IHME

Differential Privacy in 2020 Decennial Census: anticipating impact on evidence-based public policy at the state and local level

Abraham D. Flaxman

Dec 11, 2019

UNIVERSITY of WASHINGTON

Institute for Health Metrics and Evaluation

Disclosures and acknowledgements

Grant and research support: Bill and Melinda Gates Foundation; Alfred P. Sloan Foundation.

Consulting: Kaiser Permanente; Sanofi; Merck for Mothers; Agathos, Ltd (startup); and NORC (formerly National Opinion Research Council).

Thanks: Jan Vink, Cornell Institute for Social and Economic Research
--- your reformatting made my analyses much easier!

Sam Petti, Georgia Tech; David Van Riper, IPUMS;

Mike Mohrman, WA State OFM.

My argument

Thesis: Let's make **Total Count** *invariant* at **block level**

Antitheses: (1) Is this sufficiently private?

(2) Will this compromise accuracy of other statistics?

Synthesis: …

Data, capacity-building, and training needs to address rural health inequities in the Northwest United States: a qualitative study

Betty Bekemeier ✉, Seungeun Park, Uba Backonja, India Ornelas, Anne M Turner

Journal of the American Medical Informatics Association, Volume 26, Issue 8-9, August/September 2019, Pages 825–834, https://doi.org/10.1093/jamia/ocz037

Published: 17 April 2019 **Article history** ▾

Abstract

Objective

Rural public health system leaders struggle to access and use data for understanding local health inequities and to effectively allocate scarce resources to populations in need. This study sought to determine these rural public health system leaders' data access, capacity, and training needs.

Summary of Bekemeier et al needs assessment

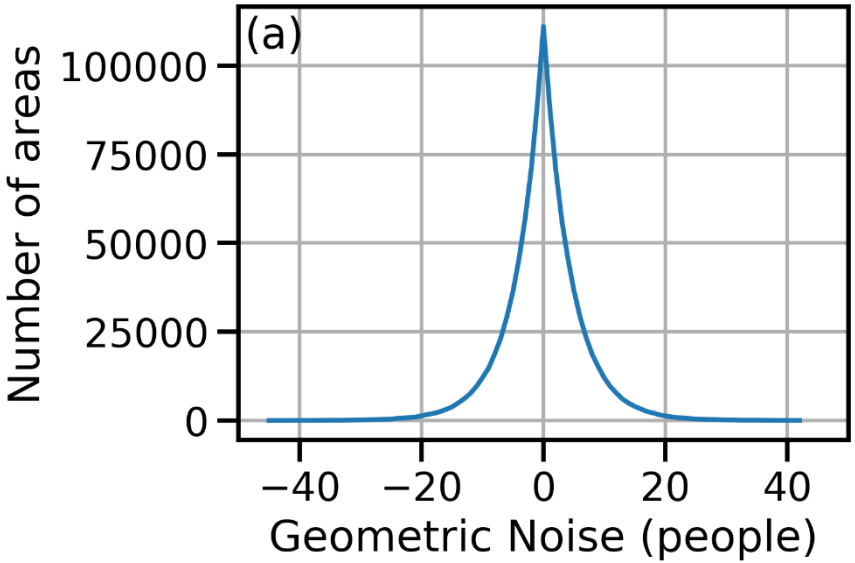
1. Limited availability or access to data
2. **Data quality issues**
3. Limited staff with expertise and resources for analyzing data

Most relevant for us is (2):

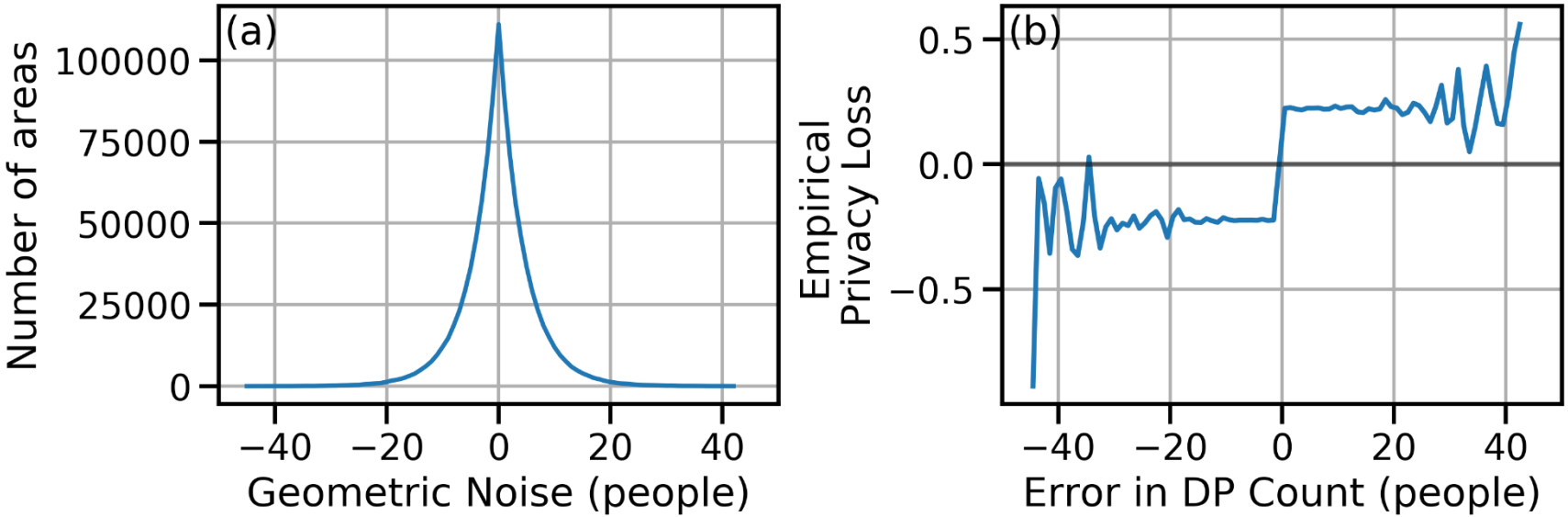
- Data perceived as unreliable or inaccurate were often considered unusable
- Outdated data sets were also a problem

