

Privacy, Census Data, and Arizona Redistricting

**an overview
with experiments**

**Moon Duchin
MGGG Redistricting Lab
Tisch College of Civic Life, Tufts University**



About MGGG Redistricting Lab

- Non-partisan scholarly research
- Community mapping support
- Map evaluation



Main funder: **National Science Foundation**
("Network Science of Census Data")

Differential privacy study funded by **Alfred P. Sloan Foundation** – joint work
with Aloni Cohen, **JN Matthews**, and Bhushan Suwal, in collaboration with
Mark Hansen, Denis Kazakov, and Peter Wayner

Pima County,
pop. 980,263

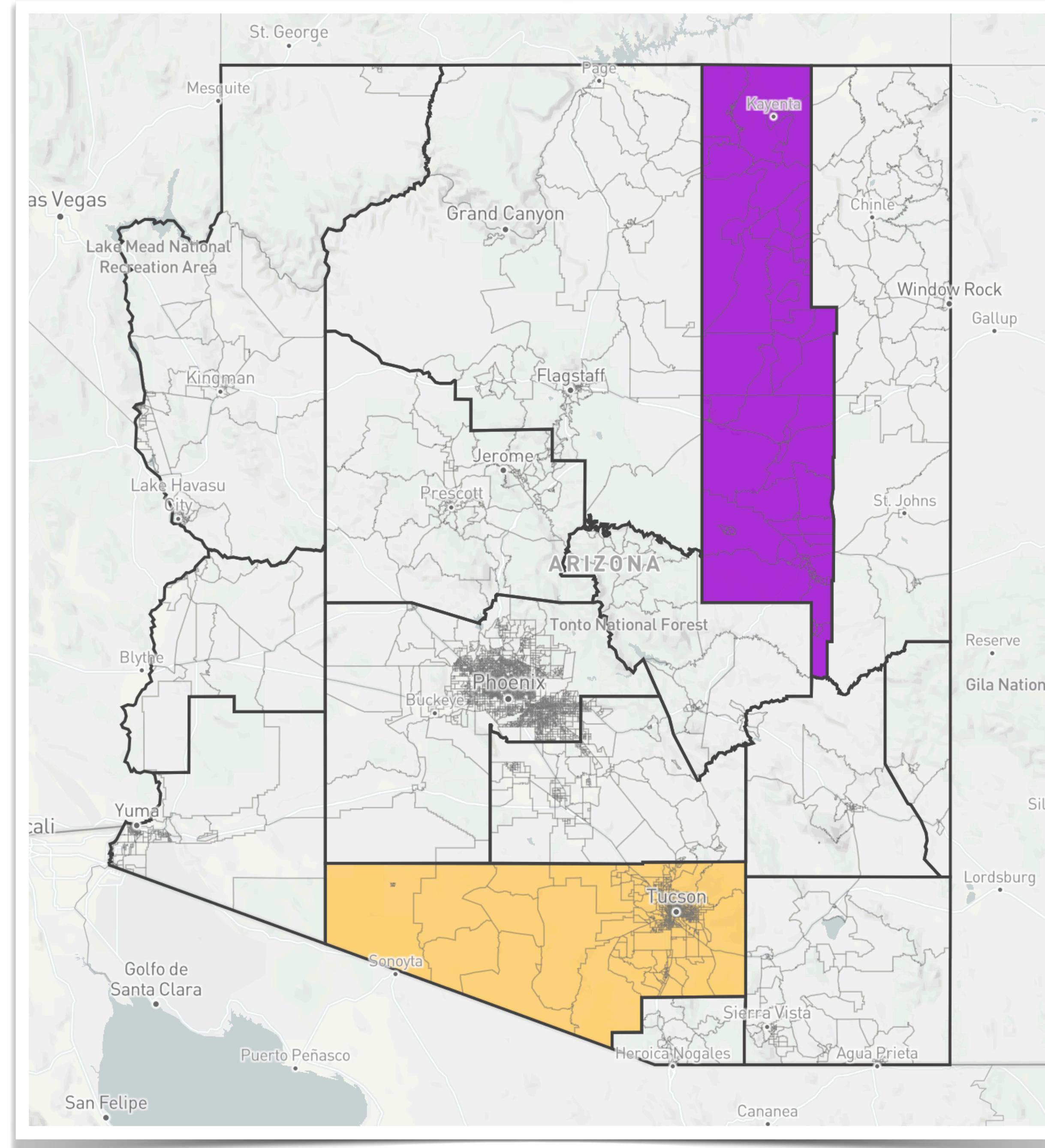
55%W, 35%H,
2.5%AMIN

Large districts
(U.S. Congress)

$7,151,502/9 \approx 794,611$

Small districts
(Navajo County commission)

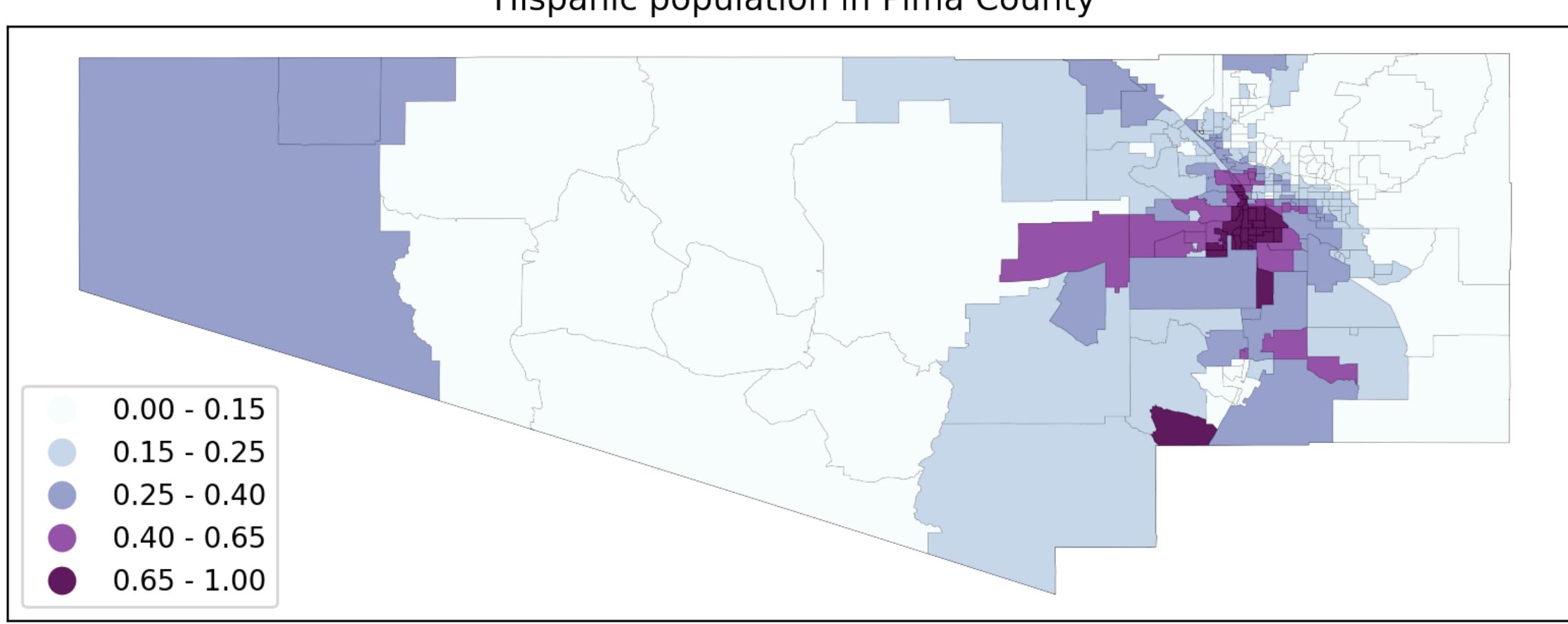
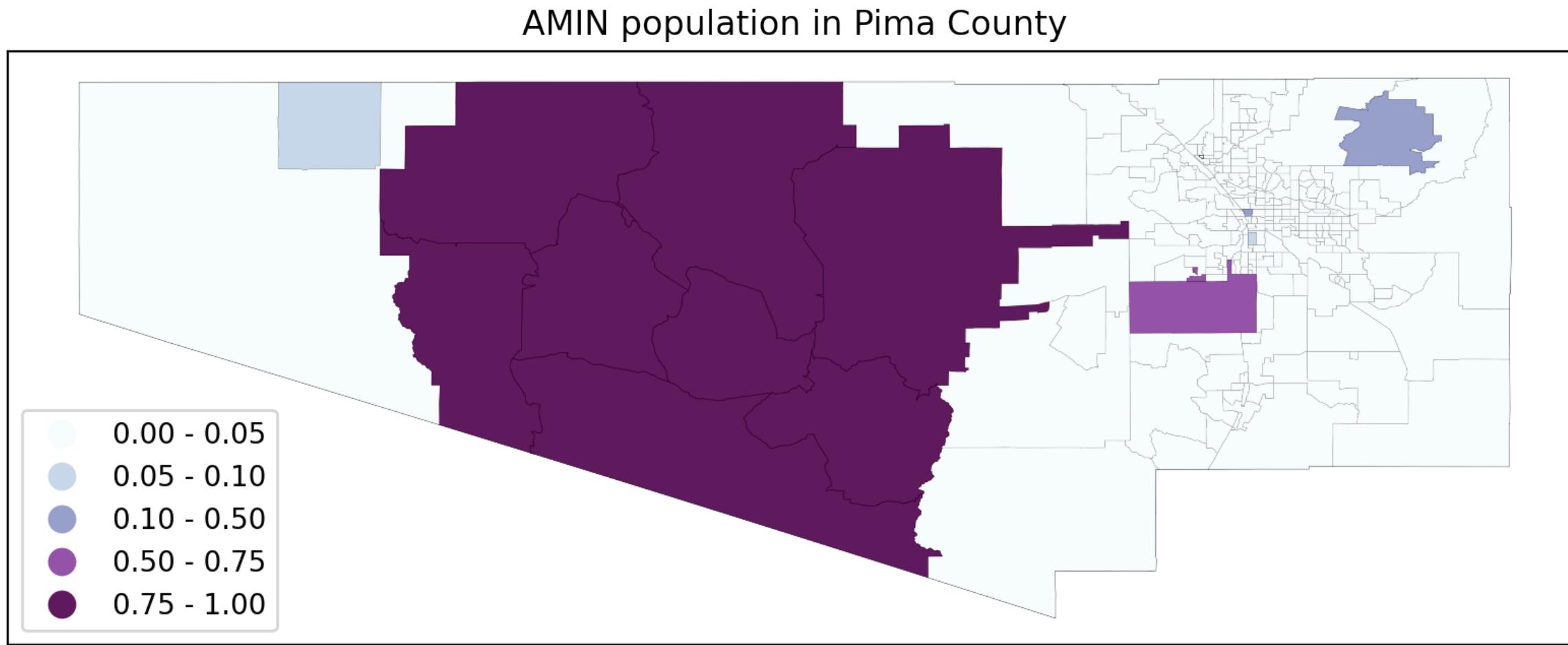
$107,449/5 \approx 21,490$



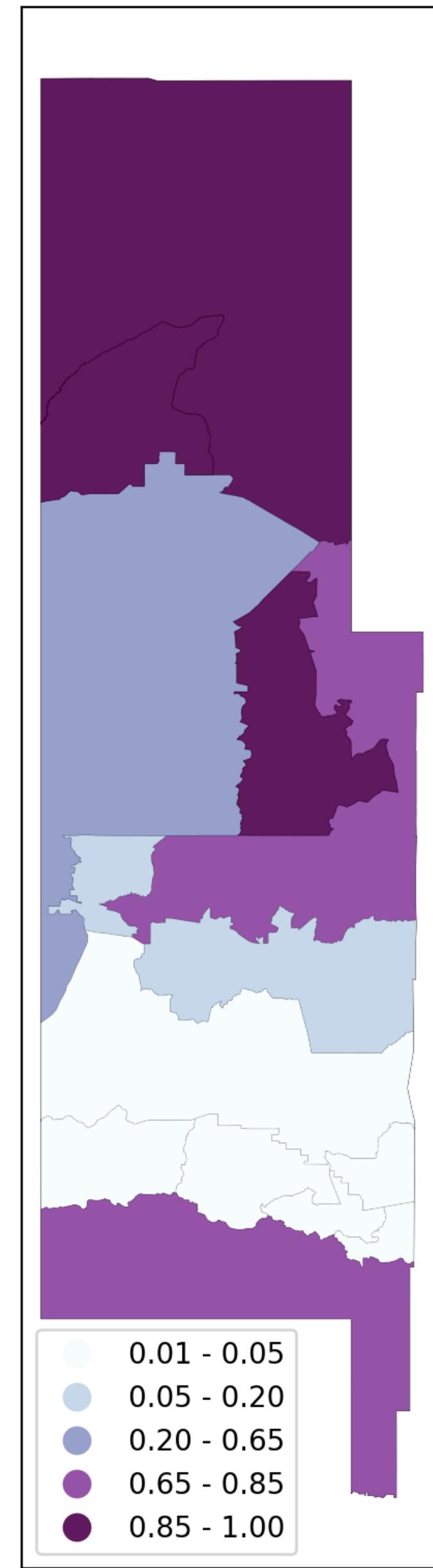
Navajo County,
pop. 107,449

44%W, 11%H,
42%AMIN

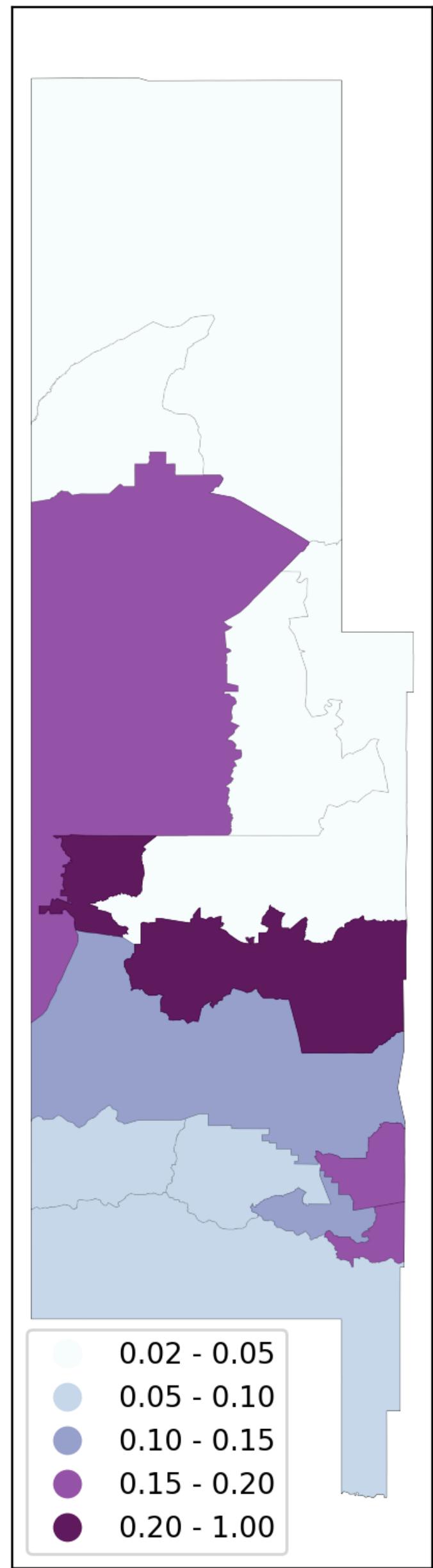
Both counties have significant diversity



AMIN population in Navajo County



Hispanic population in Navajo County



What is the risk?



