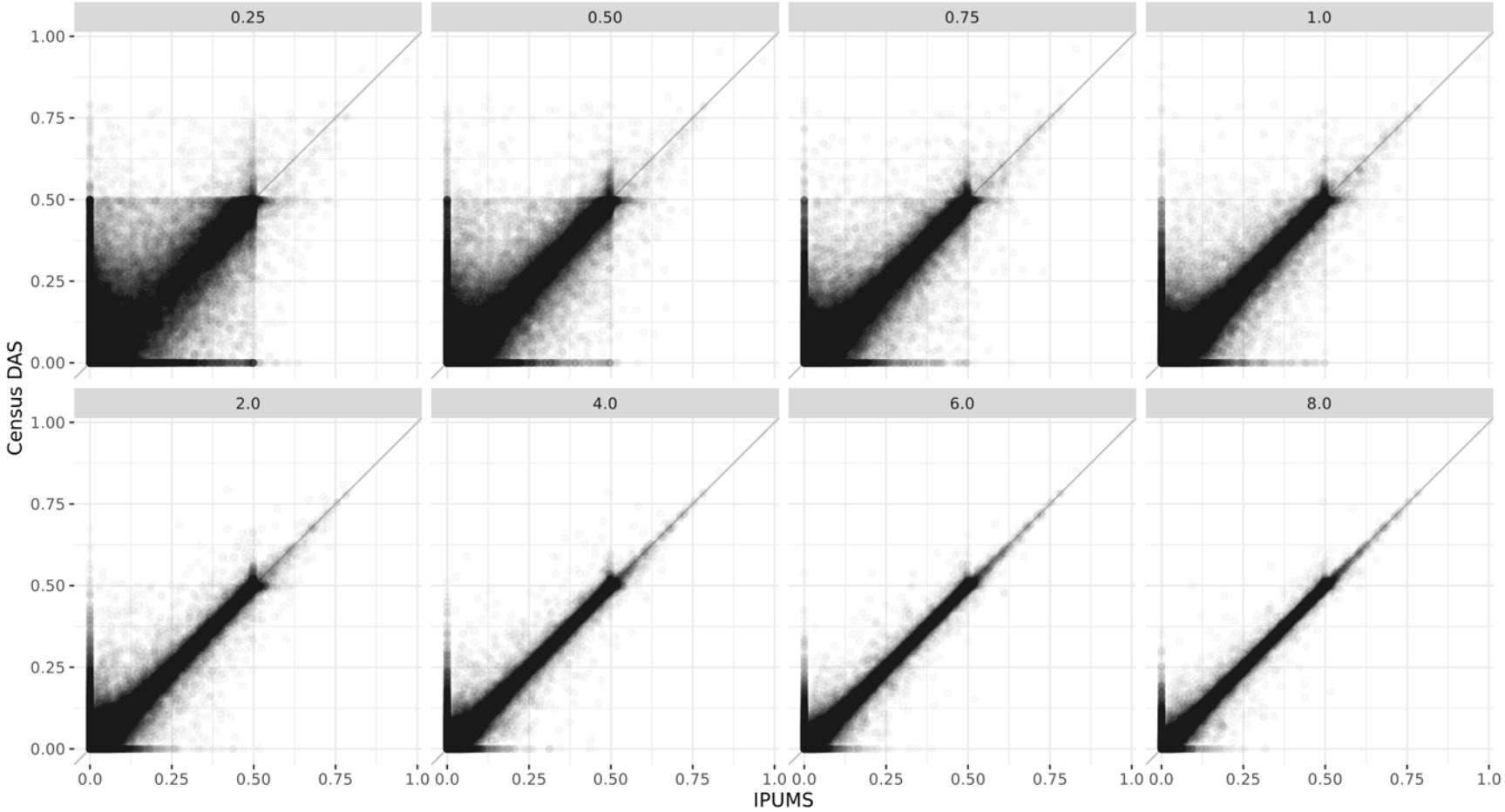
SEGREGATION

$$\mathrm{E}_i = \sum_{r=1}^{r} \pi_{ri} * ln\left(\frac{1}{\pi_{ri}}\right)$$

Iceland (2004)