I see opportunity here to address (1) and (3), also, but …

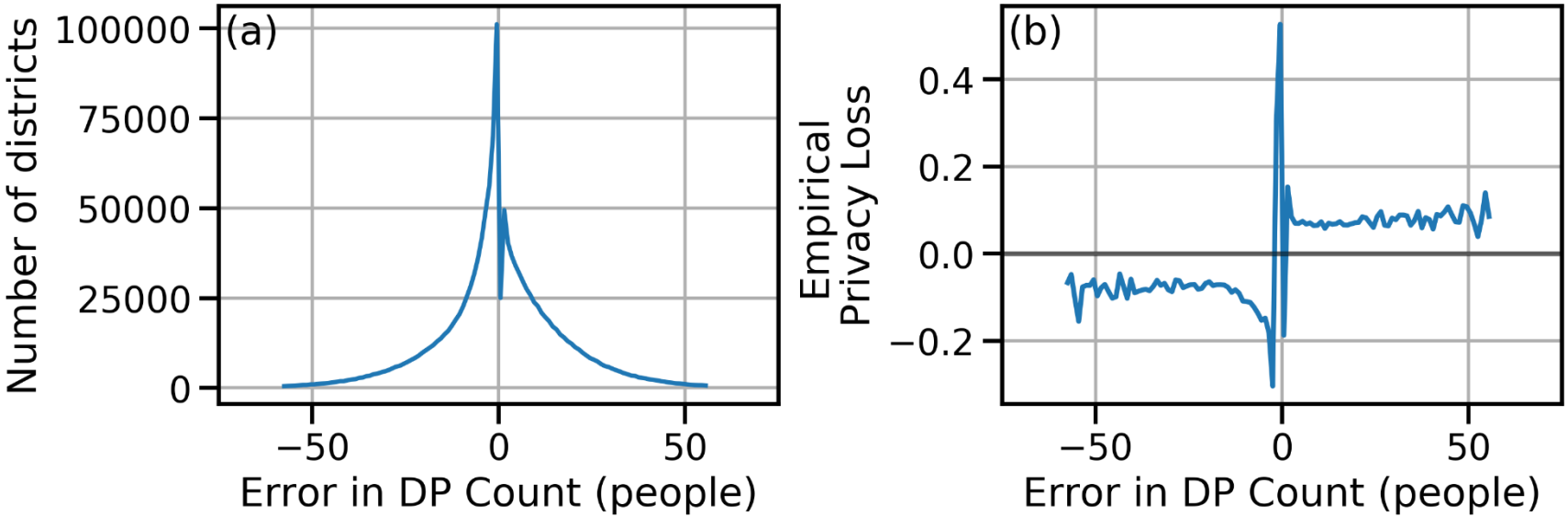
Distribution of variation added to counts (Geometric Mechanism)



Distribution of variation added to counts (Geometric Mechanism)

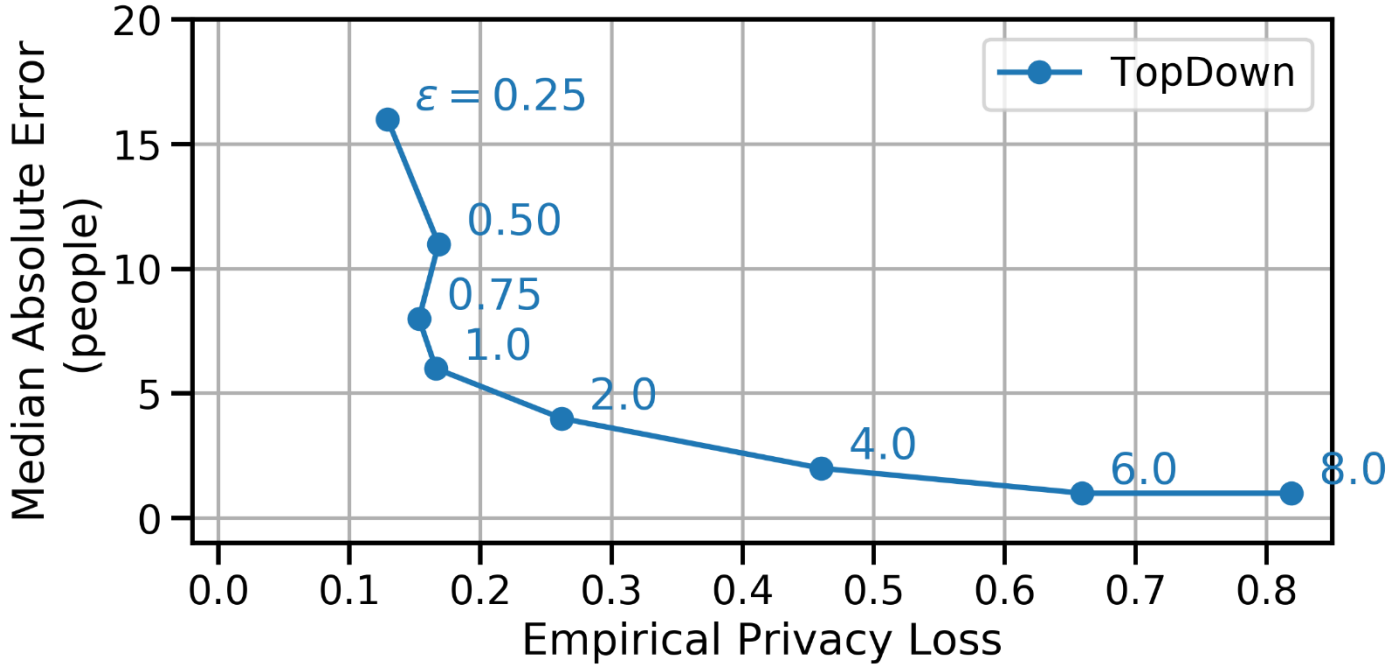


Distribution of variation added to counts (TopDown run on 1940 decennial census data)



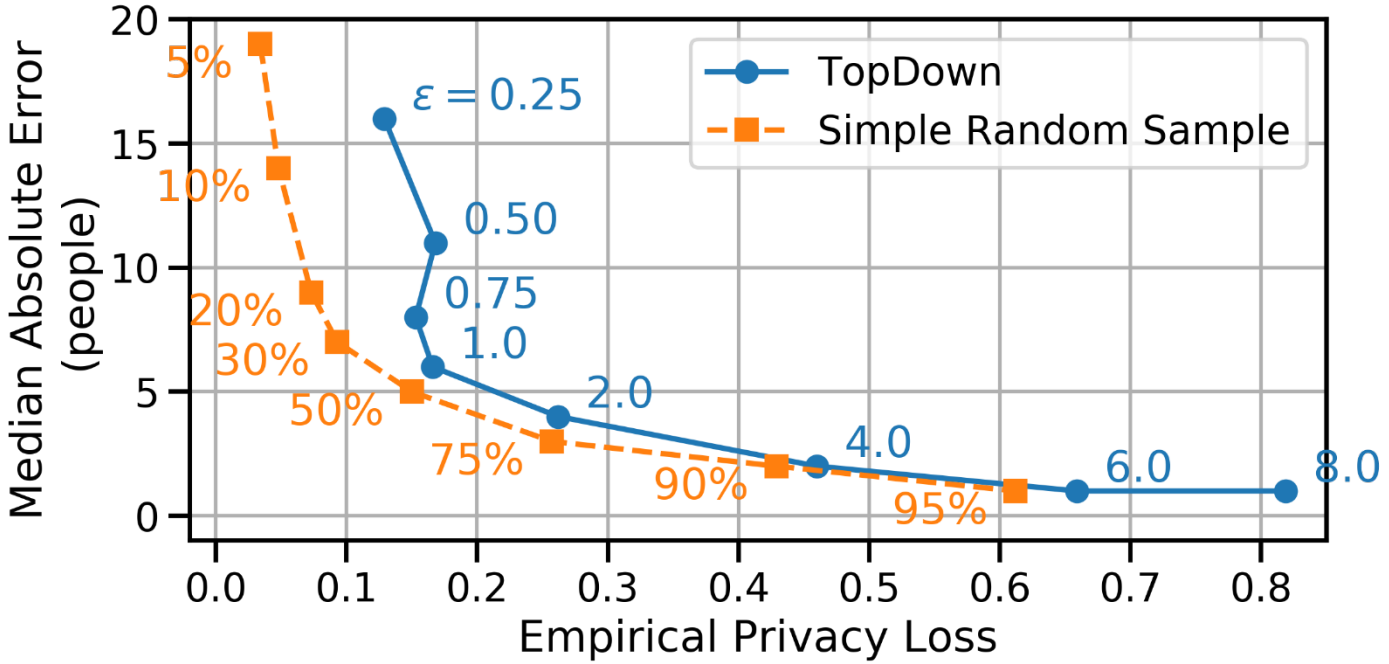
(for epsilon=1.0, for district-level stratified counts)