Reconstructing Navajo County

in <6 hours on a student-grade laptop, we recovered a complete person-by-person list of location, ethnicity, sex, age, race for every enumerated resident of Navajo County in 2010

can get whole state in a few days

our table is **100% consistent** with the aggregate numbers released by the Census

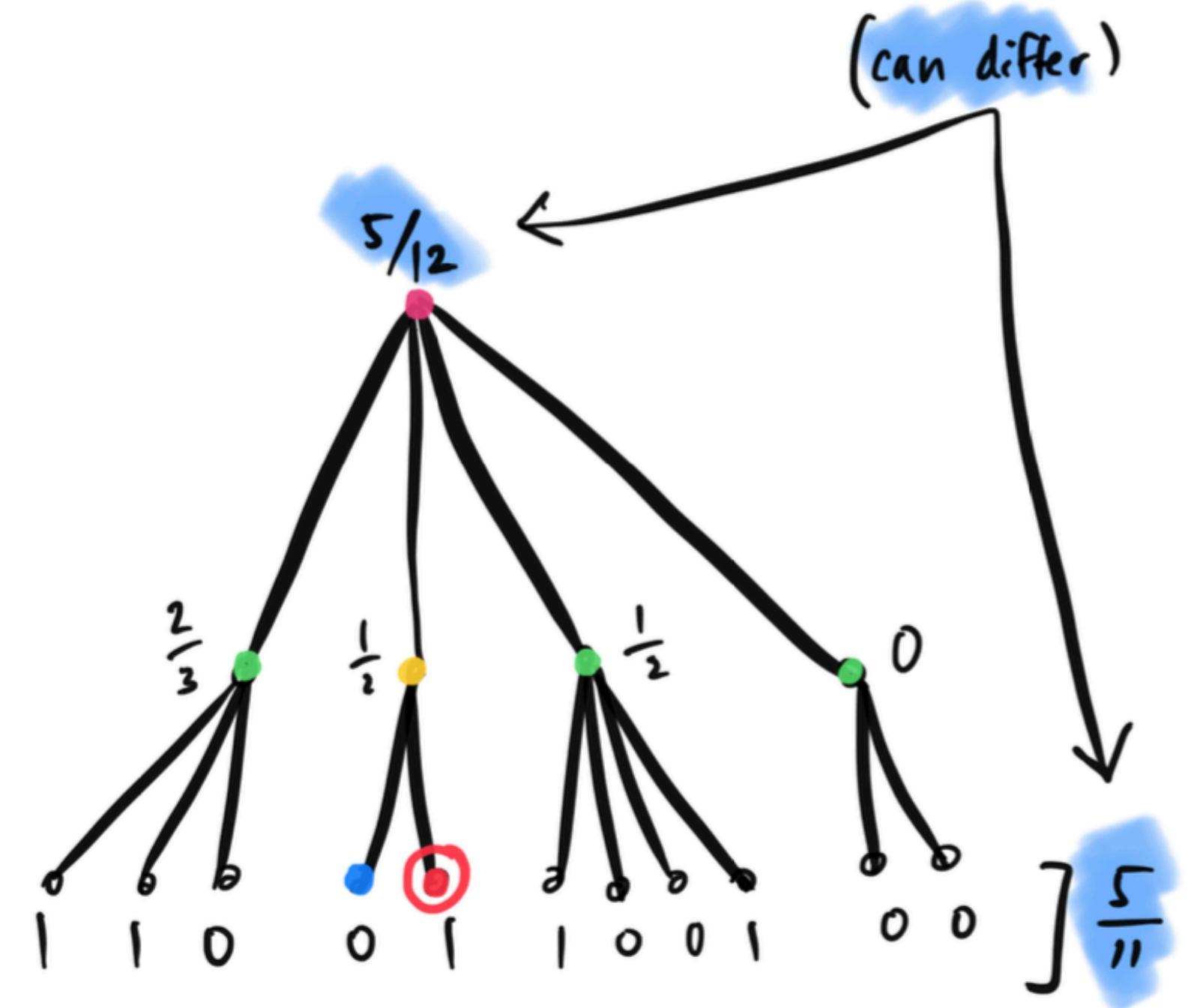
(the only inaccuracies come from the existence of multiple solutions)

pairs with easily obtained commercial data to get full **reidentification**

census_api_test.ipynb M	CensusModel.fs	reconstructr.fsproj
reconstructr# > results > 04017965300_output.csv		
1 GEOID, ETHN, SEX, AGE, RACE, SOL		
2 040179653001055, NH, M, Yrs 57, WHITE, 2.000000		
3 040179653001055, NH, M, Yrs 60, WHITE, 1.000000		
4 040179653001055, NH, F, Yrs 52, WHITE, 2.000000		
5 040179653001124, H, M, Yrs 5, OTHER, 1.000000		
6 040179653001124, H, M, Yrs 33, OTHER, 1.000000		
7 040179653001124, H, F, Yrs 10, OTHER, 1.000000		
8 040179653001124, H, F, Yrs 34, WHITE, 1.000000		
9 040179653001124, NH, M, Yrs 3, WHITE, 1.000000		
10 040179653001124, NH, M, Yrs 21, WHITE, 1.000000		
11 040179653001124, NH, M, Yrs 27, WHITE, 2.000000		
12 040179653001124, NH, M, Yrs 32, WHITE, 1.000000		
13 040179653001124, NH, M, Yrs 37, WHITE, 2.000000		
14 040179653001124, NH, M, Yrs 42, WHITE, 1.000000		
15 040179653001124, NH, M, Yrs 47, WHITE, 1.000000		
16 040179653001124, NH, M, Yrs 52, WHITE, 3.000000		
17 040179653001124, NH, M, Yrs 55, WHITE, 3.000000		
18 040179653001124, NH, M, Yrs 61, AMIN, 1.000000		
19 040179653001124, NH, M, Yrs 61, WHITE, 2.000000		
20 040179653001124, NH, M, Yrs 72, WHITE, 1.000000		
21 040179653001124, NH, M, Yrs 90, WHITE, 1.000000		
22 040179653001124, NH, F, Yrs 0, WHITE, 1.000000		
23 040179653001124, NH, F, Yrs 8, WHITE, 1.000000		
24 040179653001124, NH, F, Yrs 11, WHITE, 1.000000		
25 040179653001124, NH, F, Yrs 15, WHITE, 1.000000		
26 040179653001124, NH, F, Yrs 27, WHITE, 3.000000		
27 040179653001124, NH, F, Yrs 42, WHITE, 1.000000		
28 040179653001124, NH, F, Yrs 47, WHITE, 1.000000		
29 040179653001124, NH, F, Yrs 52, WHITE, 3.000000		
30 040179653001124, NH, F, Yrs 59, WHITE, 2.000000		
31 040179653001124, NH, F, Yrs 61, WHITE, 1.000000		
32 040179653001124, NH, F, Yrs 64, WHITE, 1.000000		
33 040179653001124, NH, F, Yrs 69, WHITE, 1.000000		
34 040179653001124, NH, F, Yrs 75, WHITE, 1.000000		
35 040179653001124, NH, F, Yrs 86, WHITE, 1.000000		
36 040179653001125, H, M, Yrs 13, WHITE, 1.000000		
37 040179653001125, NH, M, Yrs 3, WHITE, 1.000000		
38 040179653001125, NH, M, Yrs 6, WHITE, 1.000000		
39 040179653001125, NH, M, Yrs 10, WHITE, 1.000000		
40 040179653001125, NH, M, Yrs 19, WHITE, 1.000000		
41 040179653001125, NH, M, Yrs 24, WHITE, 1.000000		
42 040179653001125, NH, M, Yrs 34, WHITE, 2.000000		
43 040179653001125, NH, M, Yrs 35, WHITE, 1.000000		

test: conda) 0 △ 0 csv | 04017965300_output.csv CSVLint Query

What is differential privacy?

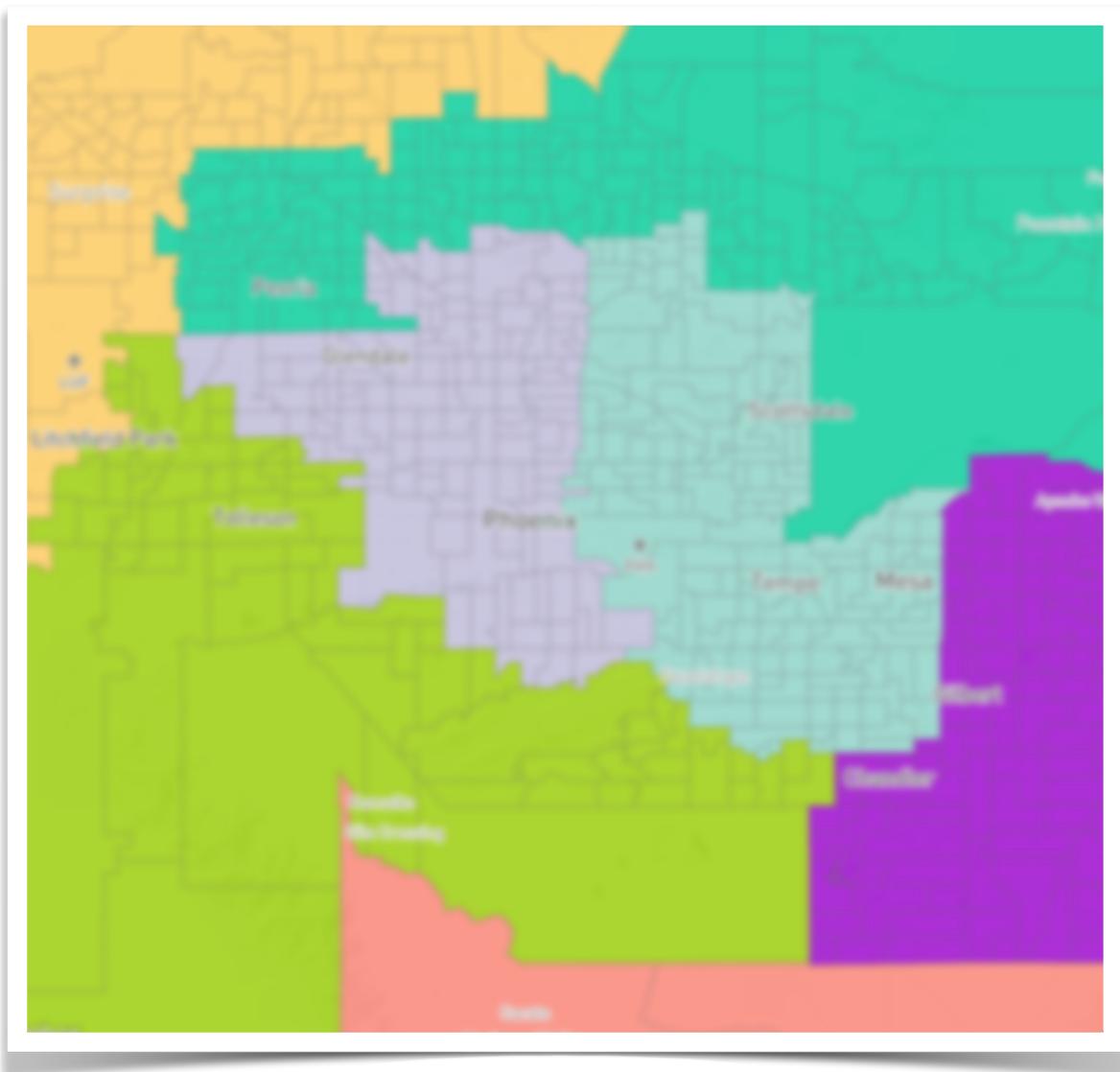
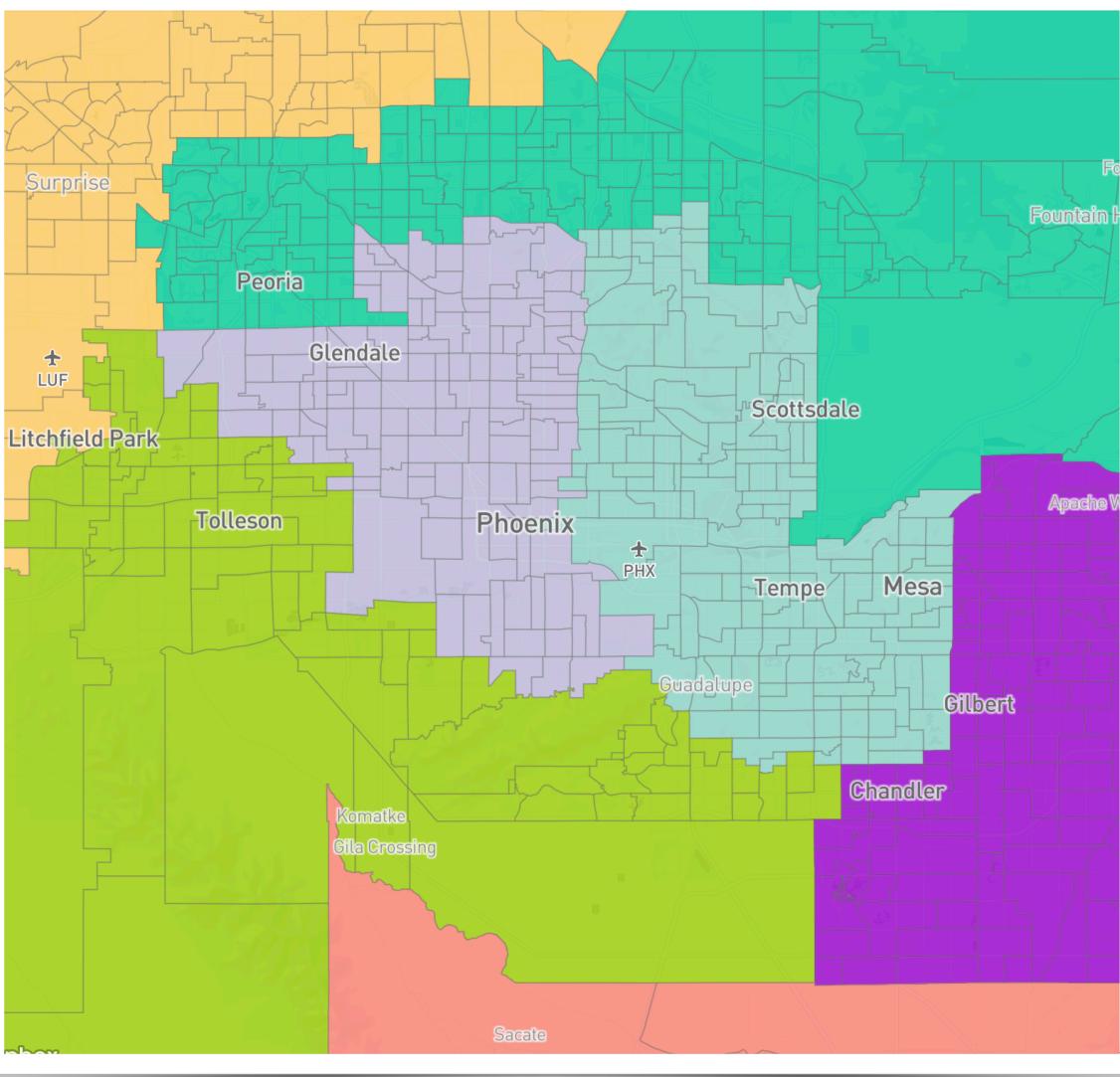