Four-group diversity index under different levels of noise injection

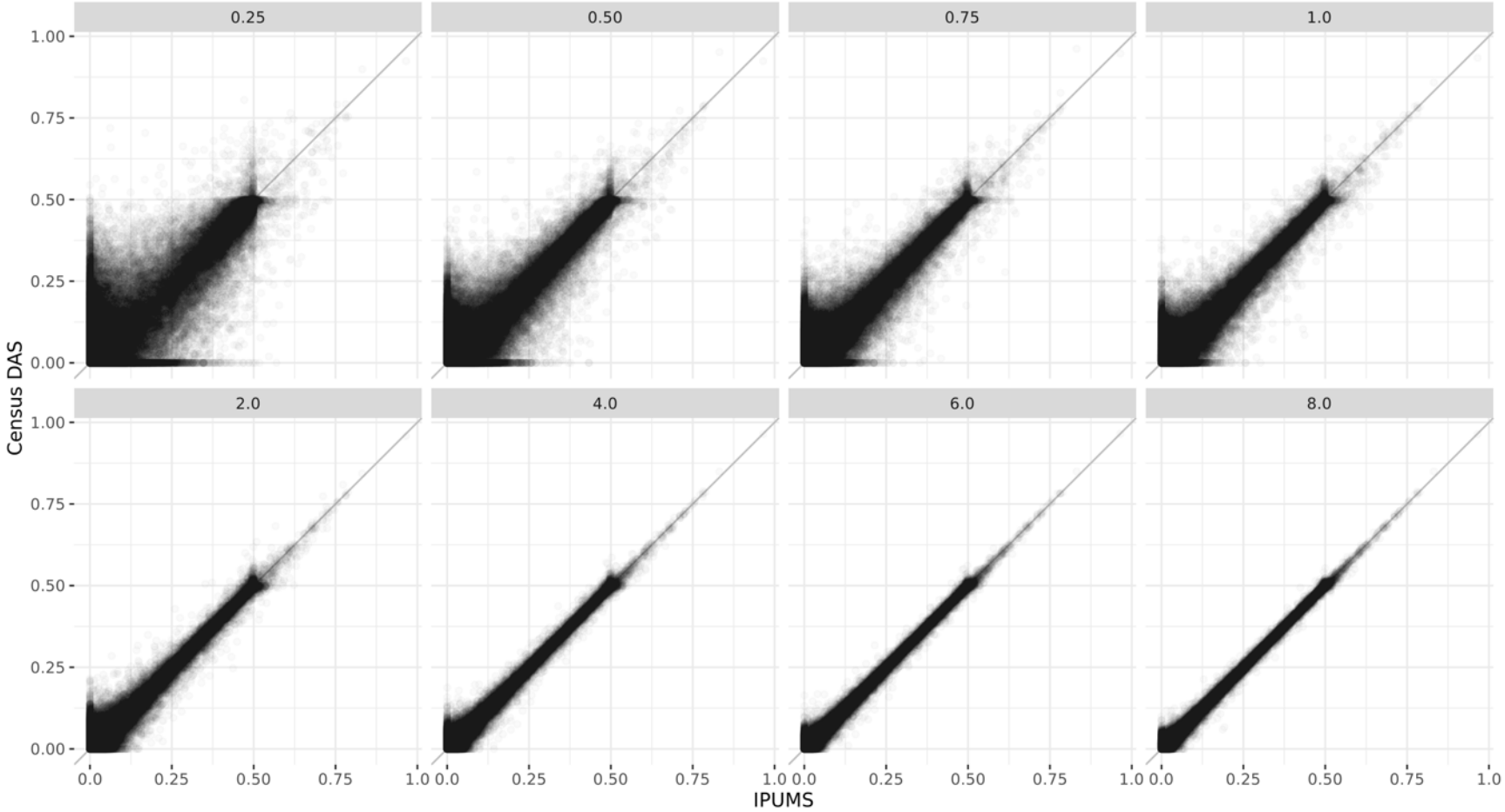
EDs containing 50 or more persons



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

Four-group diversity index under different levels of noise injection

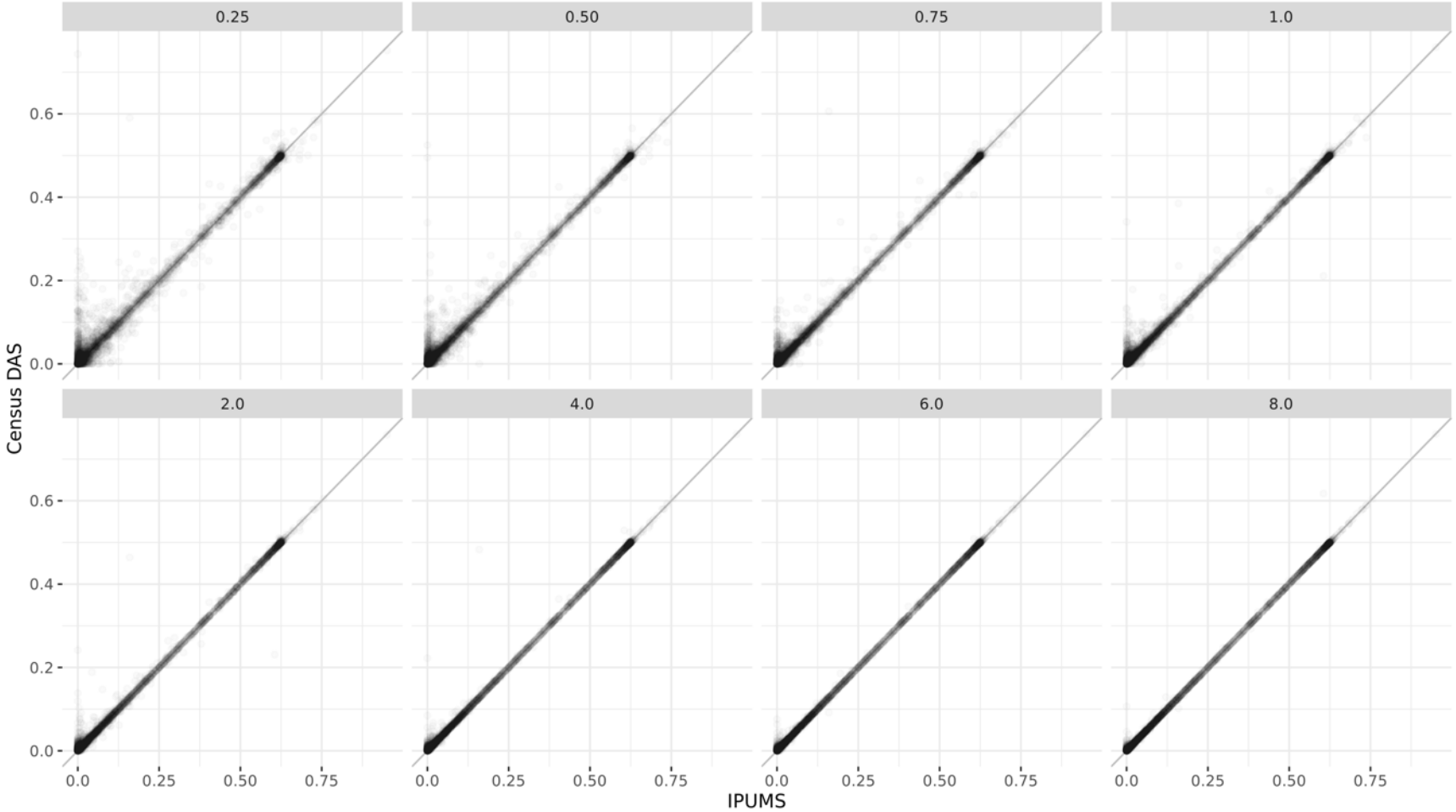