Empirical Privacy Loss decreases as a function of epsilon, but only for epsilon of at least 1.0



(for county-level stratified counts)

Empirical Privacy Loss decreases as a function of epsilon, but only for epsilon of at least 1.0



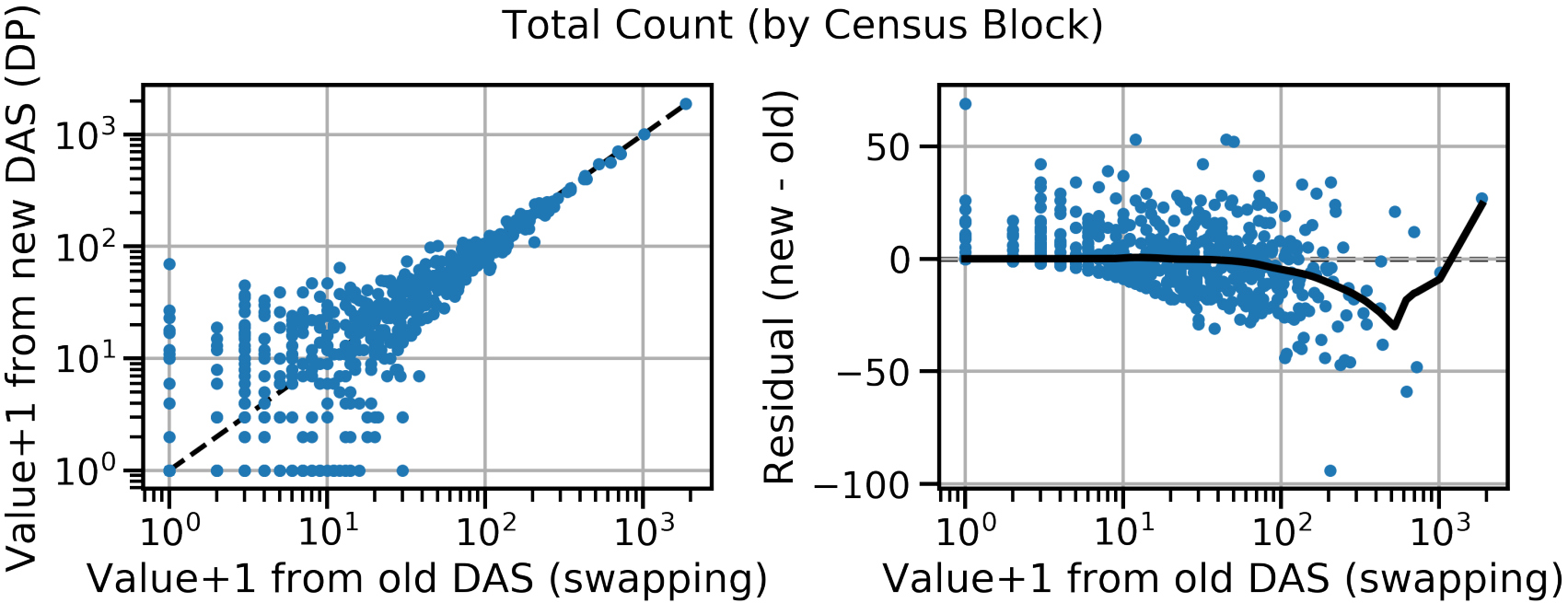
(for county-level stratified counts)

Evidence for decision-making: number of people

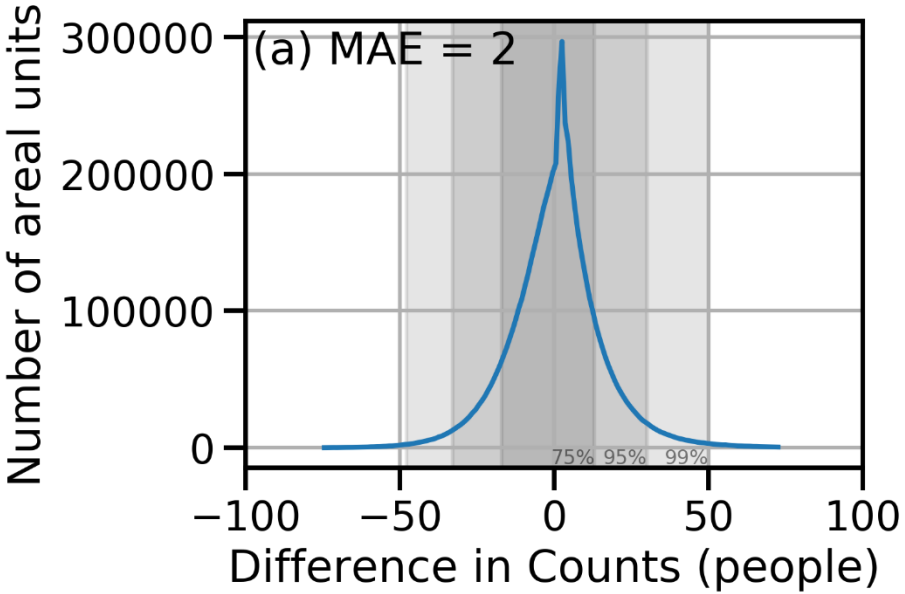
In my own discussions with census data users in state and local government, here are some examples of things they want to know:

1. How many people have been in contact with travelers returning from [country with outbreak]? (epidemic response)
2. How many people should be evacuated in case of forest fire? (emergency preparedness)
3. **How many people are in this city/county and what share of state revenue will that correspond to next year? (budgeting)**