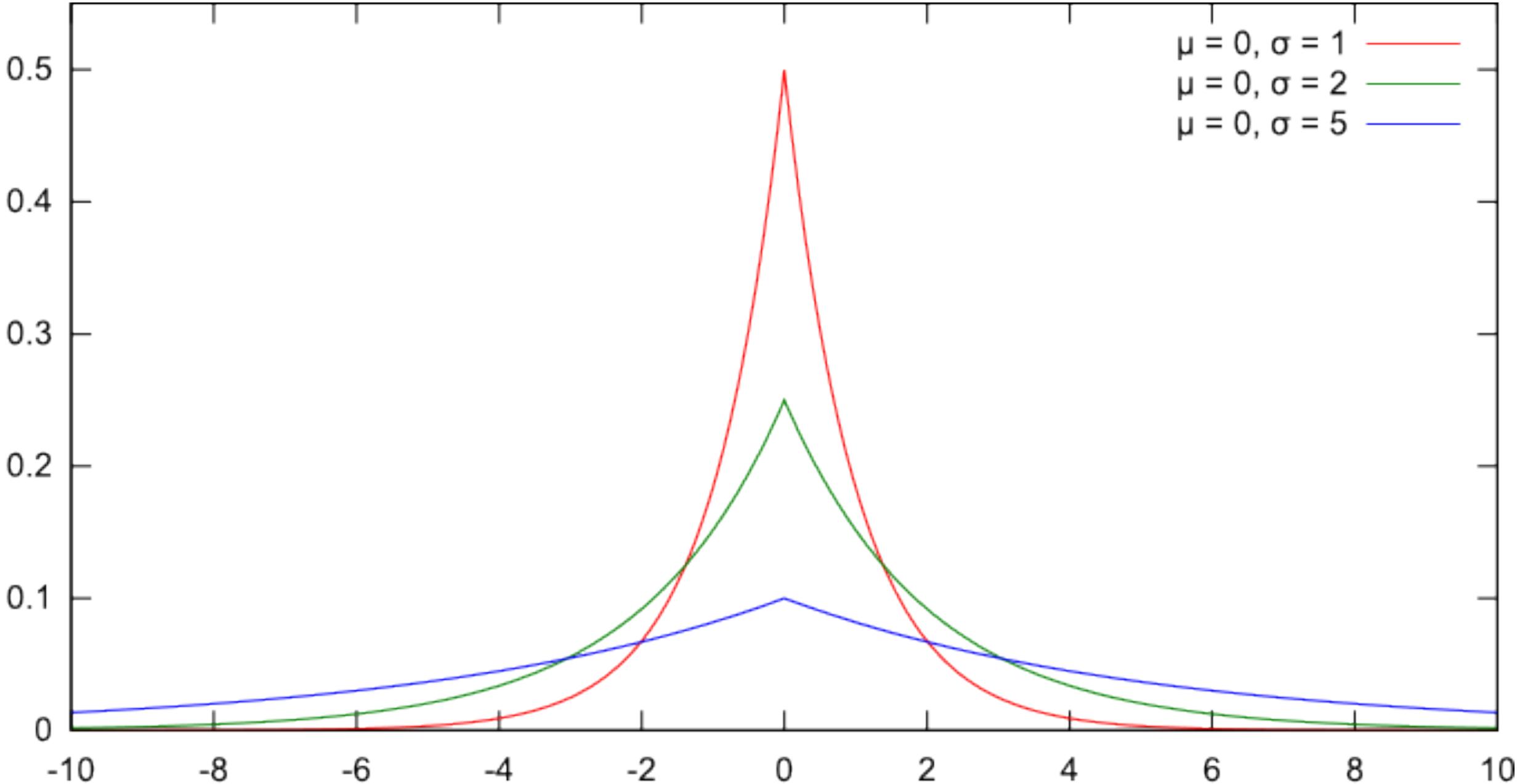
$$\begin{aligned}\text{Error} = & \frac{1}{2}L_3 + -\frac{1}{2}L_3 \\ & + \frac{1}{12}L_2 + \frac{1}{4}L_2 + \frac{1}{12}L_2 + -\frac{5}{12}L_2 \\ & + \frac{5}{12}L_1\end{aligned}$$

Punishes inhomogeneity in each sibling group!

Idea: for **privacy**, add **noise**

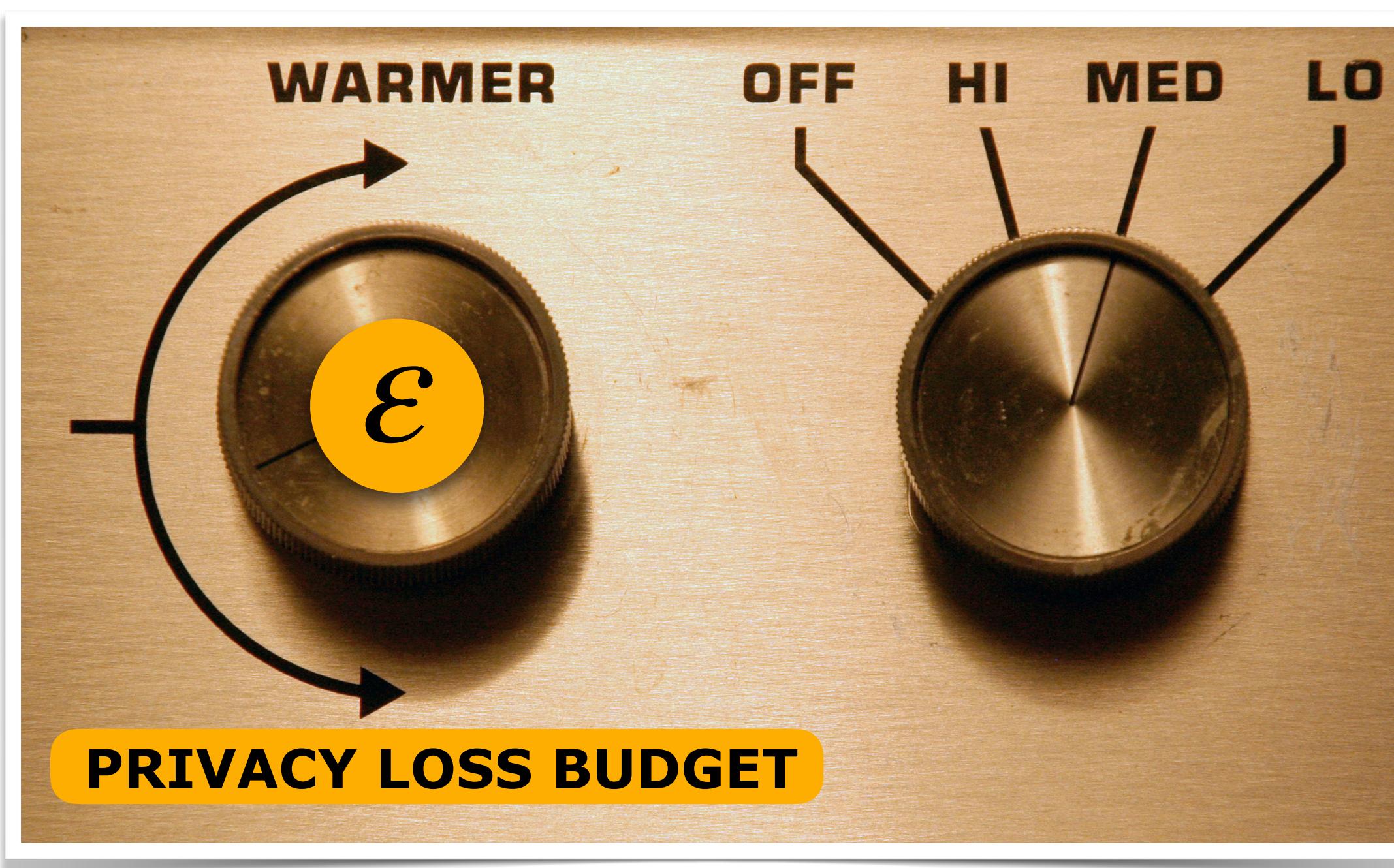
make the numbers fuzzier so
exact reconstruction is impossible





we'll draw **random numbers** to add to every count in the Census redistricting release (PL 94-171)

“differential privacy” essentially means that you have control over the knobs – can **calibrate** the tradeoff between privacy and accuracy



Census “**TopDown**” algorithm

two main things to know:

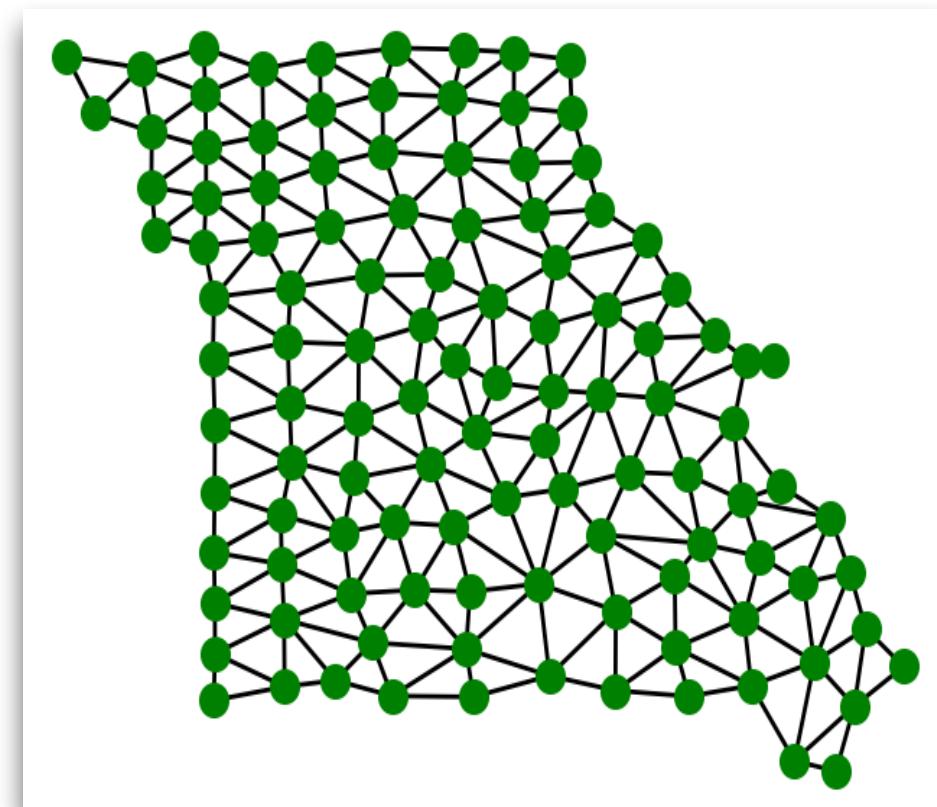
- (1) it uses the geographical **hierarchy**, from top to bottom
- (2) after adding random noise, there's a **processing** phase to make the numbers satisfy various plausibility constraints

top

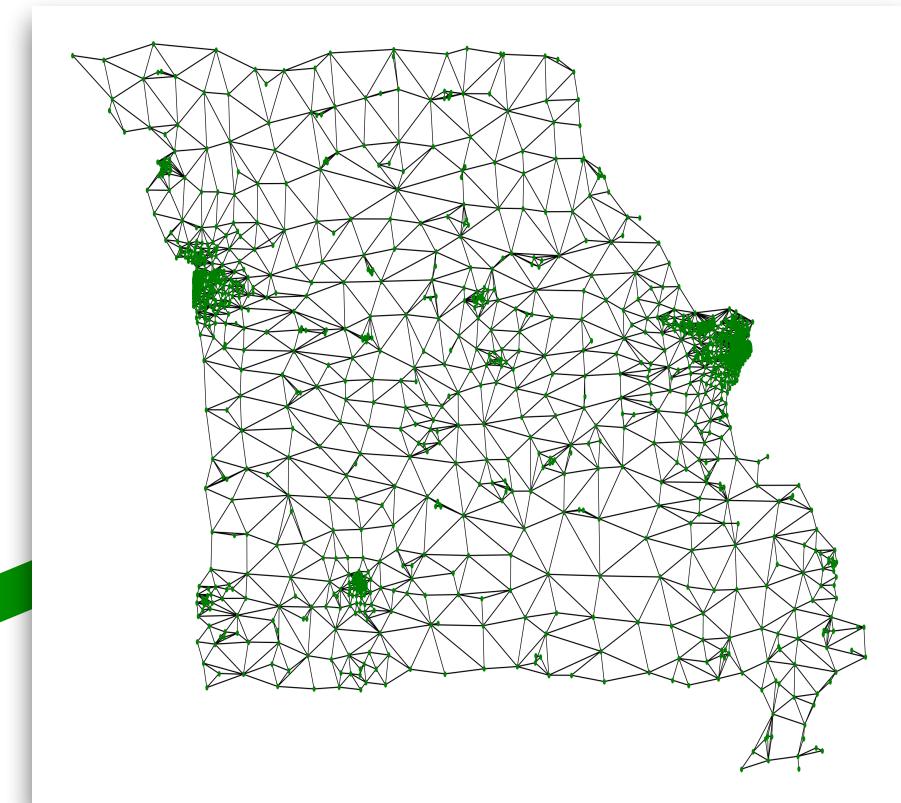
down



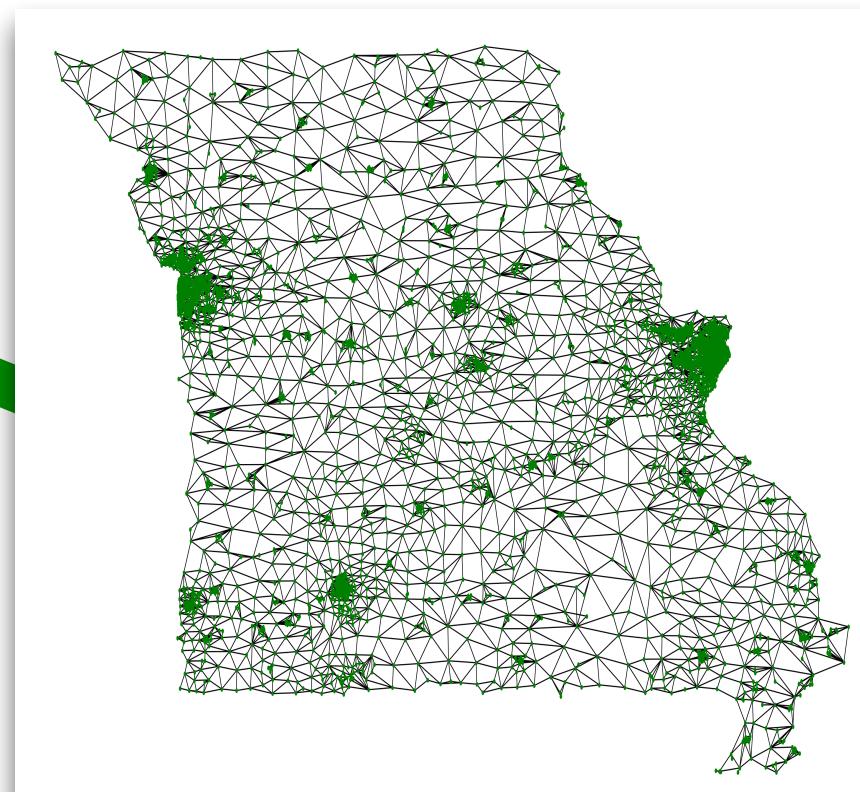
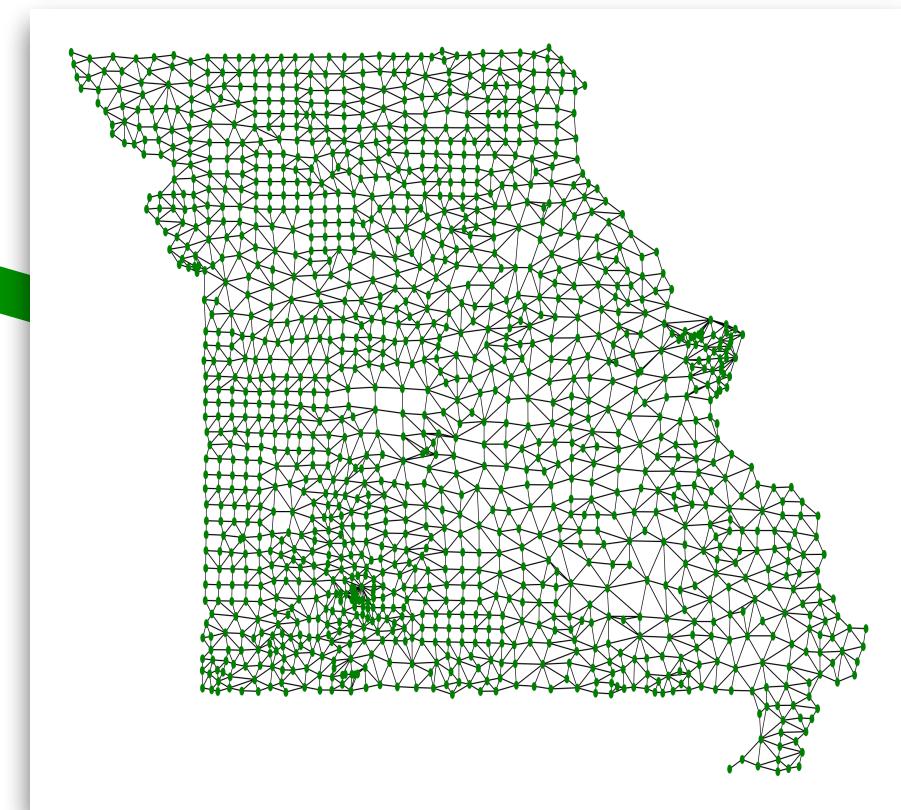
tracts