EDs containing 500 or more persons



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

Four-group diversity index under different levels of noise injection

US counties

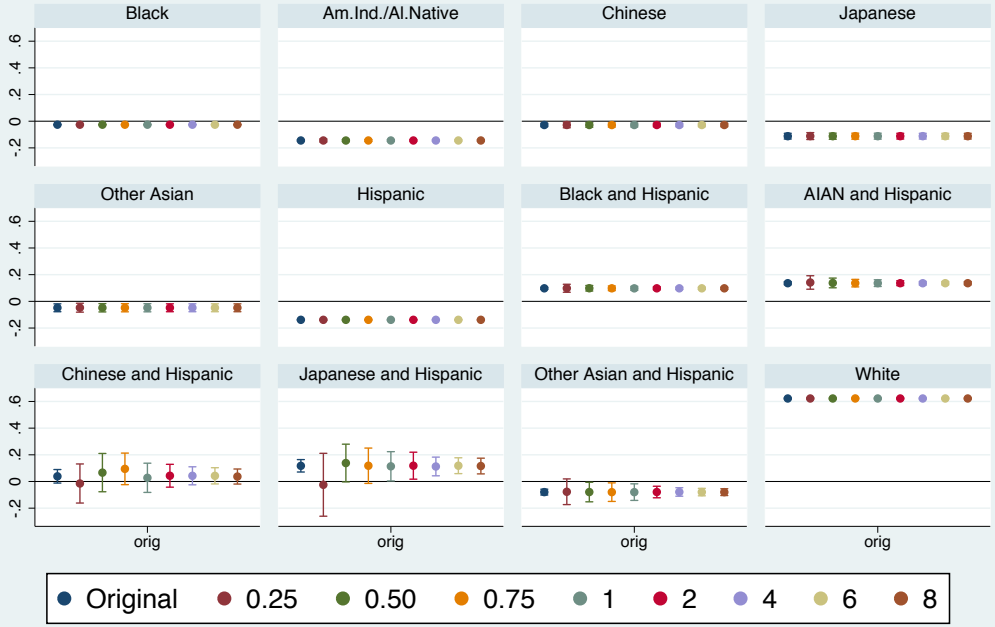


REGRESSIONS USING MICRODATA

We run the following regression for each individual state and each level of epsilon:

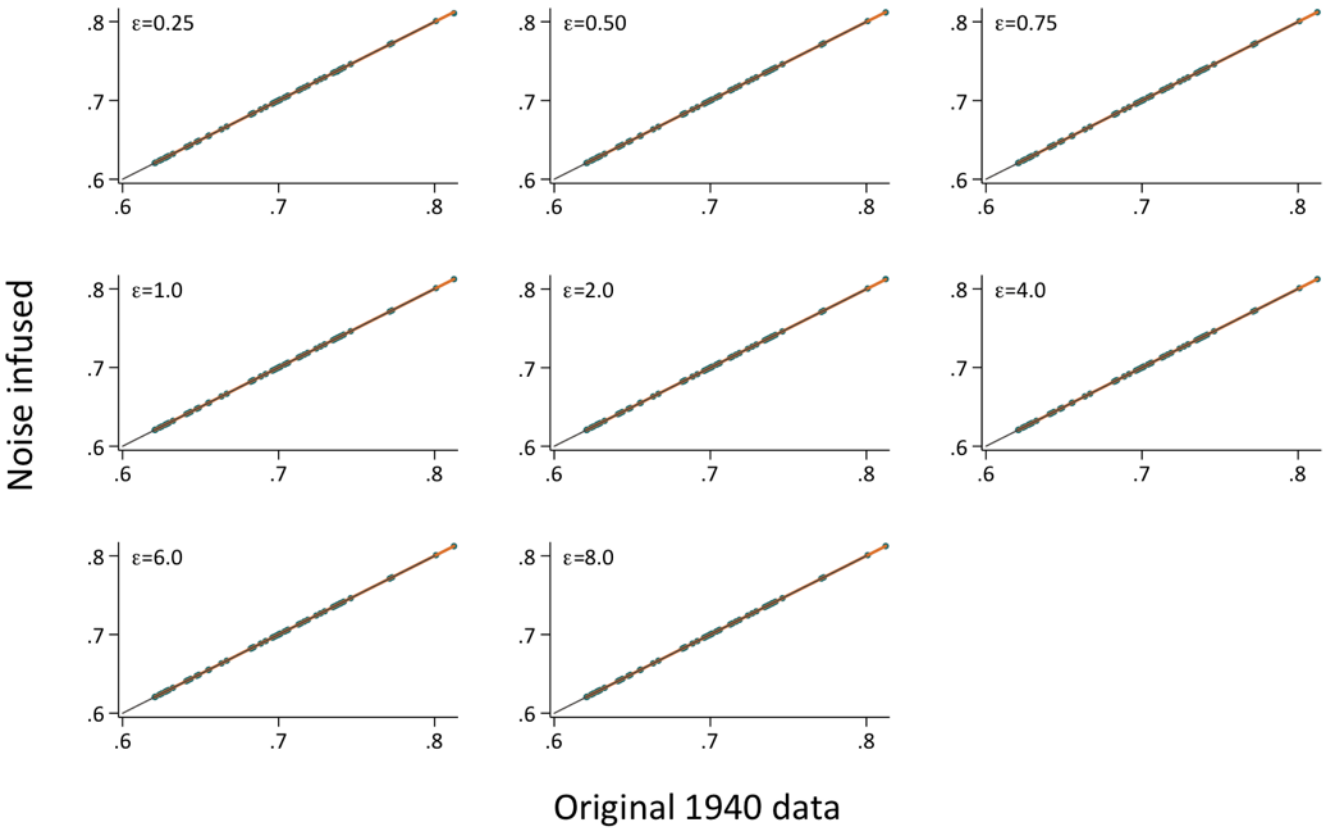
$$Voting\ Age = \beta_0 + \beta_1 Black + \beta_2 AIAN + \beta_3 Chinese + \beta_4 Japanese + \beta_5 OtherAsian + \beta_6 Hispanic + \boldsymbol{\beta} Hispanic * (Black + AIAN + Chinese + Japanese + OtherAsian)$$