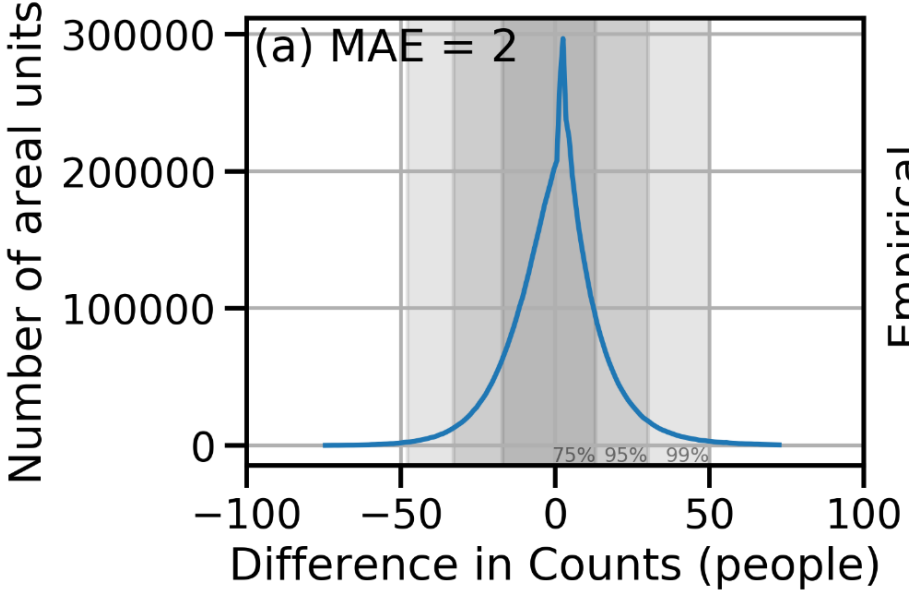
TopDown and Total Counts

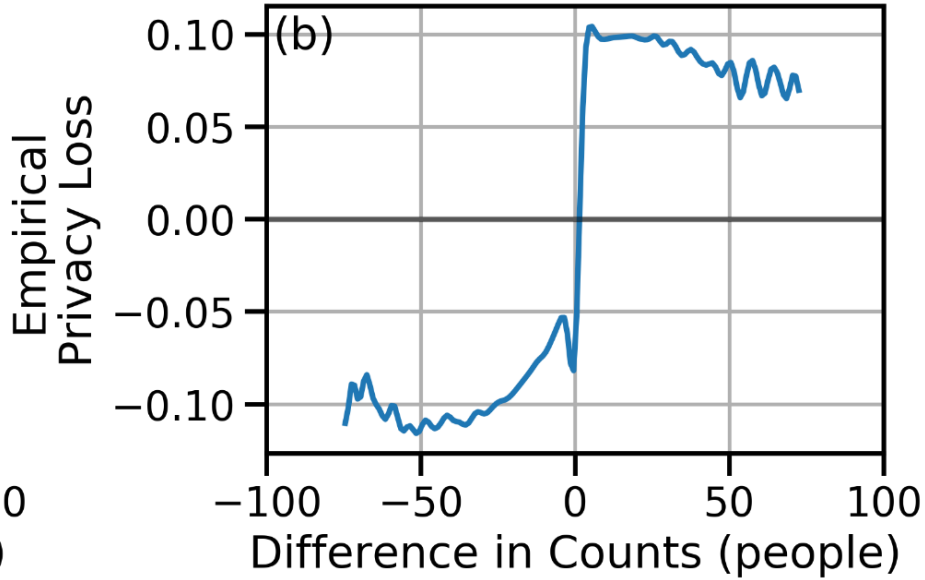


TopDown and Total Counts



TopDown and Total Counts





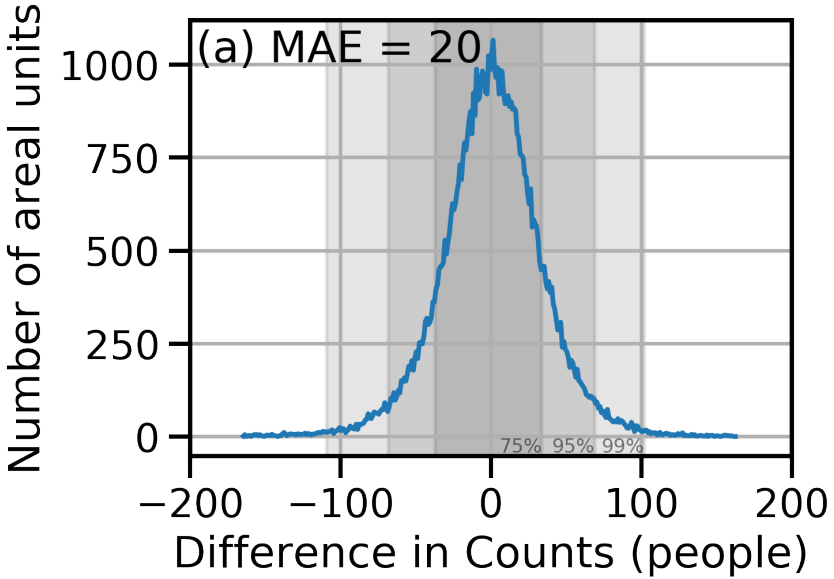
Relevant units of aggregation

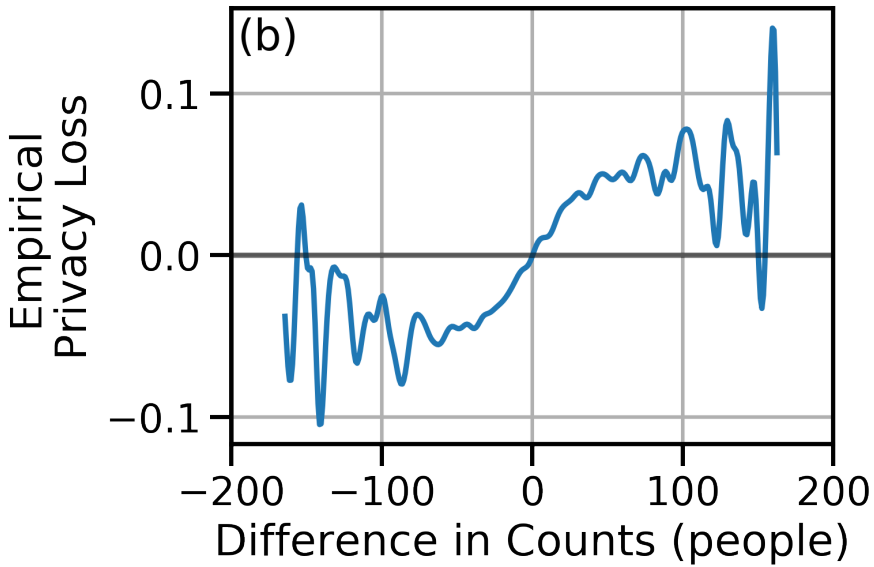
- Census Tract (example of all-cause mortality from King County BoD)