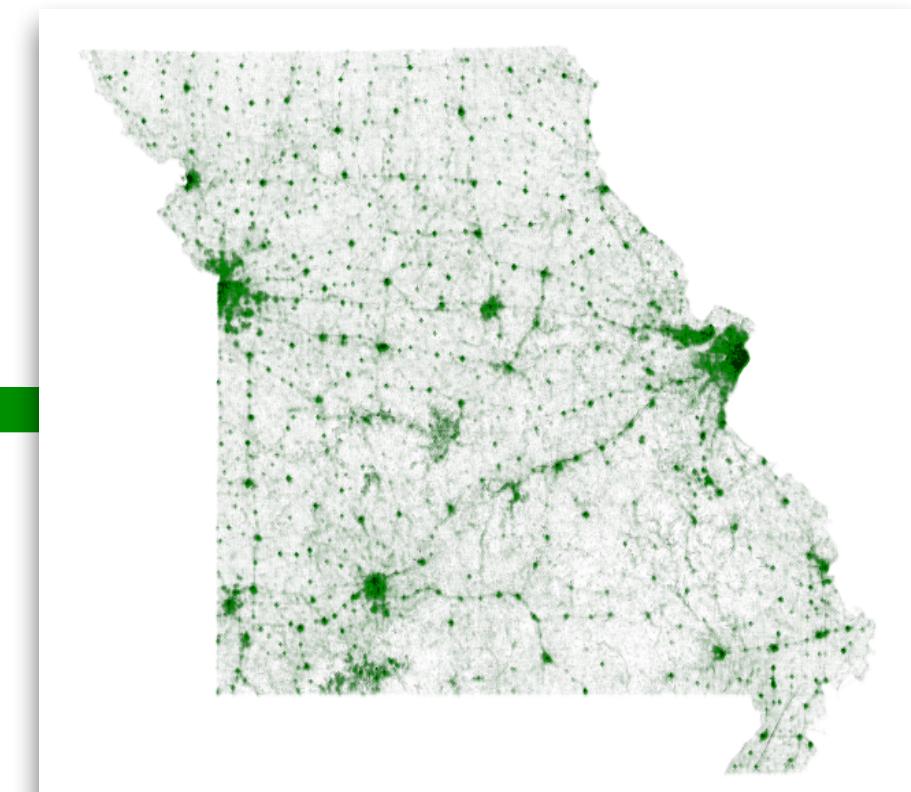
counties



county subunits



block groups



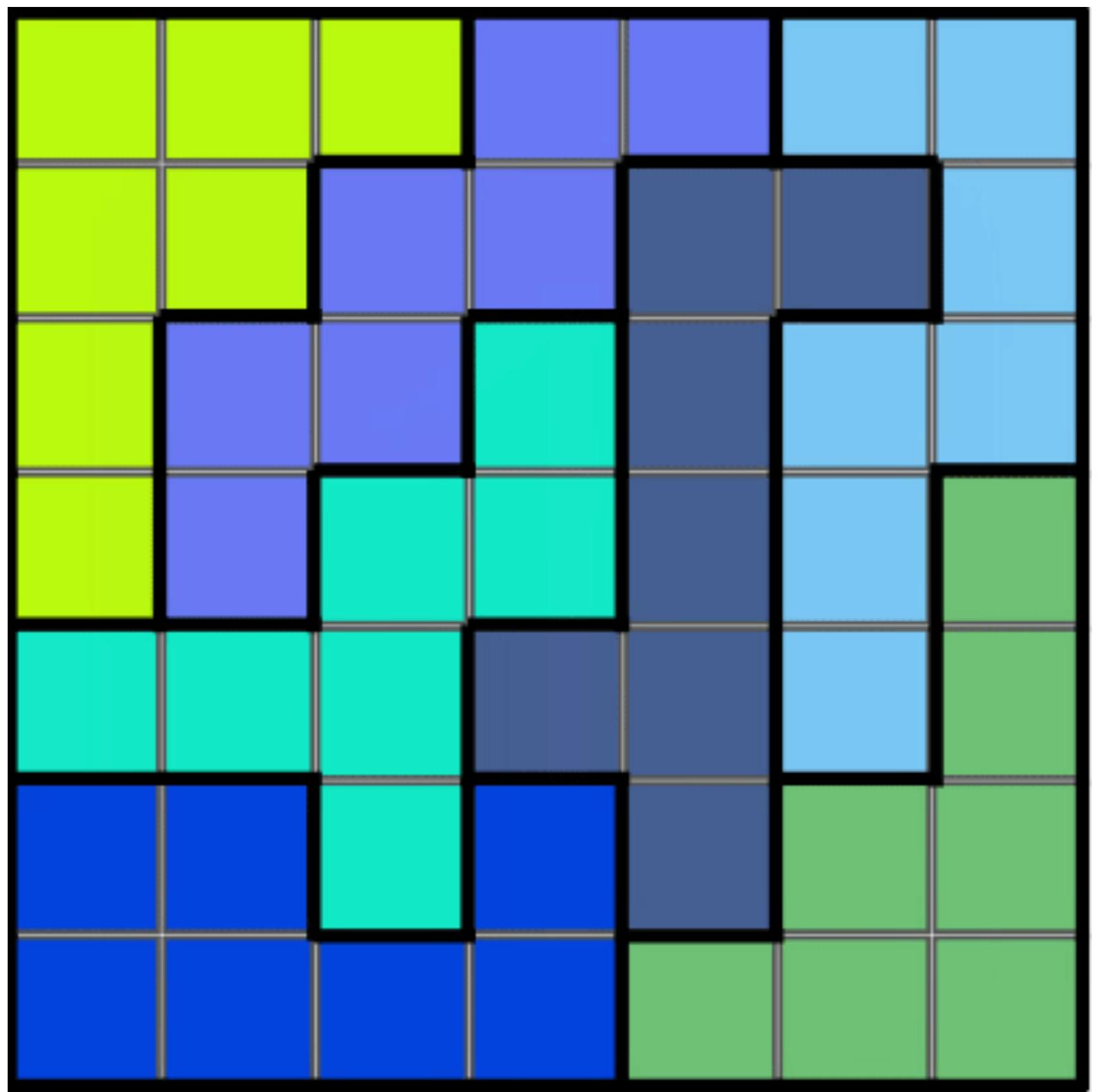
blocks

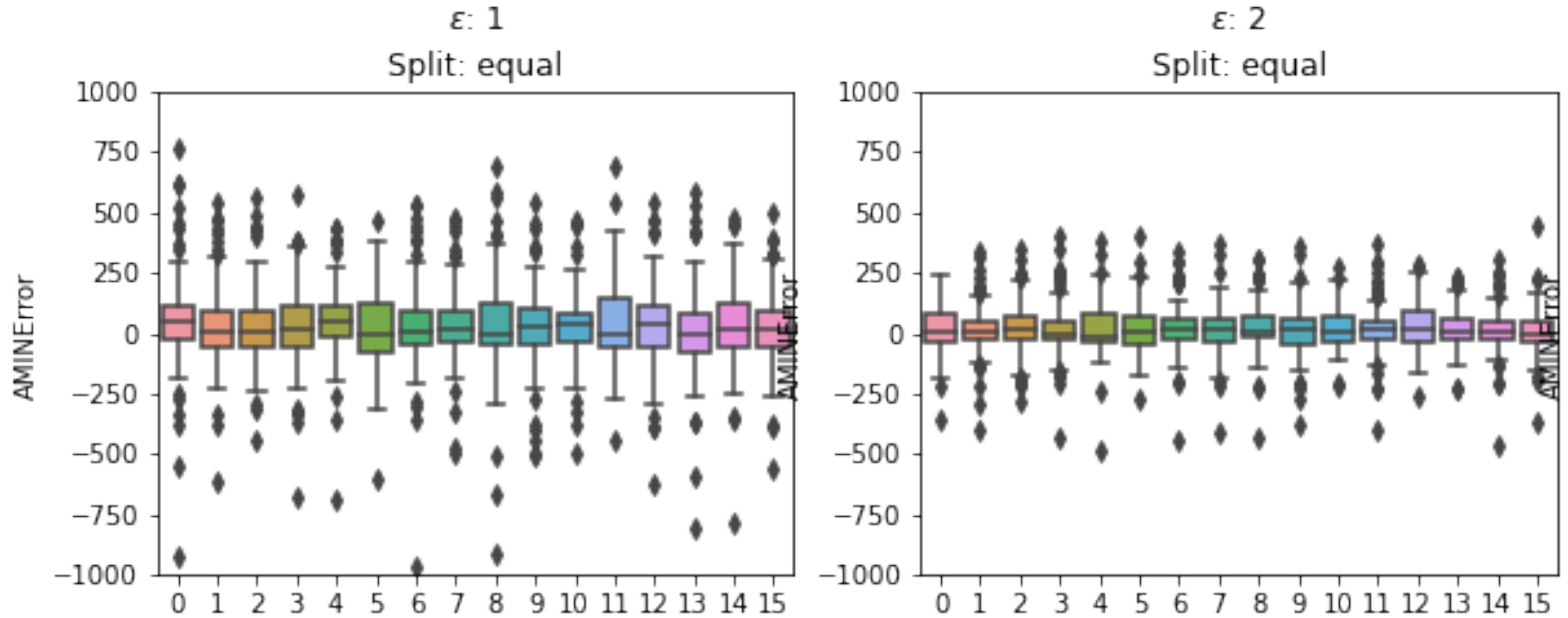
let's see some

experiments

we'll use a simplified model called "**ToyDown**" – see mggg.org/dp

Do districts lose Native population?

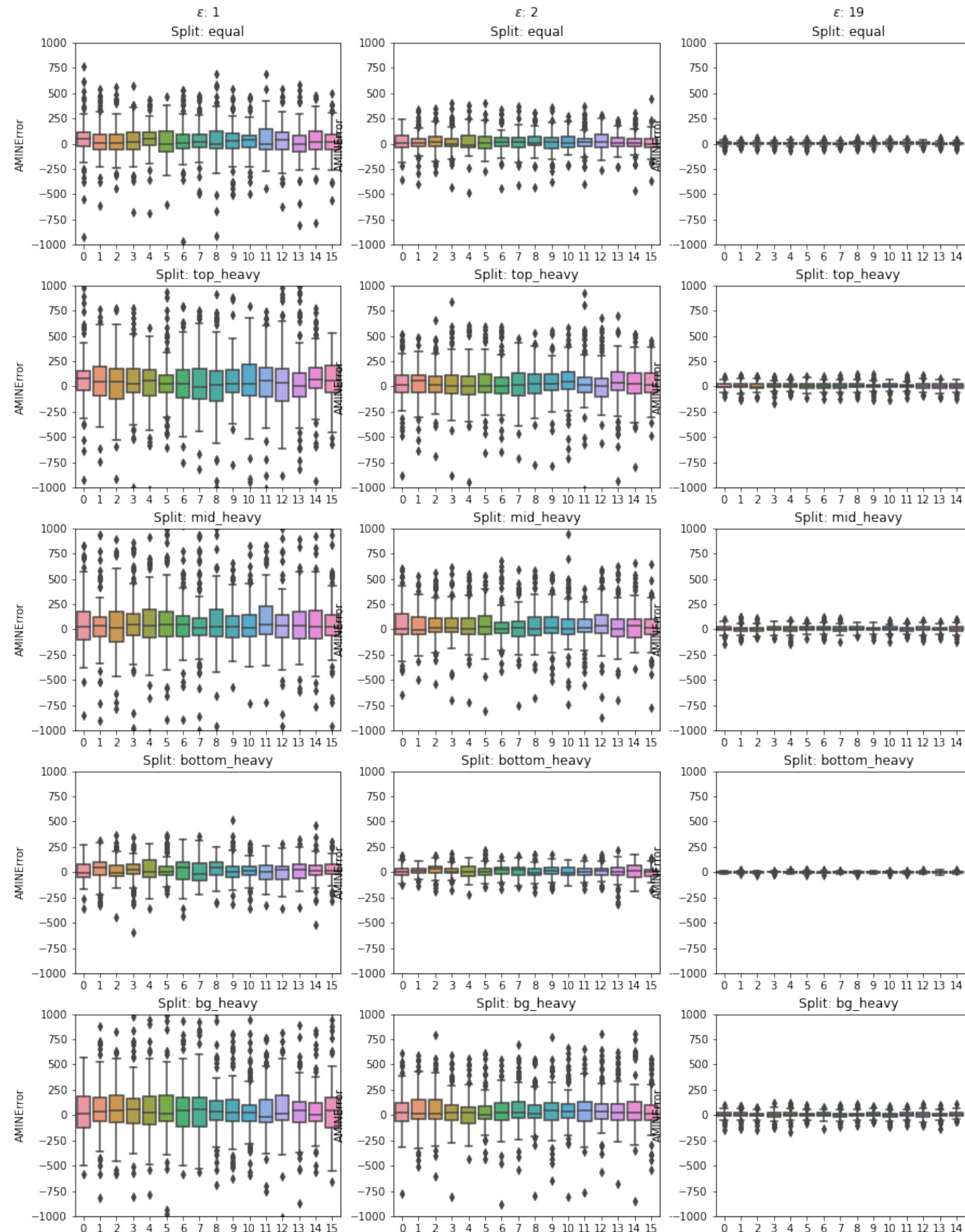




population distortions already very small (half percent) with $\varepsilon = 1, 2$

...truly tiny at $\varepsilon = 19$

$\epsilon = 1, 2, 19$



Navajo County

k=5 districts, population 20K

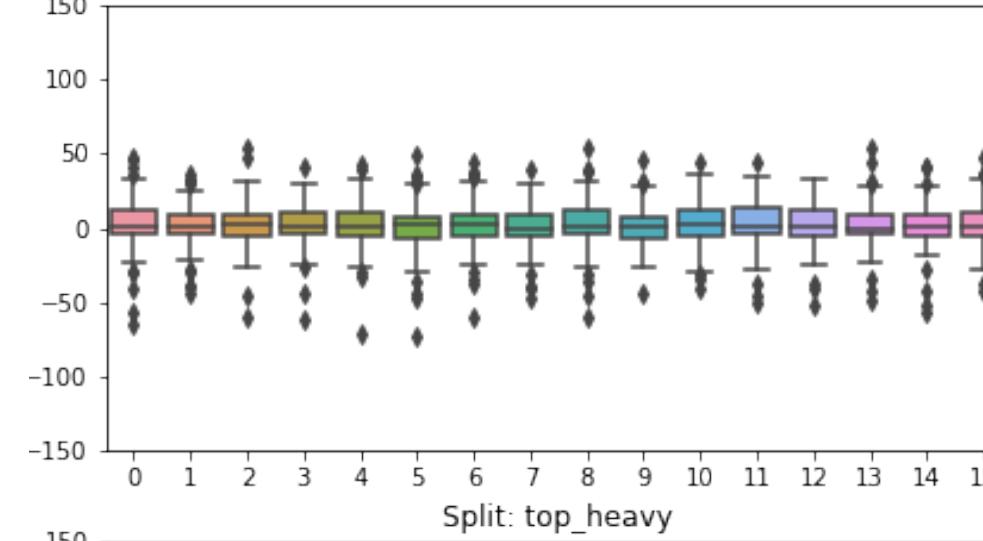
these plots show the discrepancy introduced by top-down style differential privacy

we made 100 random districts and noised them 16 times, then measured the error in the American Indian/Native American population total