Coefficients from Regression Model with Different Epsilons



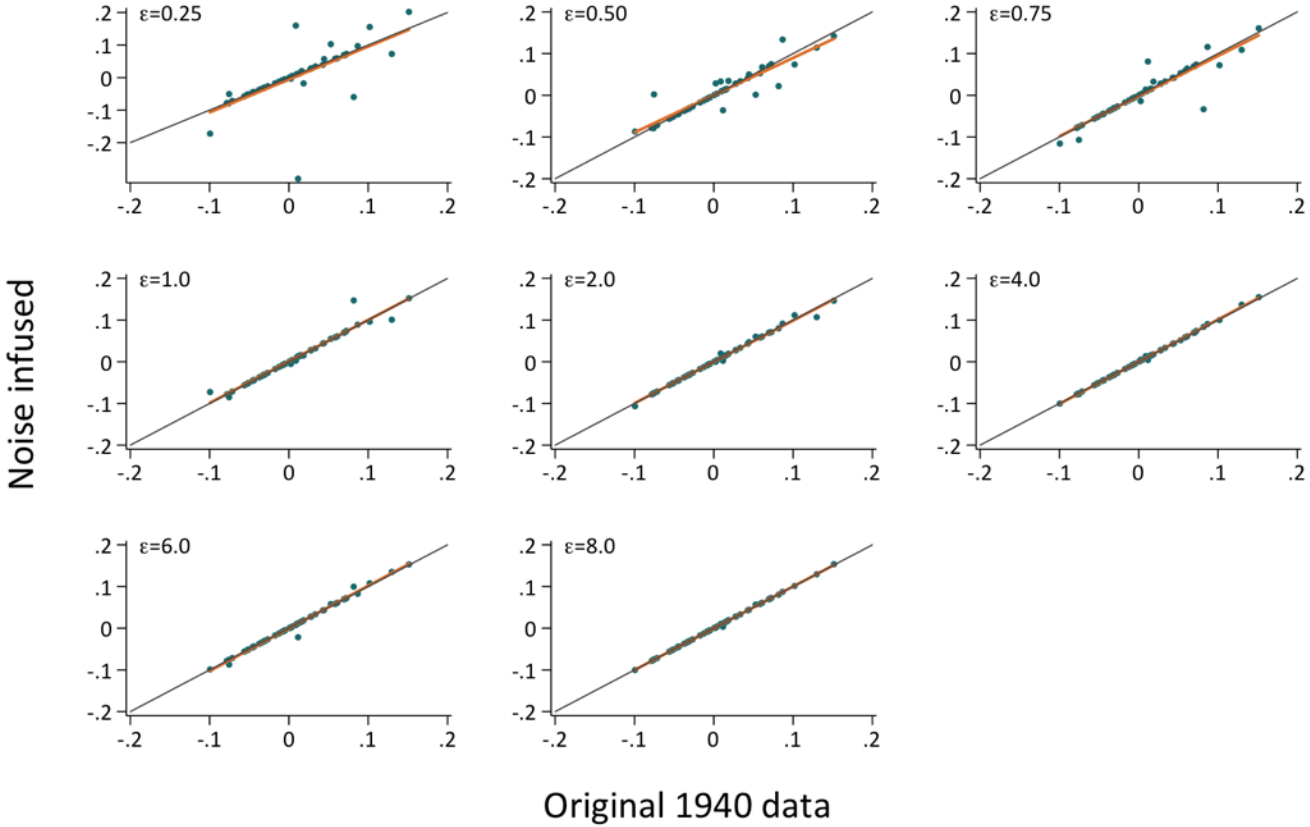
Note. Data source: IPUMS

Probability of Being of Voting Age - White (n=118,803,172)