- City - Dollars from state
- County - (in WA) at least

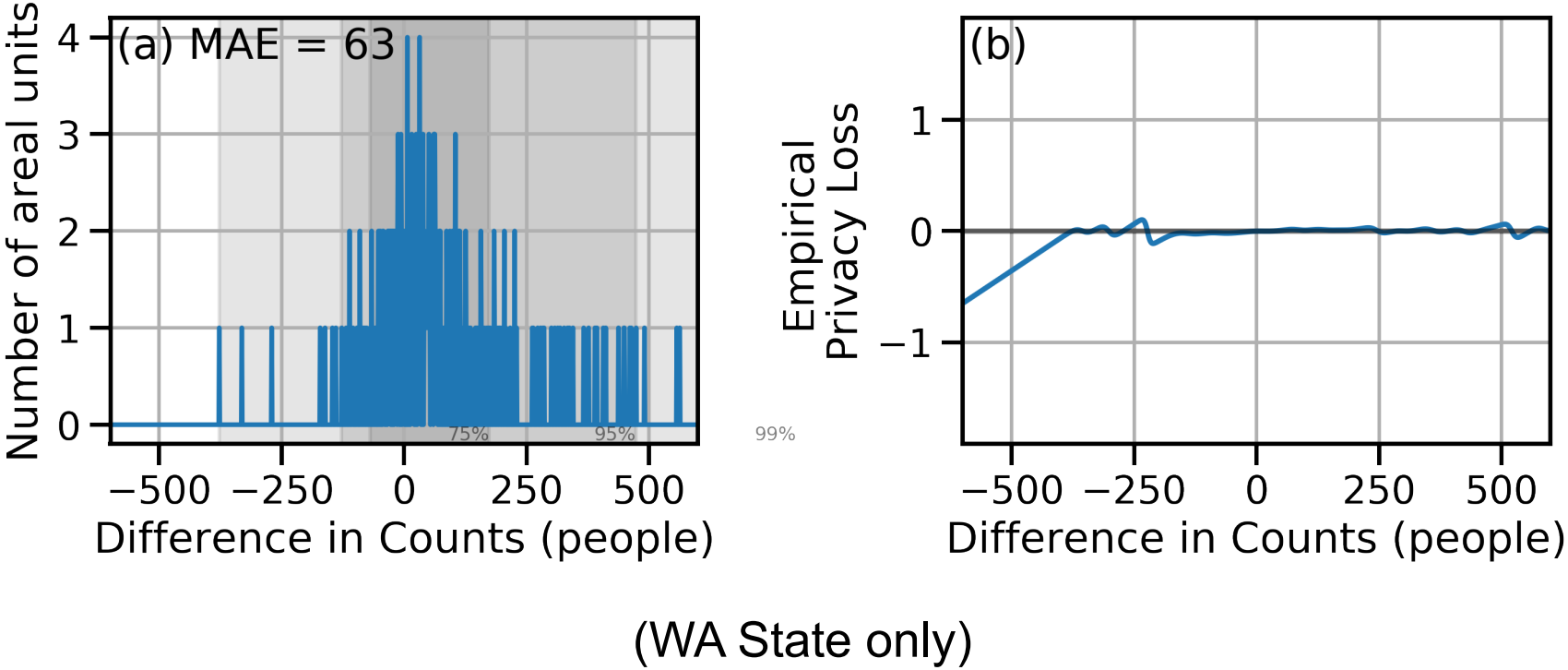
- State – dollars from federal government