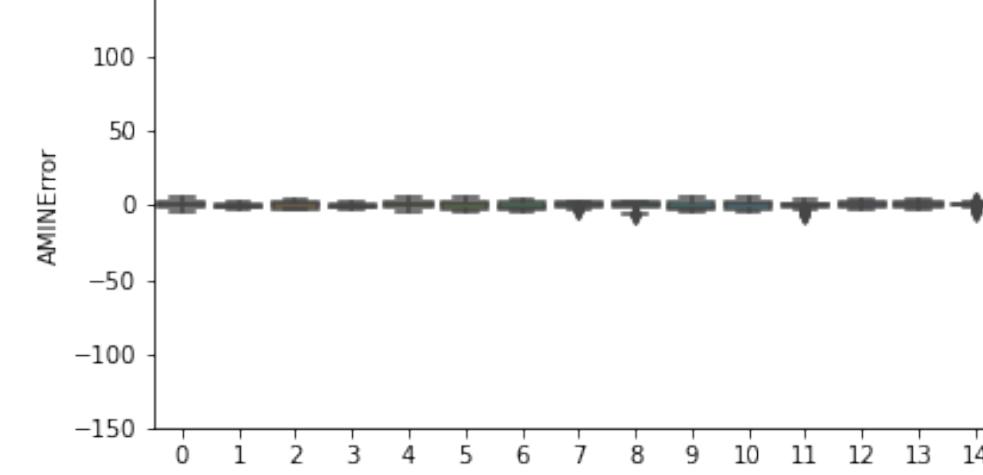
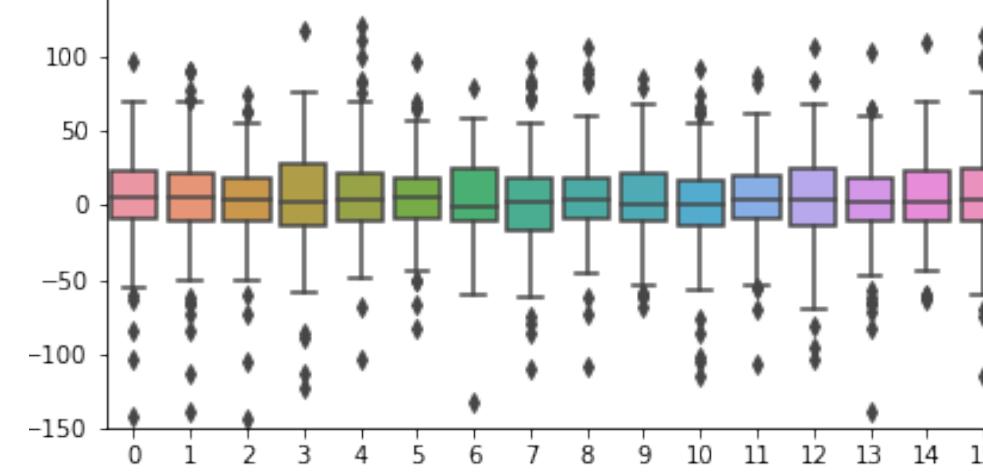
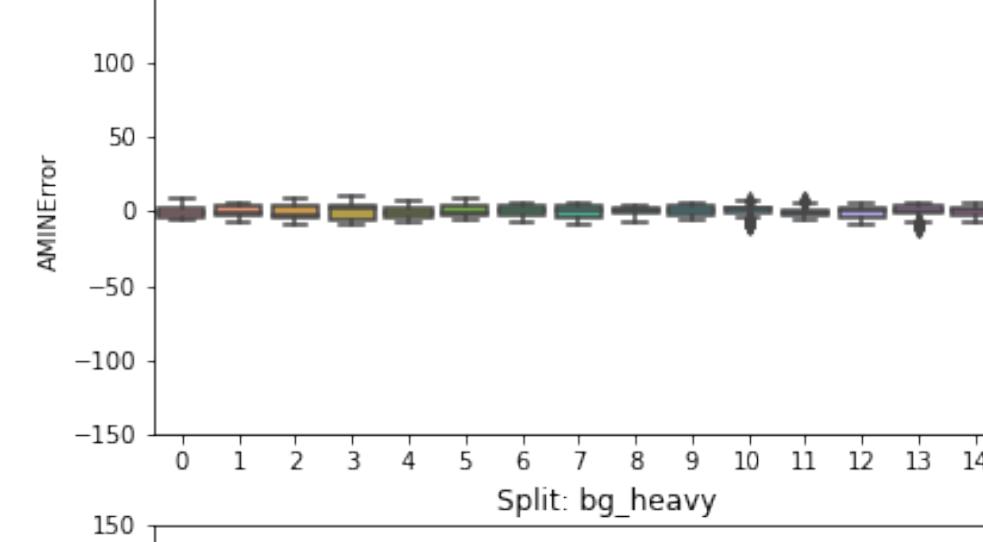
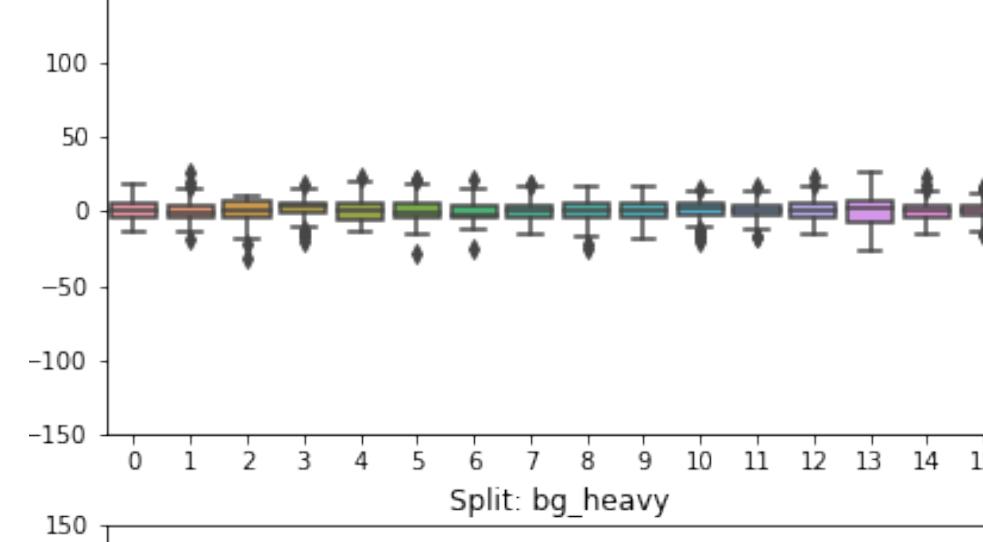
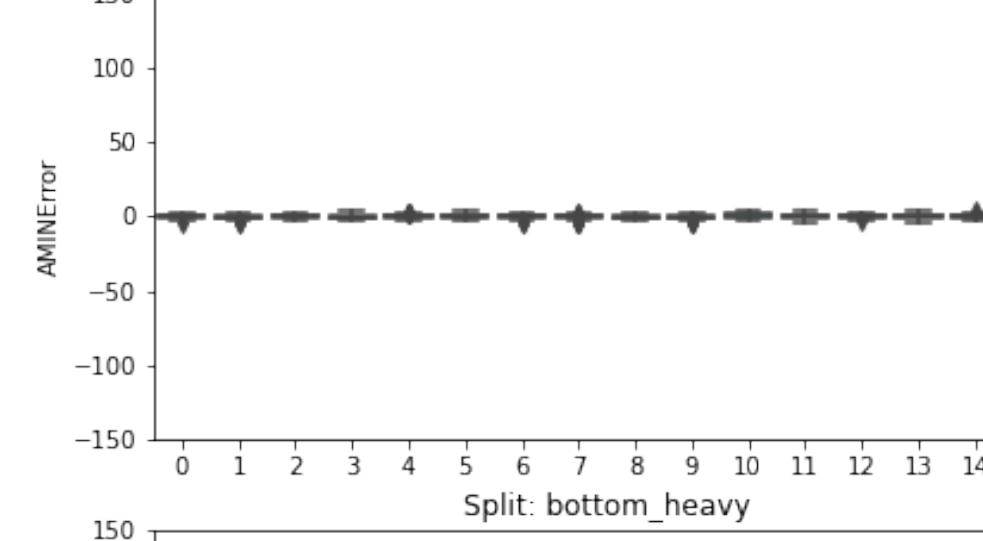
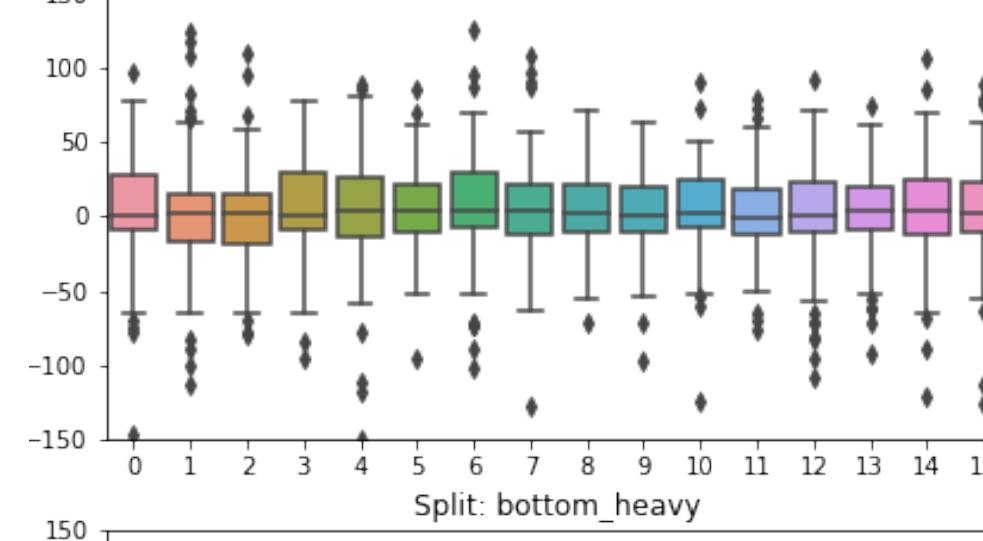
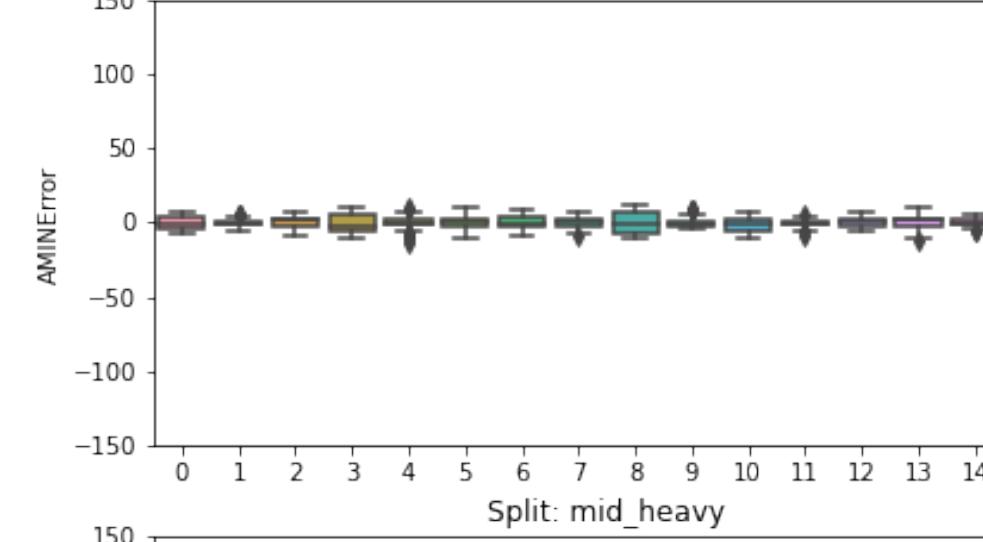
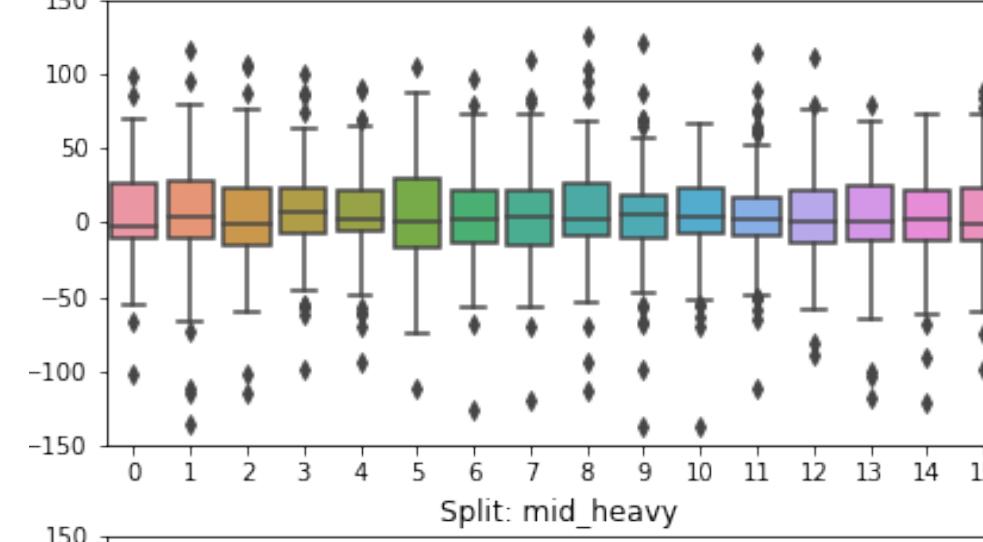
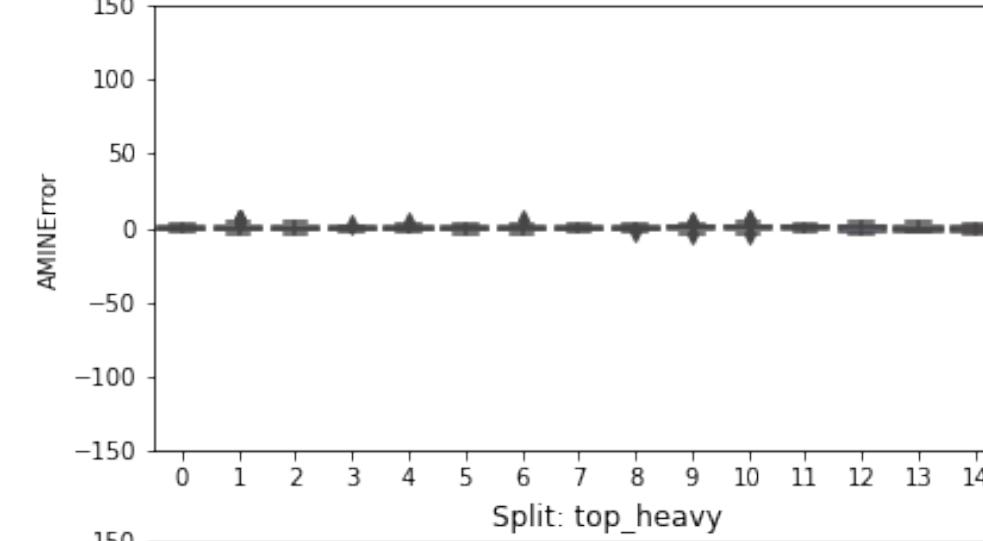
even with $\epsilon = 1$, the typical discrepancy is under 500