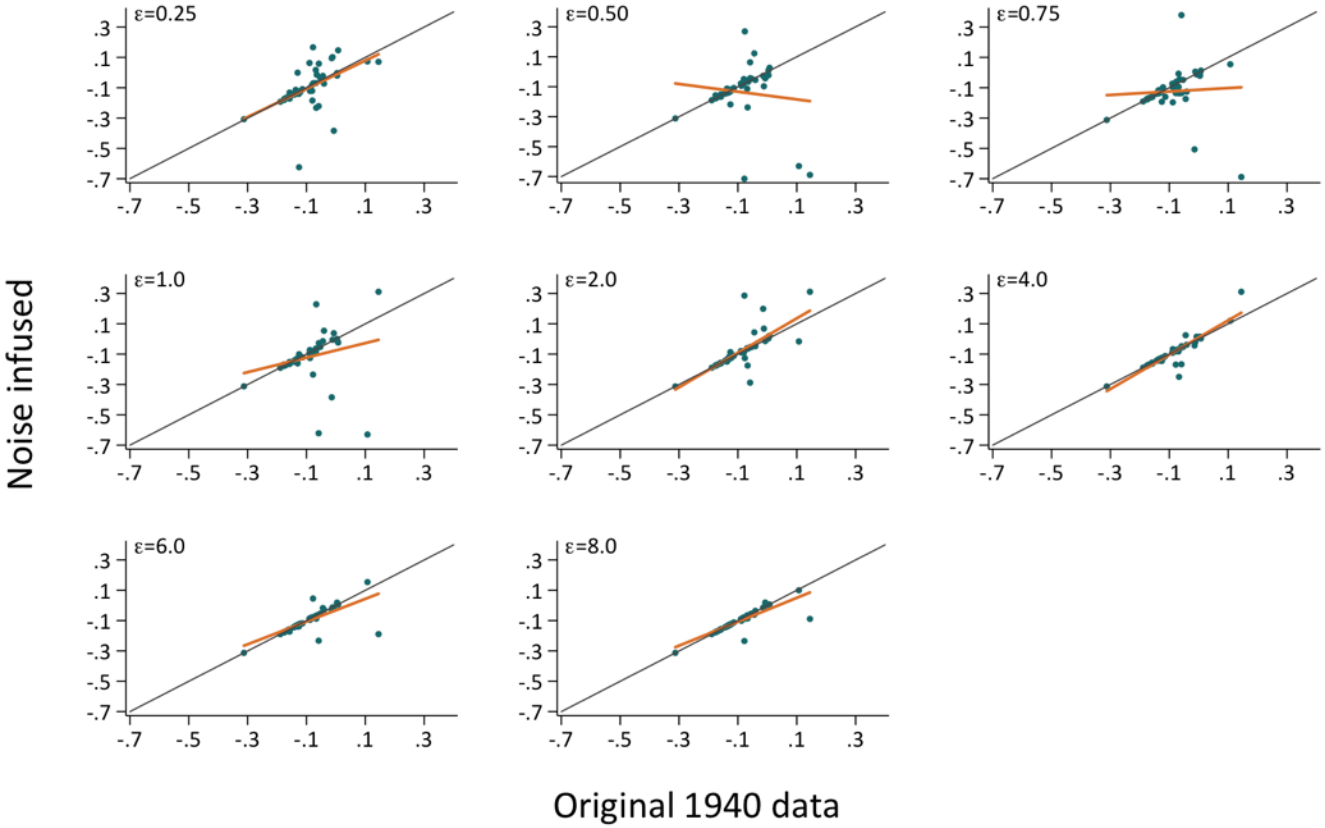
Coefficients from state-specific regressions. Omitted category is White Non-Hispanic. Best-fit line in orange. 45-degree line added for reference.
Voting Age = black+aian+chinese+japanese+other_asian+hispanic+hispanic(black+aian+chinese+japanese+other_asian)*

Probability of Being of Voting Age - Black (n=12,663,581)



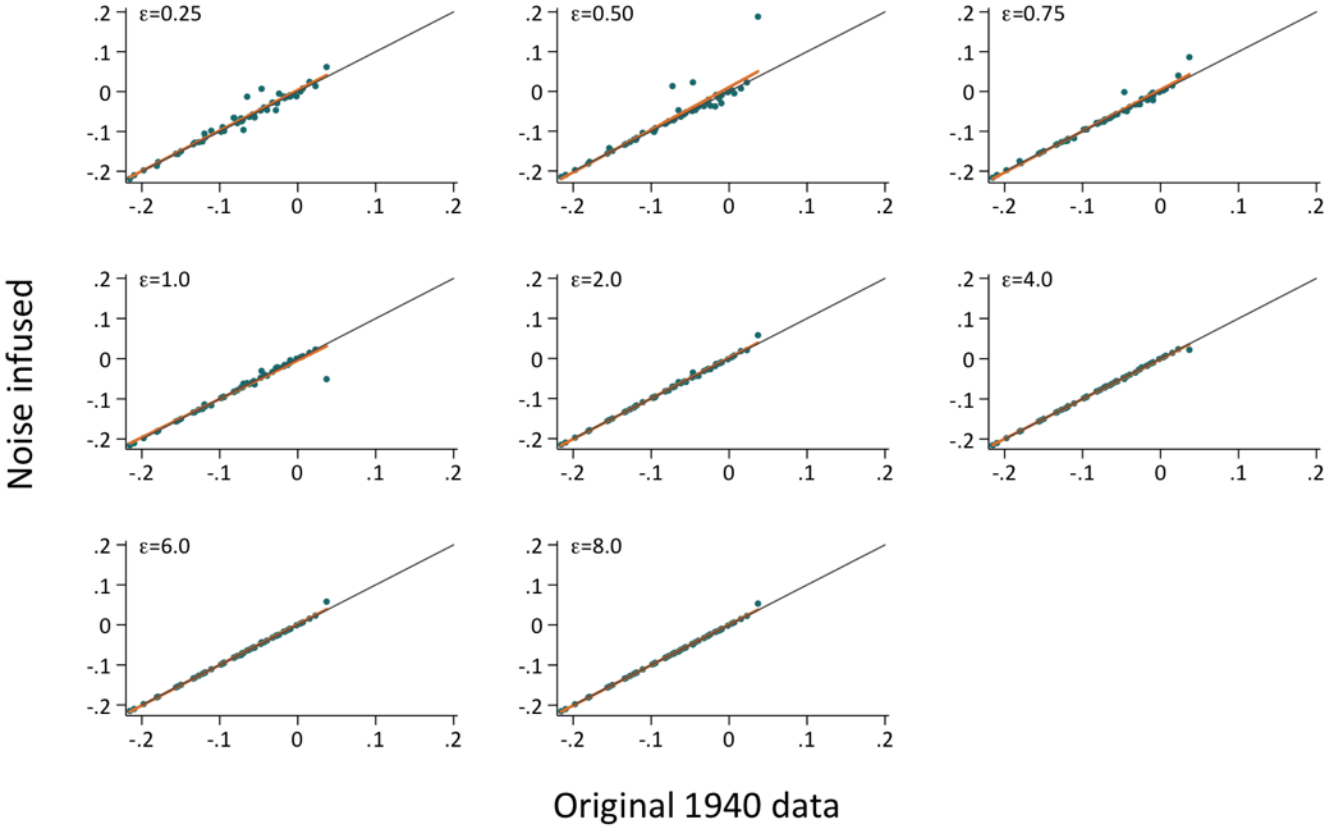
Coefficients from state-specific regressions. Omitted category is White Non-Hispanic. Best-fit line in orange. 45-degree line added for reference.
Voting Age = black+aian+chinese+japanese+other_asian+hispanic+hispanic(black+aian+chinese+japanese+other_asian)*