Relevant units of aggregation: Census Tract



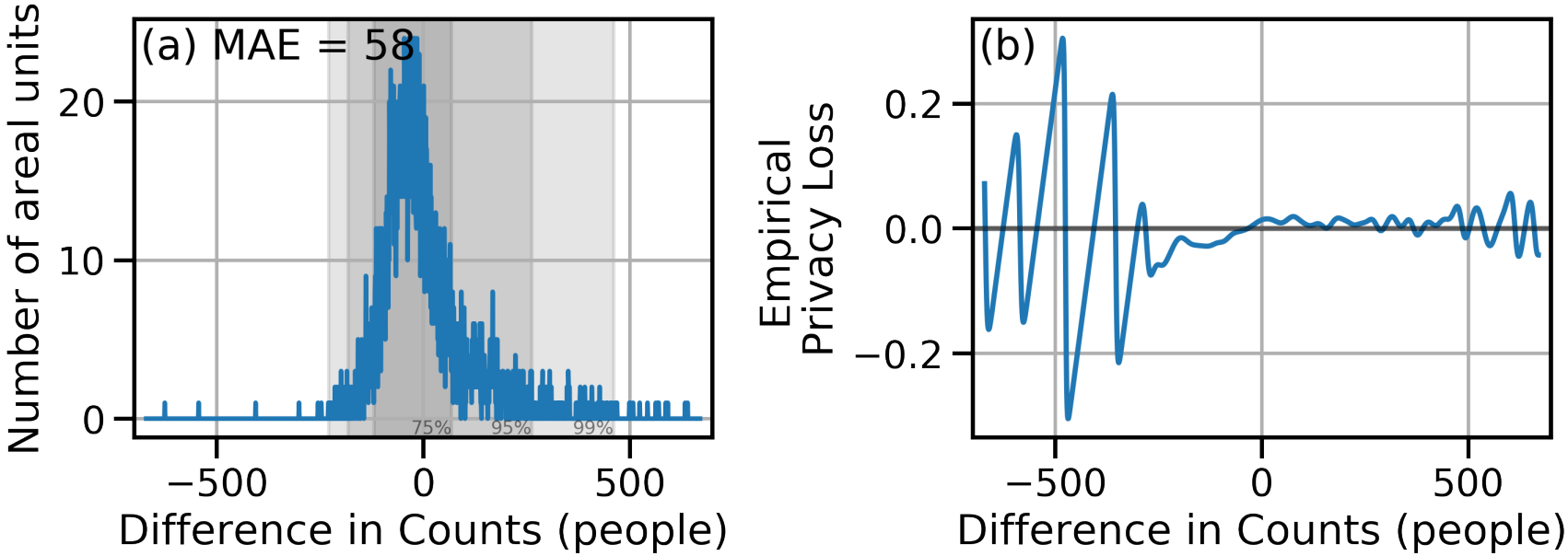


Relevant units of aggregation: City

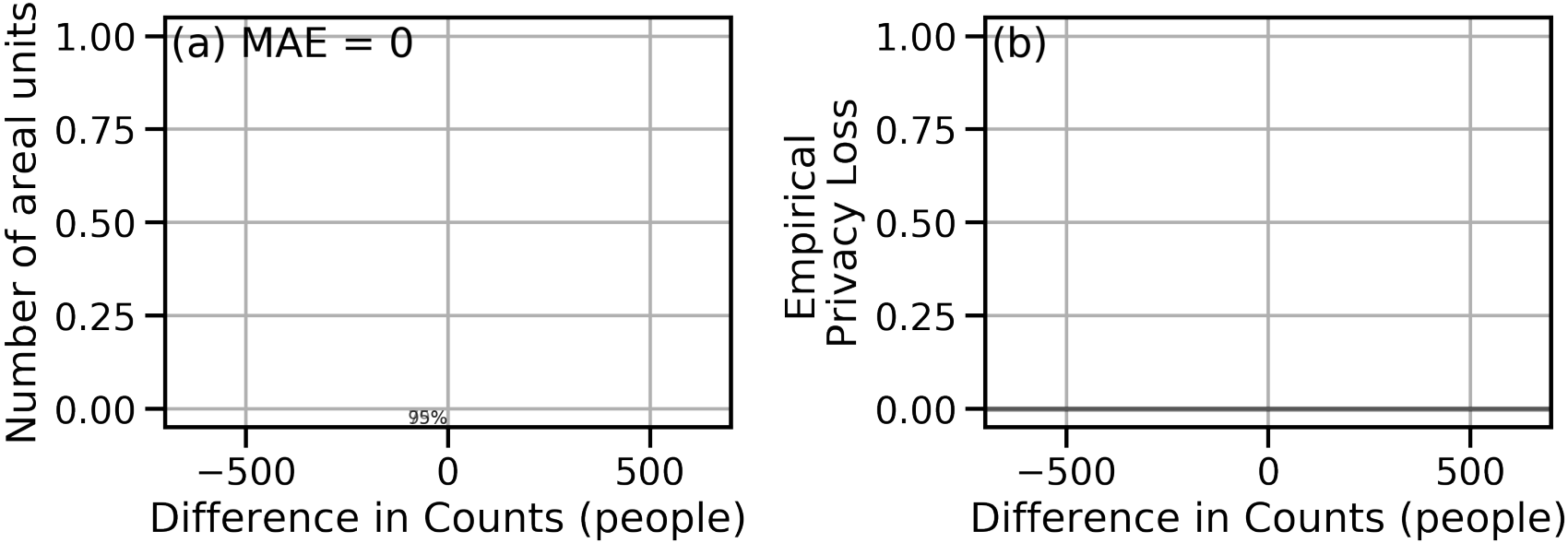


(WA State only)