with $\epsilon = 19$, the typical discrepancy is under 5 people

District Type: block_recom
Split: equal



District Type: bg_recom
Split: equal



Navajo County

**built from blocks
vs. block groups**

k=5 districts, population 20K

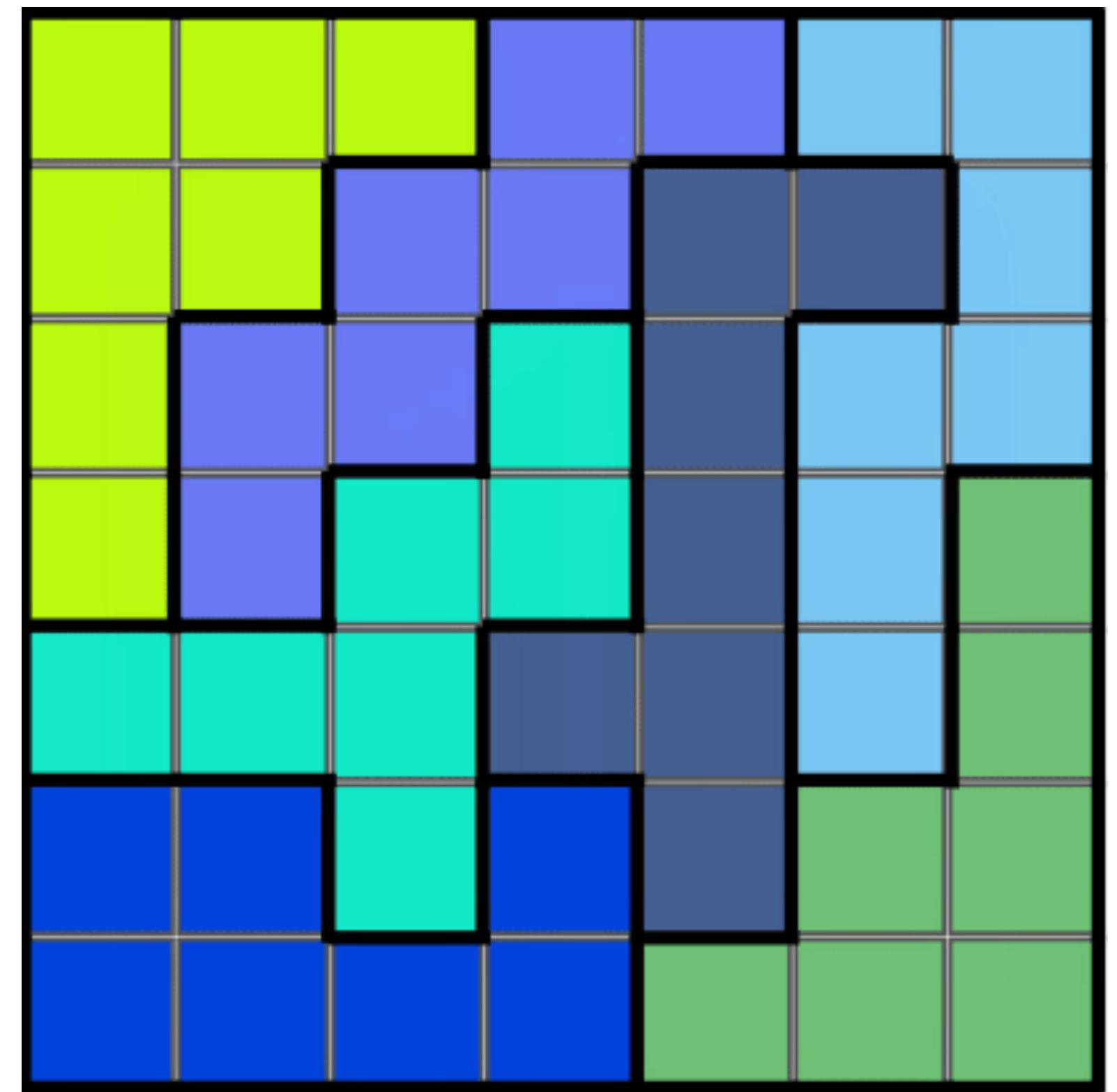
these plots show the discrepancy introduced by top-down style differential privacy

we made 100 random districts and noised them 16 times, then measured the error in the American Indian/Native American population total

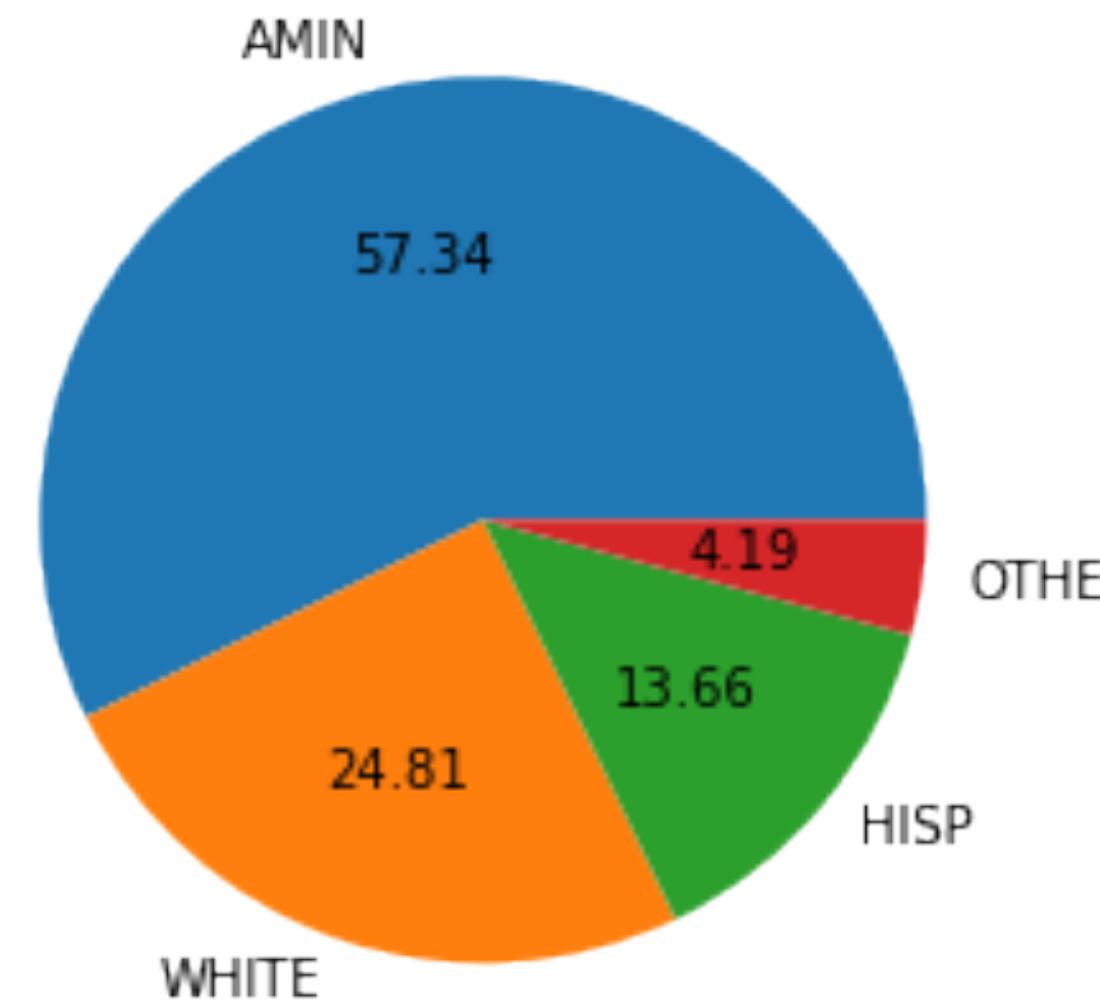
construction matters!

far better accuracy on districts built from larger pieces

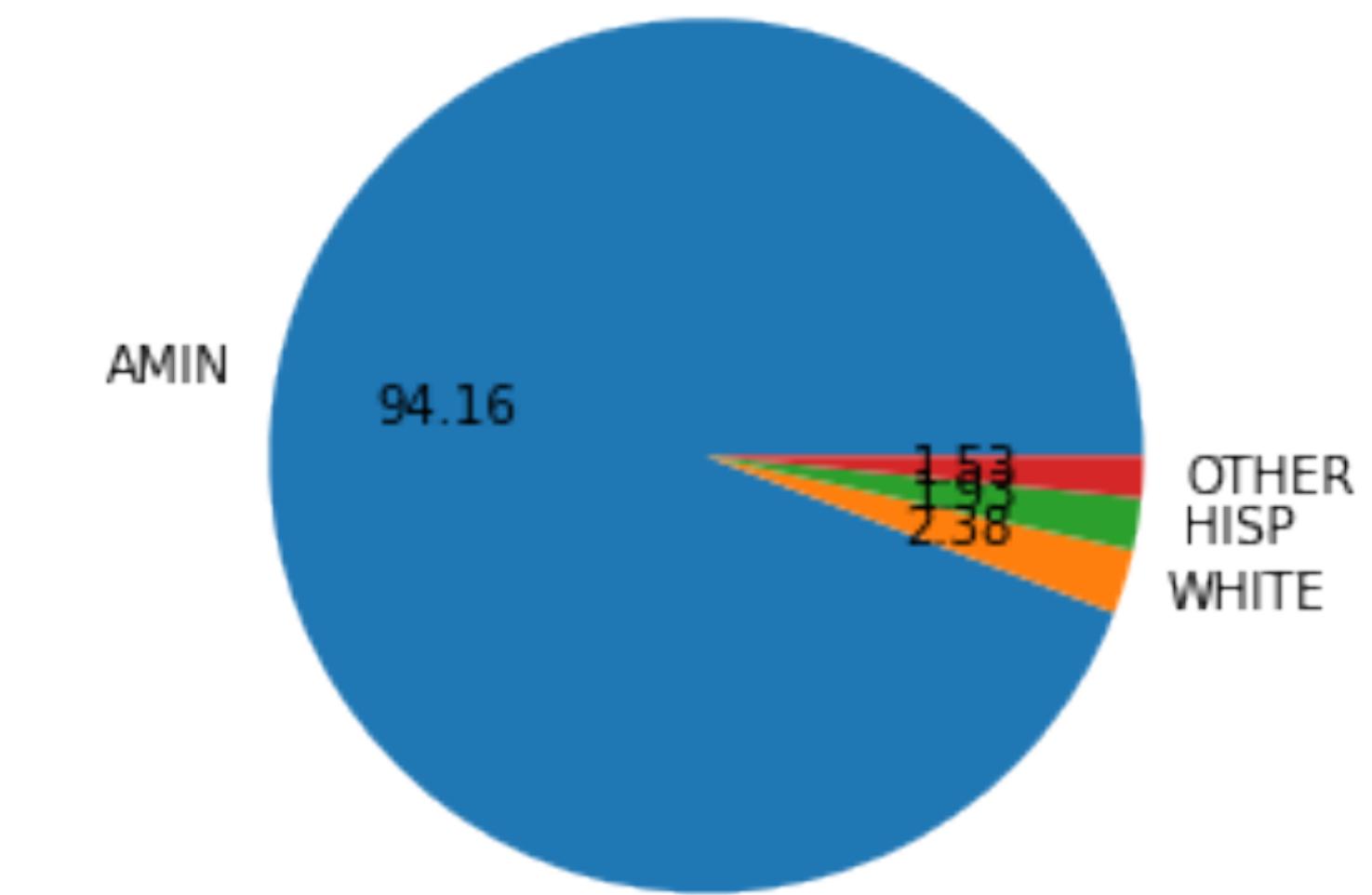
Do districts change their overall racial composition?



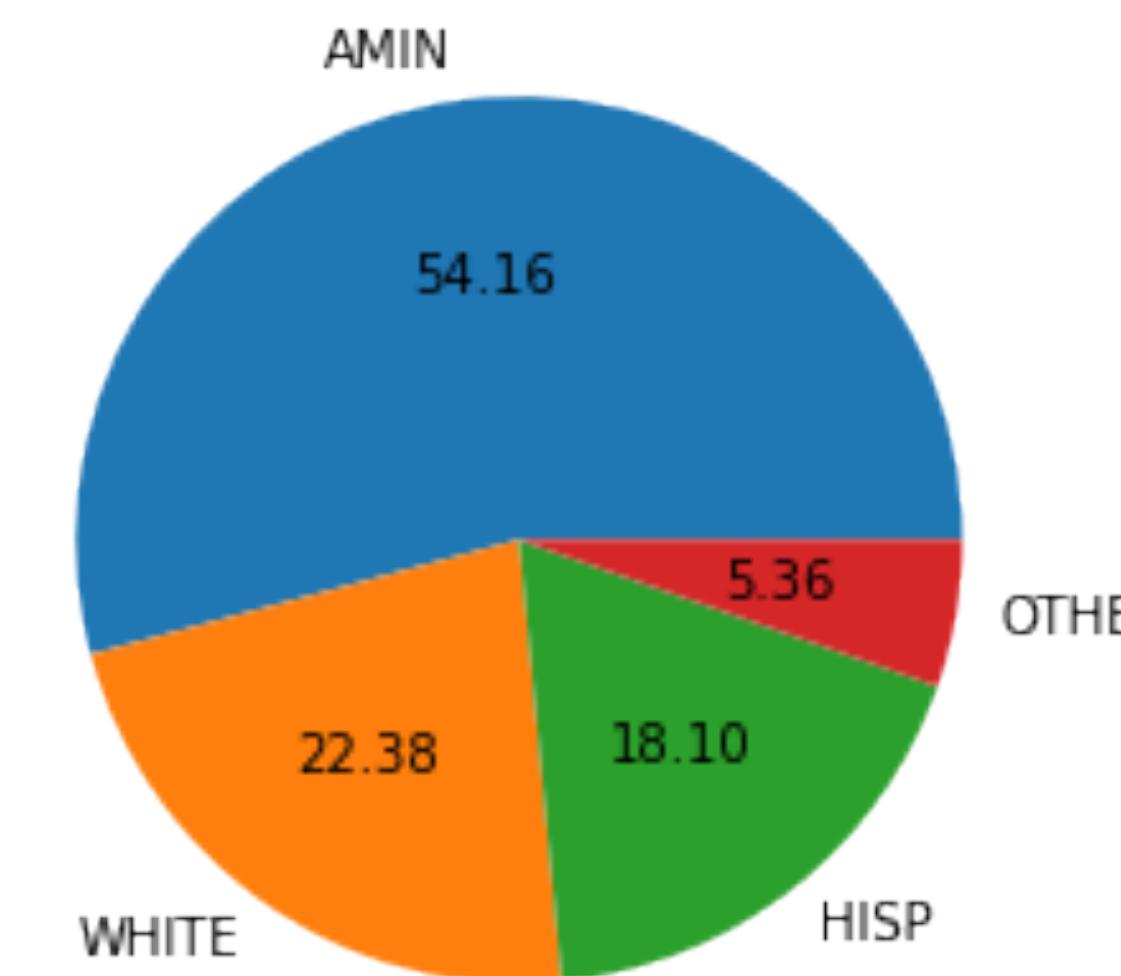
we will noise these
16 times with $\varepsilon = 2$
and equal allocation
over the
geographical levels