Probability of Being of Voting Age - AI/AN (n=340,744)



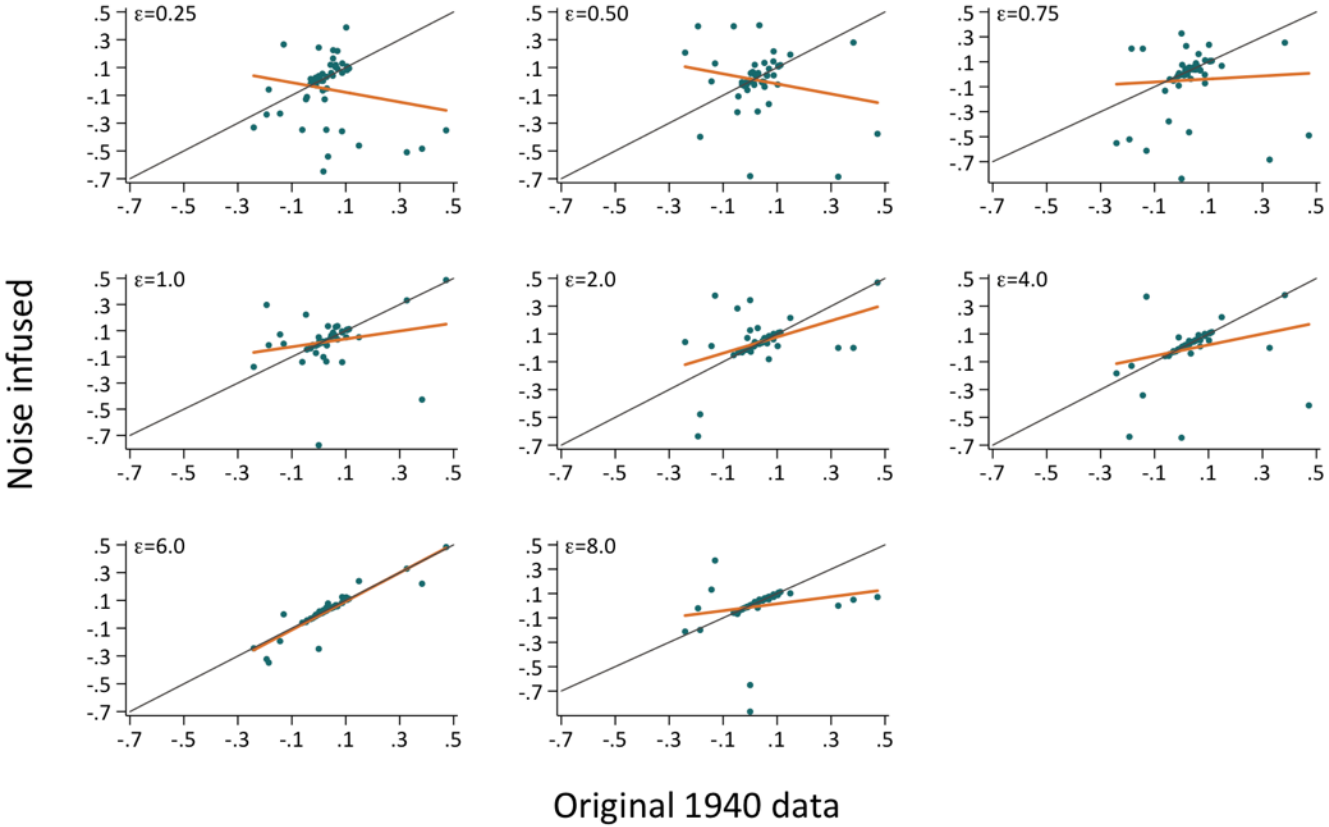
Coefficients from state-specific regressions. Omitted category is White Non-Hispanic. Best-fit line in orange. 45-degree line added for reference.
Voting Age = black+aian+chinese+japanese+other_asian+hispanic+hispanic(black+aian+chinese+japanese+other_asian)*