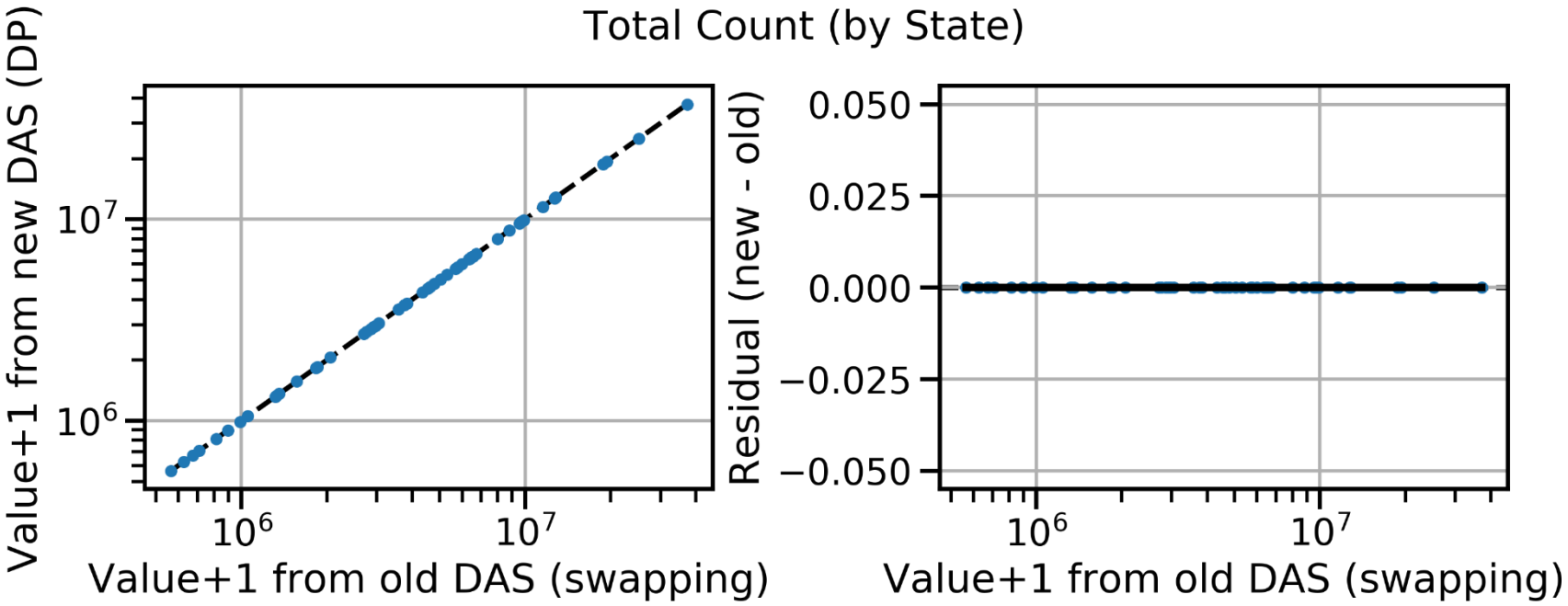
Relevant units of aggregation: County



Relevant units of aggregation: State



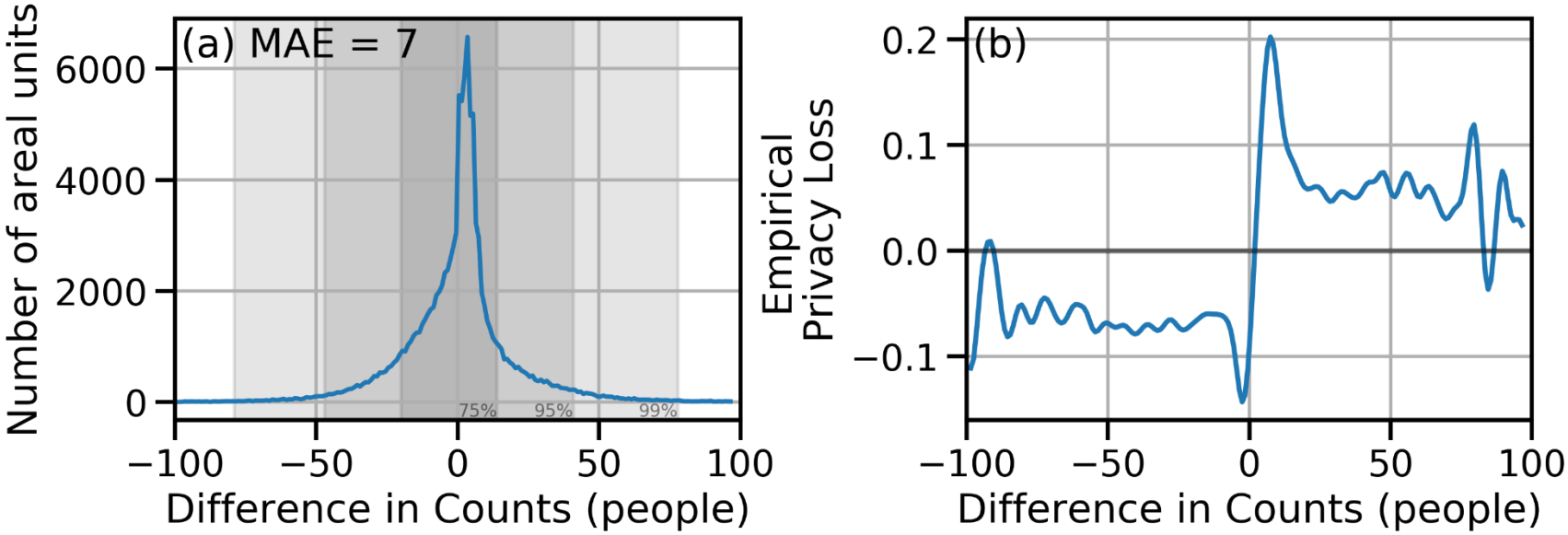
Relevant units of aggregation: State



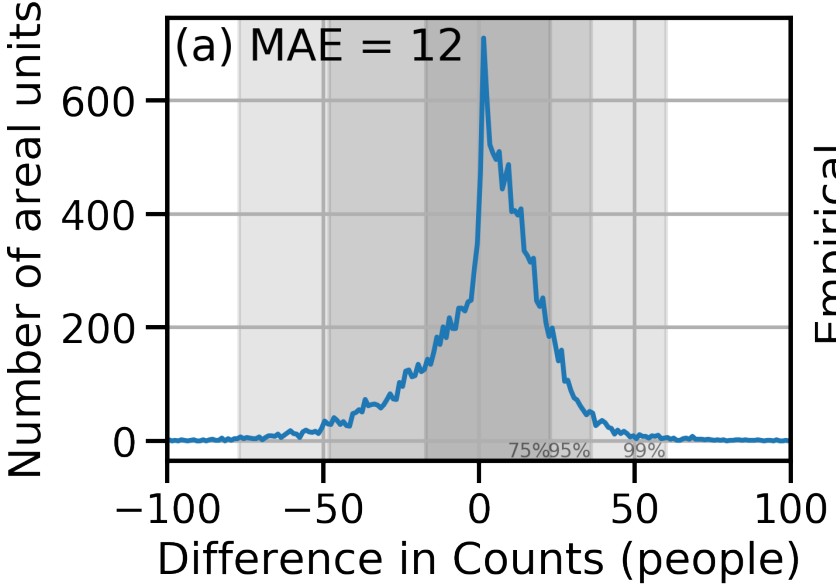
Quality assurance and Group Quarters

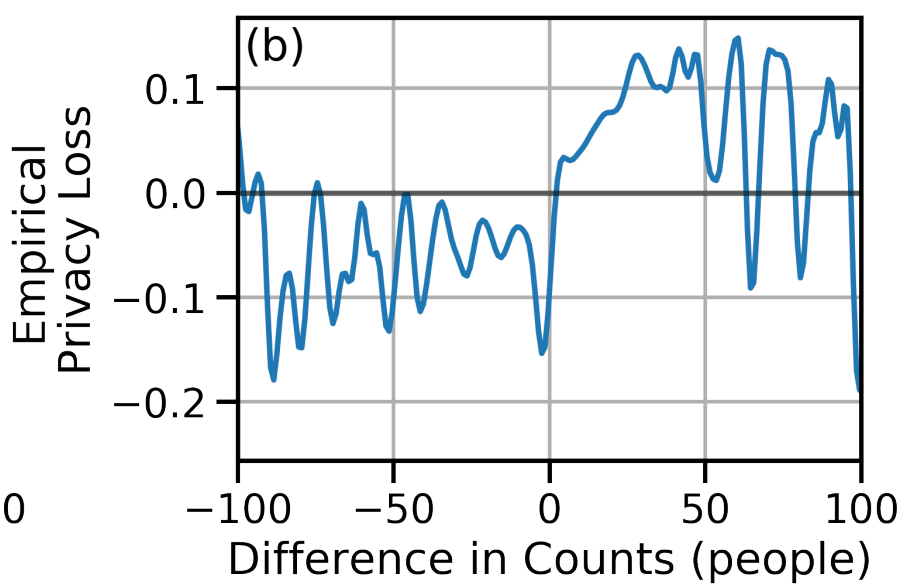
- And correction of census counts, and alternative enumerations in WA state

Quality assurance and Group Quarters (census blocks with non-zero all-GQ counts)

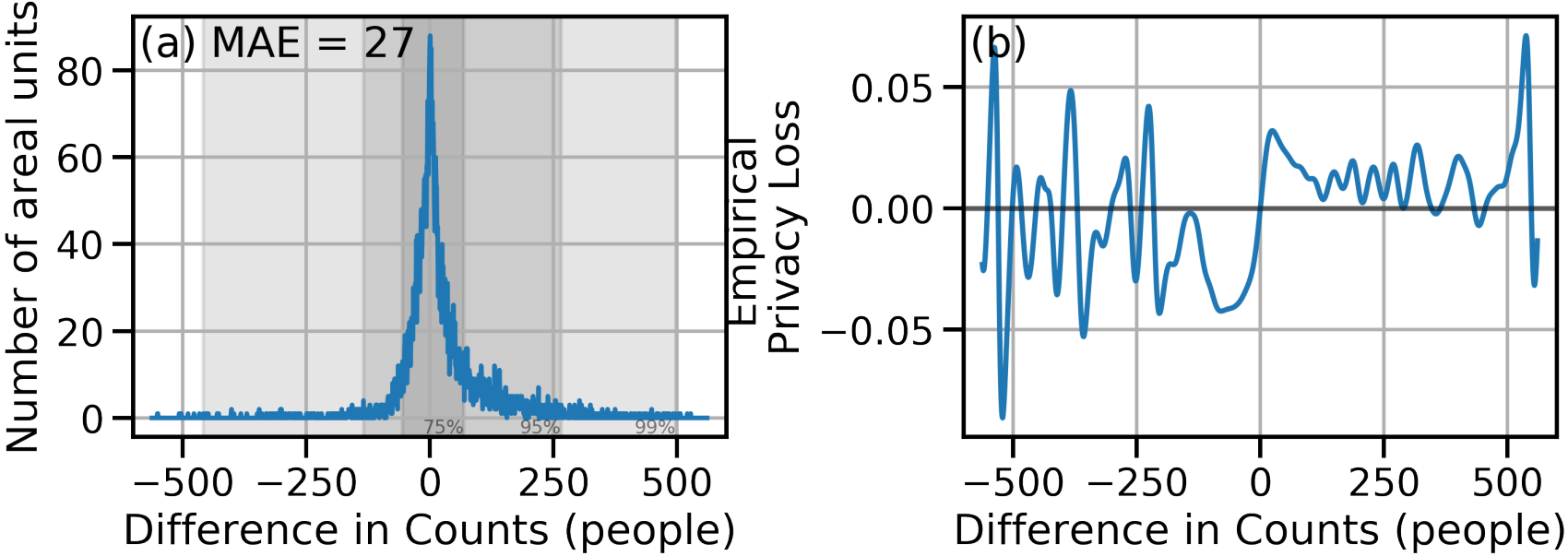


Quality assurance and Group Quarters (census blocks non-zero male-65+-nursing-home counts)