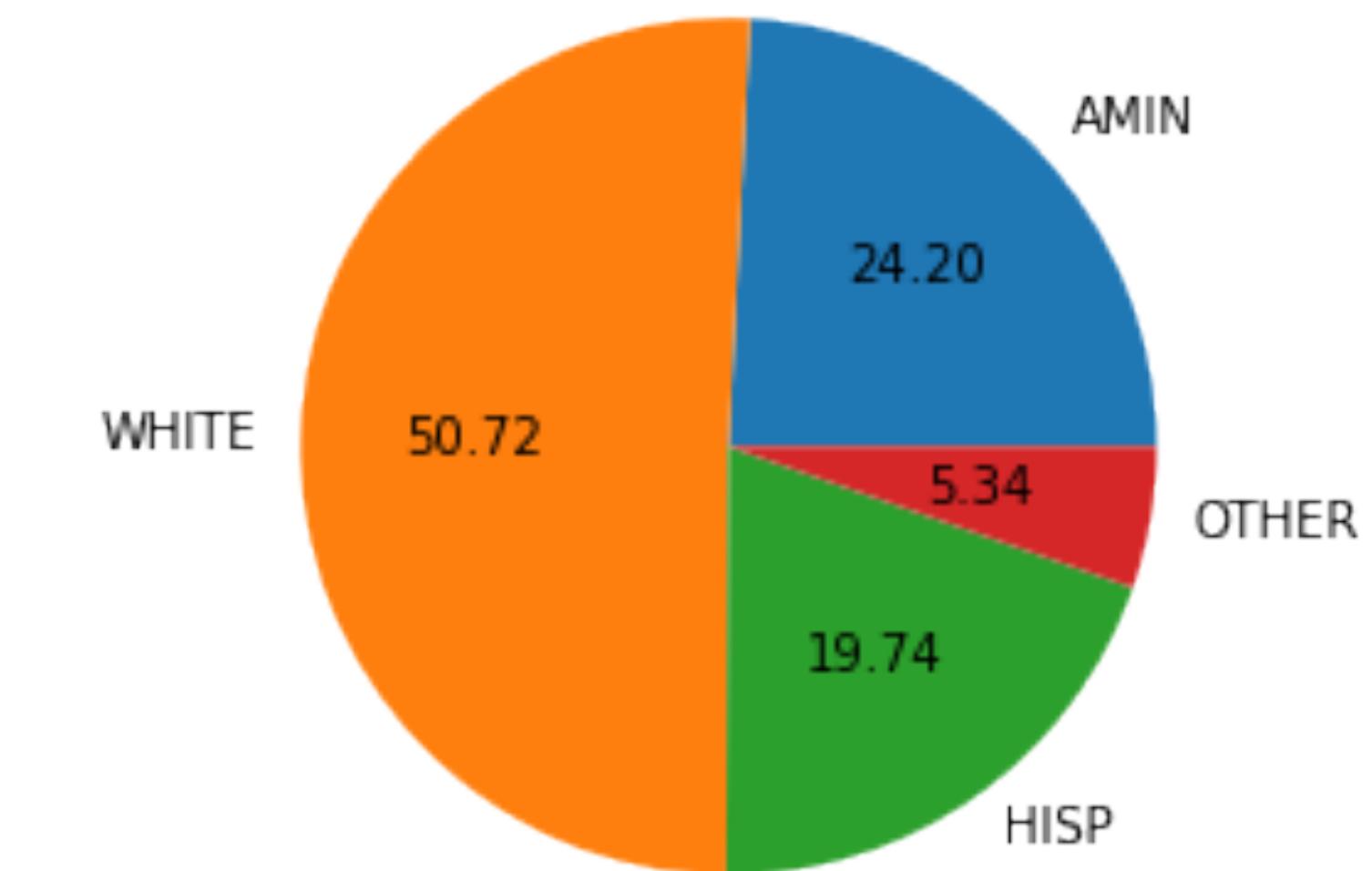
random district #2



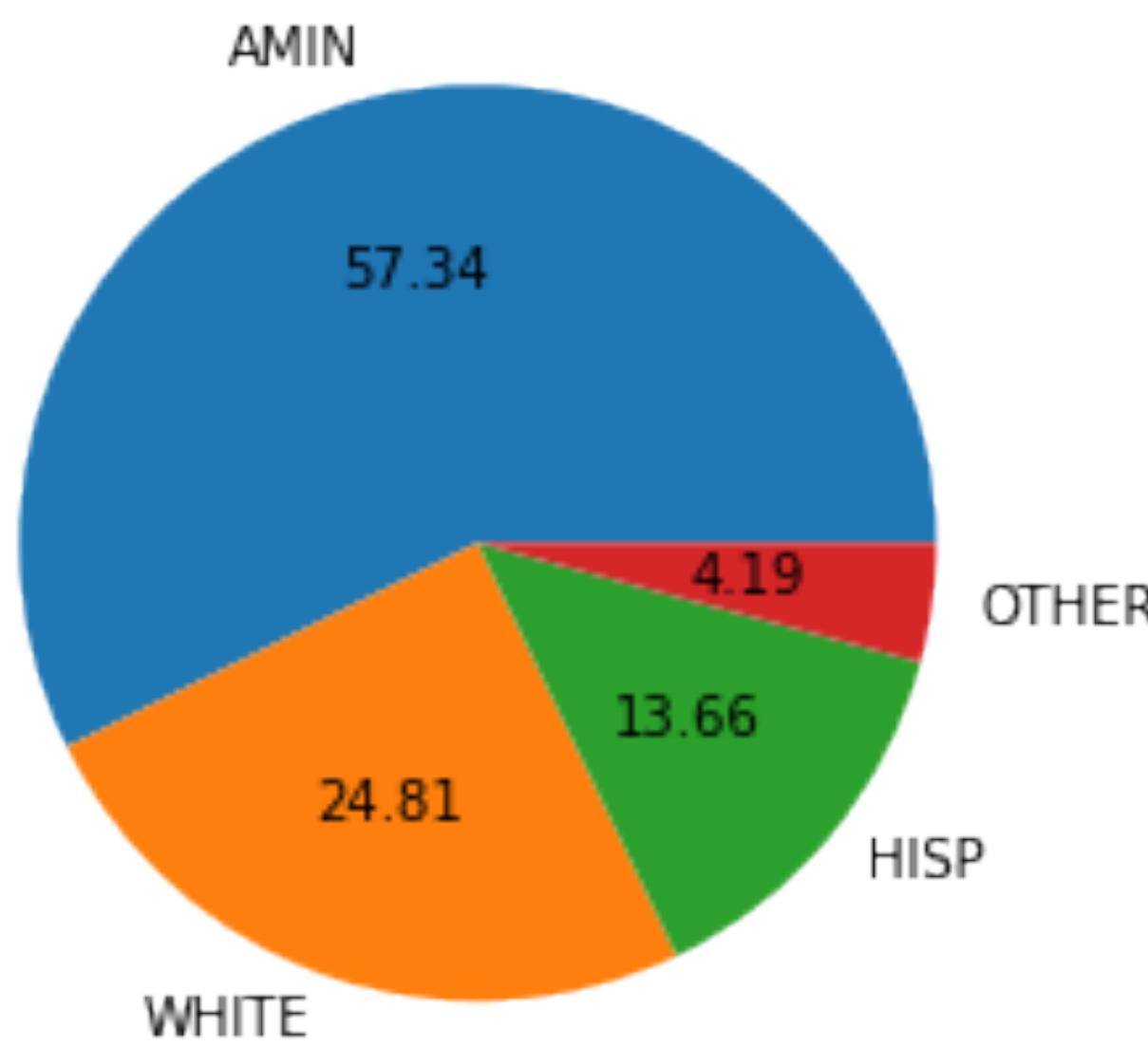
random district #9



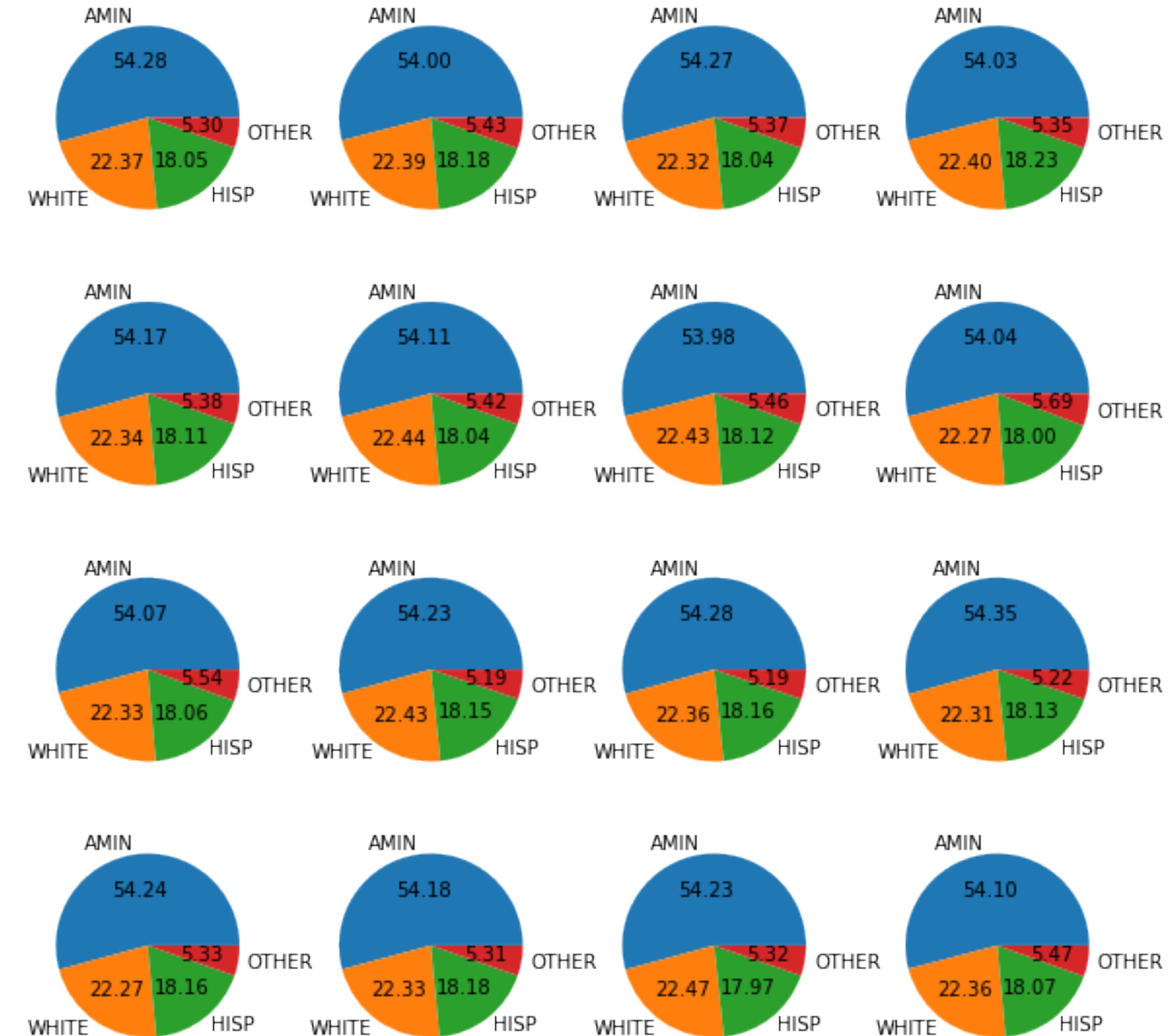
random district #13

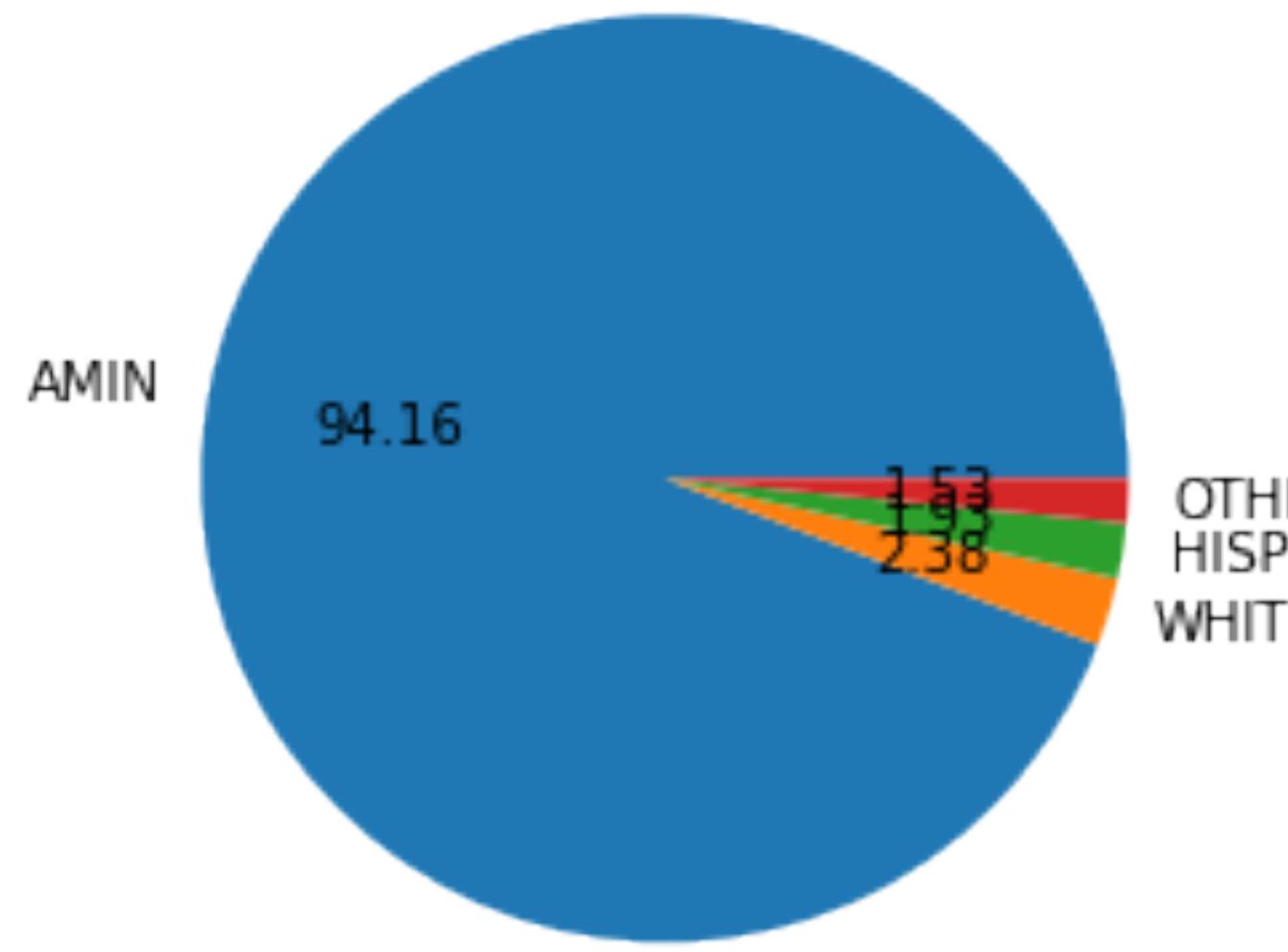


random district #46