Probability of Being of Voting Age - Hispanic (n=2,202,645)



Coefficients from state-specific regressions. Omitted category is White Non-Hispanic. Best-fit line in orange. 45-degree line added for reference.
Voting Age = black+aian+chinese+japanese+other_asian+hispanic+hispanic(black+aian+chinese+japanese+other_asian)*

Probability of Being of Voting Age - Black & Hisp. (n=58,074)



Coefficients from state-specific regressions. Omitted category is White Non-Hispanic. Best-fit line in orange. 45-degree line added for reference.
Voting Age = black+aian+chinese+japanese+other_asian+hispanic+hispanic(black+aian+chinese+japanese+other_asian)*

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

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. 2018. "Disclosure Avoidance for Block Level Data and Protection of Confidentiality in Public Tabulations." US Census Bureau. https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#.

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.

Ashmead, Robert, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. 2019. "Effective Privacy after Adjusting for Invariants with Applications to the 2020 Census." US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0941_Effective_Privacy_after_Adjusting_for_Constraints__With_applications_to_the_2020_Census.pdf.

Bambauer, Jane, Krishnamurty Muralidhar, and Rathindra Sarathy. n.d. "Fool's Gold: An Illustrated Critique of Differential Privacy." *Vanderbilt Journal of Entertainment & Technology Law* 16: 55.

boyd, danah. 2019. "Differential Privacy in the 2020 Decennial Census and the Implications for Available Data Products." *ArXiv:1907.03639 [Cs]*, July. http://arxiv.org/abs/1907.03639.

Census Bureau, US. 2019. “2018 End-to-End Test Disclosure Avoidance System Design Specification, Version 1.2.8.” US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/2019-04-11-2018-End-to-End-Test-Disclosure-Avoidance-System-Design-Specification.pdf.

———. n.d. “Differentially Private 1940 Census Data.” US Census Bureau. https://www2.census.gov/census_1940/.

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. “Issues Encountered Deploying Differential Privacy.” In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–137. WPES’18. New York, NY, USA: ACM. https://doi.org/10.1145/3267323.3268949.

Iceland, John. 2004. “The Multigroup Entropy Index (Also Known as Theil’s H or the Information Theory Index).” US Census Bureau. https://www2.census.gov/programs-surveys/demo/about/housing-patterns/multigroup_entropy.pdf.

Leclerc, Philip. 2019. “Guide to the Census 2018 End-to-End Test Disclosure Avoidance Algorithm and Implementation.” US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0938_2018_E2E_Test_Algorithm_Description.pdf.

McClure, David, and Jerome P. Reiter. 2012. “Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data.” *Trans. Data Privacy* 5 (3): 535–552.

Reiter, Jerome P. 2019. “Differential Privacy and Federal Data Releases.” *Annual Review of Statistics and Its Application* 6 (1): 85–101. https://doi.org/10.1146/annurev-statistics-030718-105142.

Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. “Differential Privacy and Census Data: Implications for Social and Economic Research.” *AEA Papers and Proceedings* 109 (May): 403–8. https://doi.org/10.1257/pandp.20191107.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2018. “IPUMS USA: Version 8.0 Extract of 1940 Census for U.S. Census Bureau Disclosure Avoidance Research [Dataset].” IPUMS. https://doi.org/10.18128/D010.V8.0.EXT1940USCB.

Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O’Brien, Thomas Steinke, and Salil Vadhan. 2018. “Differential Privacy: A Primer for a Non-Technical Audience.” *Vanderbilt Journal of Entertainment & Technology Law*, no. 1 (2019): 209–76.