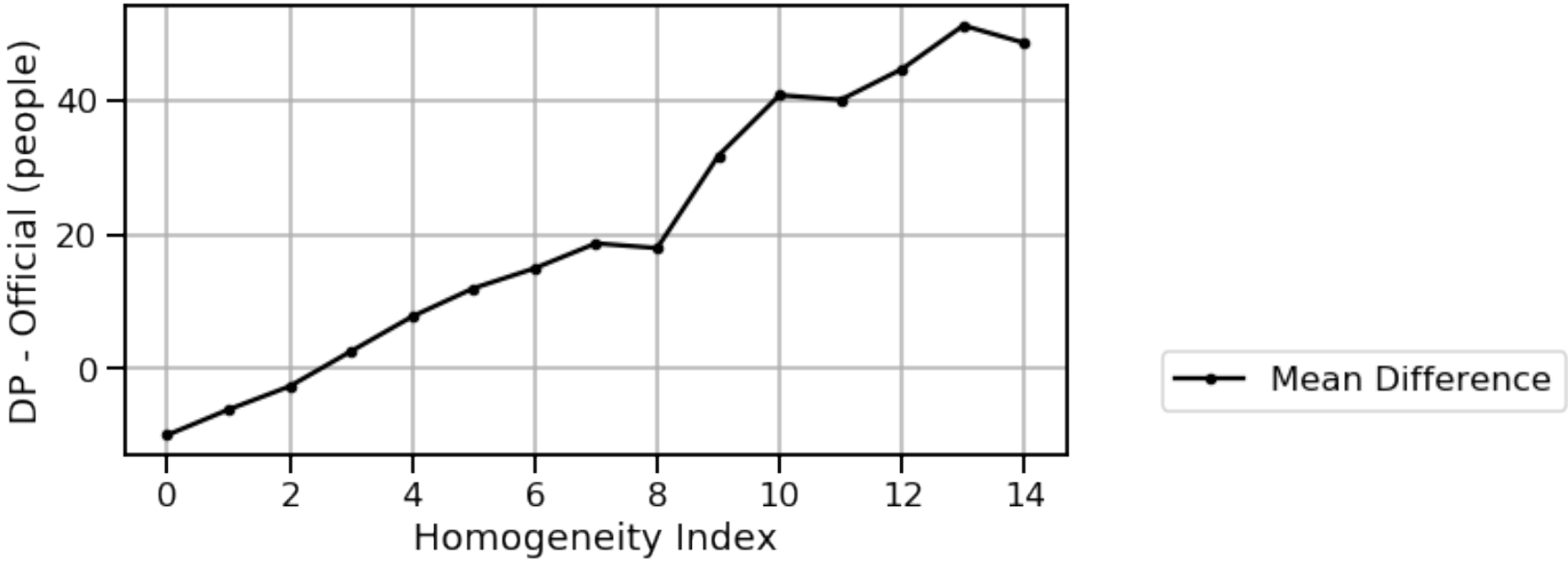
Quality assurance and Group Quarters (census blocks non-zero male-18-to-64-correctional cnts)



“Bias”

Difference between Swapping and DP has predictable structure: the more homogeneous the census tract, the larger the DP count compared to the Swapping count

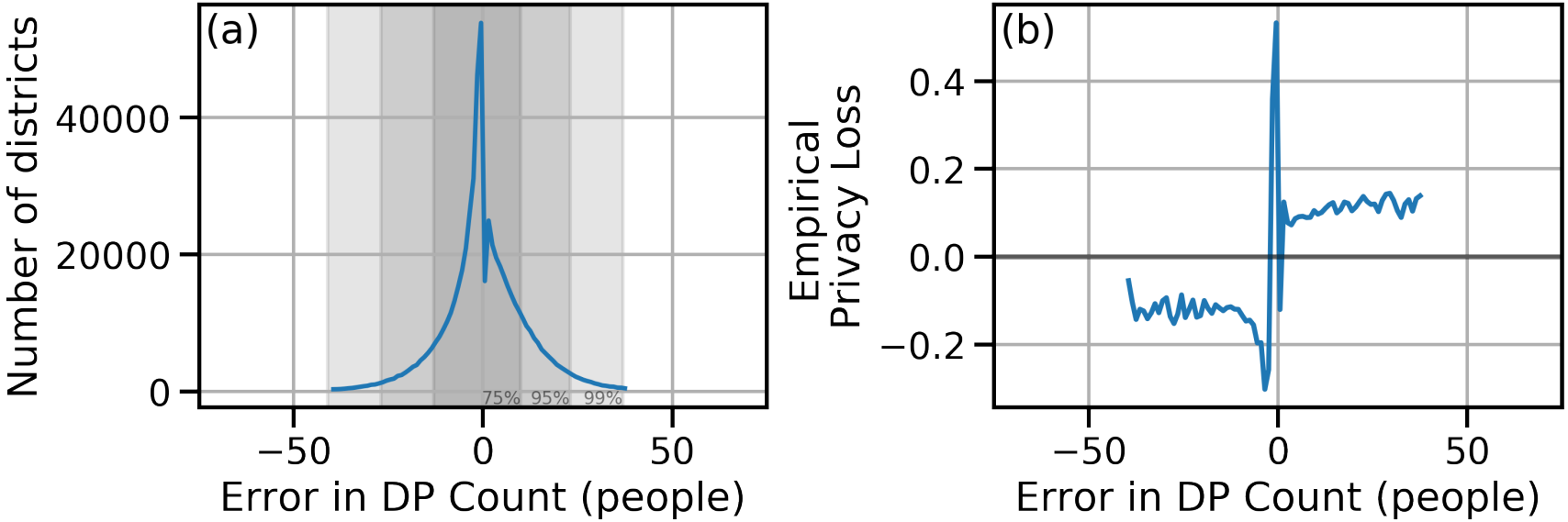
Relationship between homogeneity and average difference for non-empty census tracts



Invariants as a countermeasure for bias

TopDown has a way to make this go away: invariants. Demonstration products have held total count invariant at *state level.* With David Van Riper, I tried making total count invariant at *enumeration district level*. It seems to have worked!

When total count is invariant on enumeration districts, privacy loss is still small (in 1940)



To wrap up back where we started

Bekemeier et al needs assessment identified:

1. **Limited availability or access to data**
2. Data quality issues
3. **Limited staff with expertise and resources for analyzing data**

Most relevant for us is (2), but we also have an opportunity to address (1) and (3) through release of I.P.D. data, for each county.

We should also release the imprecise counts (pre-optimization) in a “replication archive” (not for typical use by rural LHD, perhaps, but useful.)

My Recommendations

1. Include total count at census block level as an invariant or address bias in some other way

2. Publish (a) “county-by-county synthetic microdata” files and (b) “replication archive histogram-with-uncertainty” files

Thank you

Thanks to:

- Jan Vink (Cornell), David Van Riper (IPUMS), Mike Mohrman (WA OFM) for being data heros;
- Samantha Petti, ACO/Math PhDc at GATech, who did all the work understanding the code;
- Simson Garfinkel and Philip Leclerc at Census Bureau who graciously answered many questions from me and Sam;
- Social researchers like many of you, who have been patient despite being anxious about these big changes happening for the 2020 US Census.