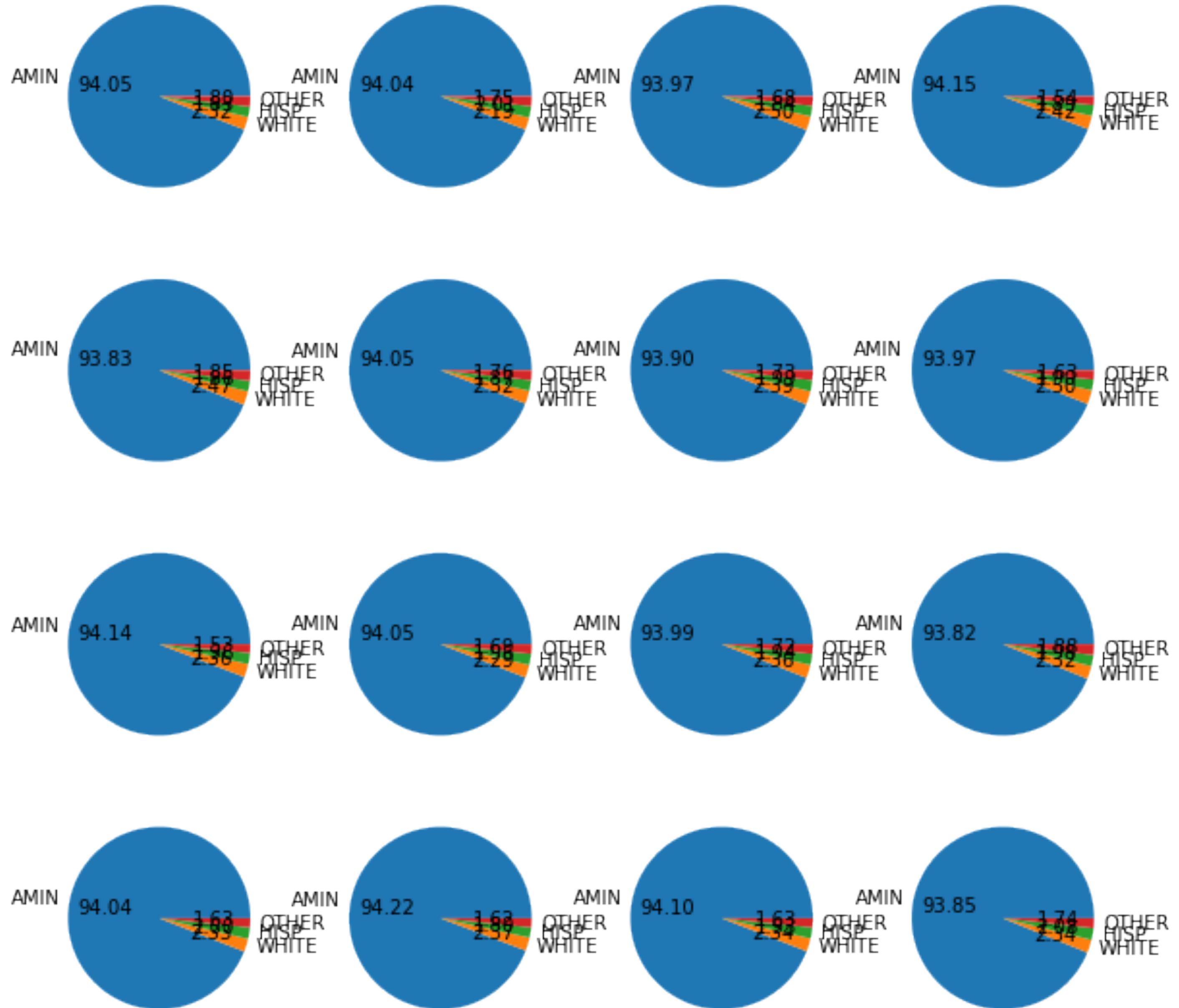


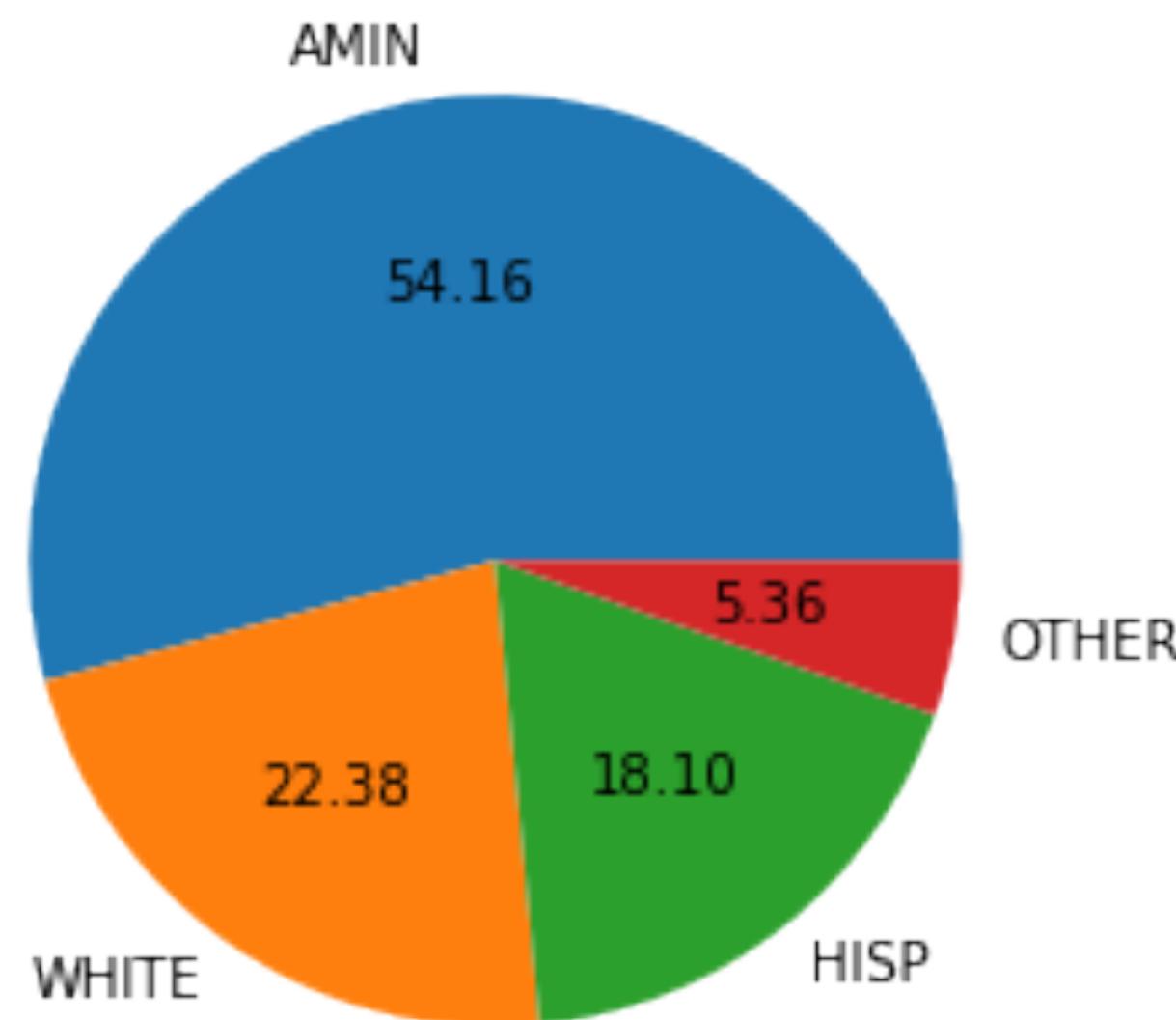
random district #2



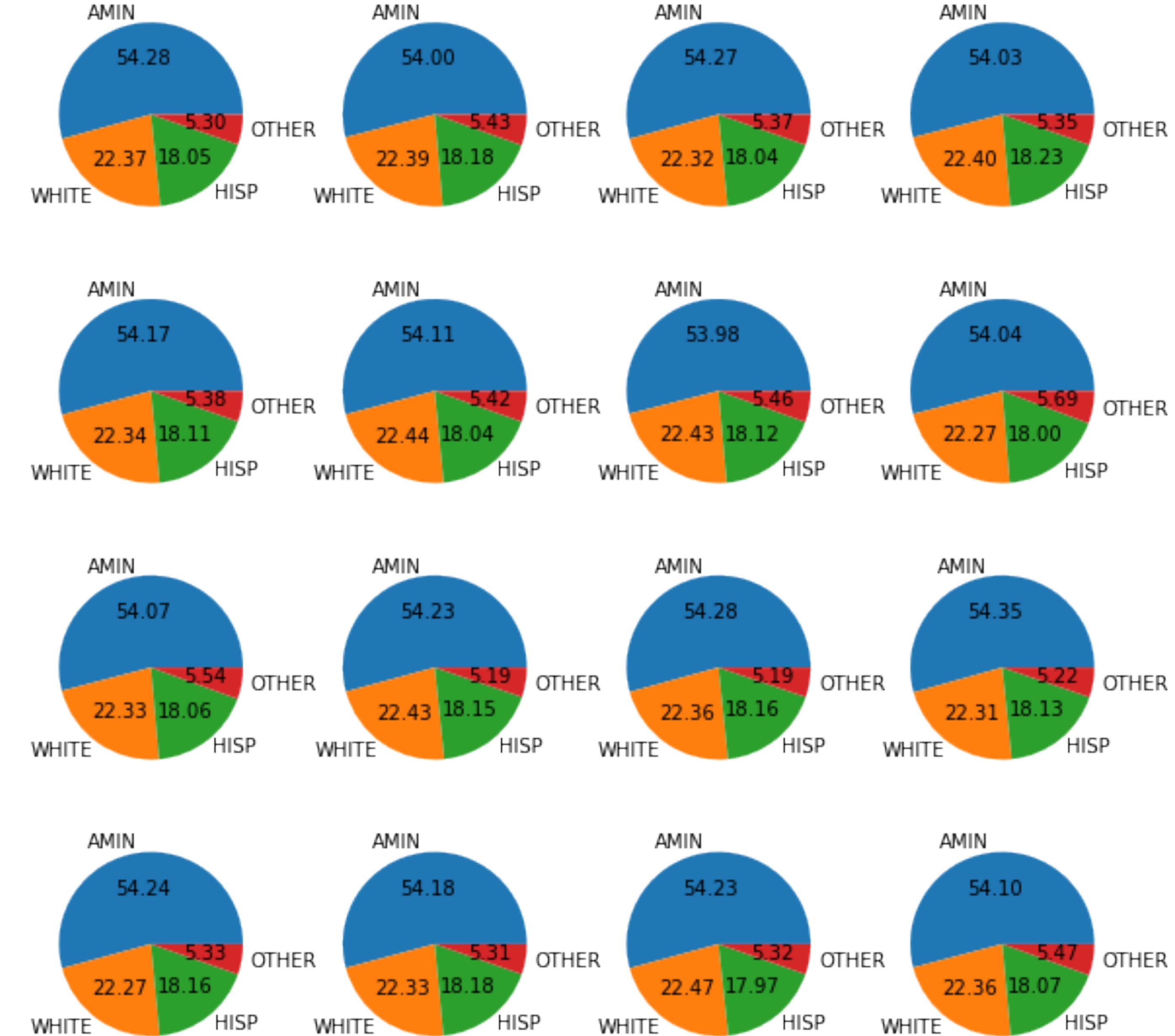


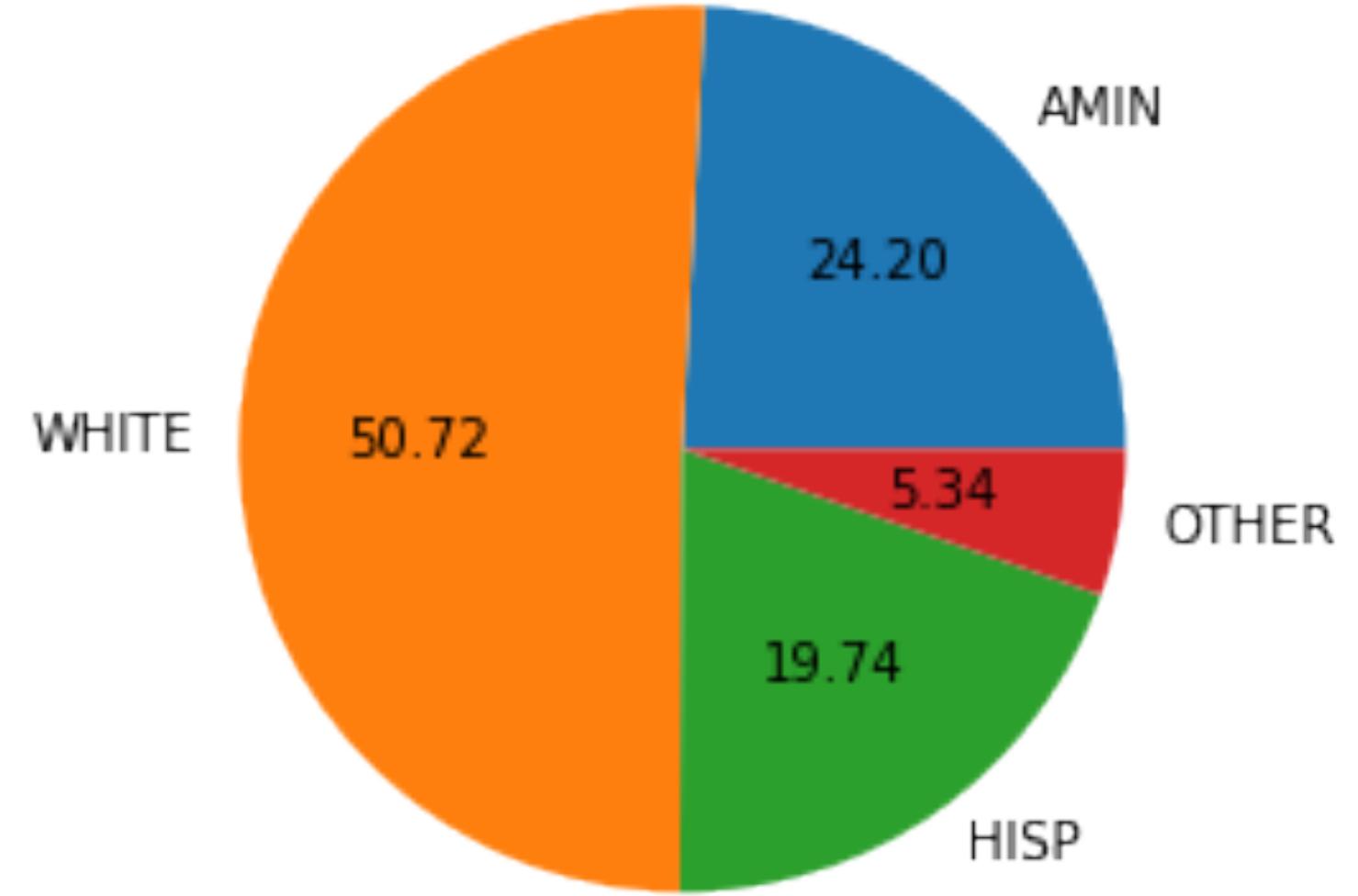
random district #9



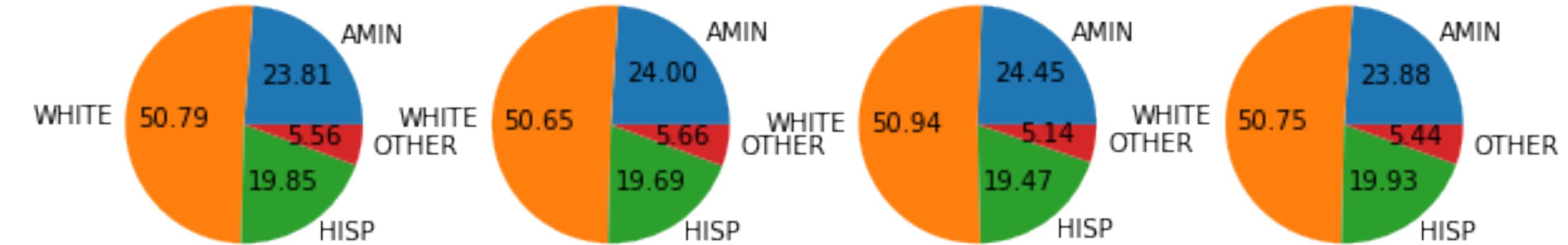


random district #13

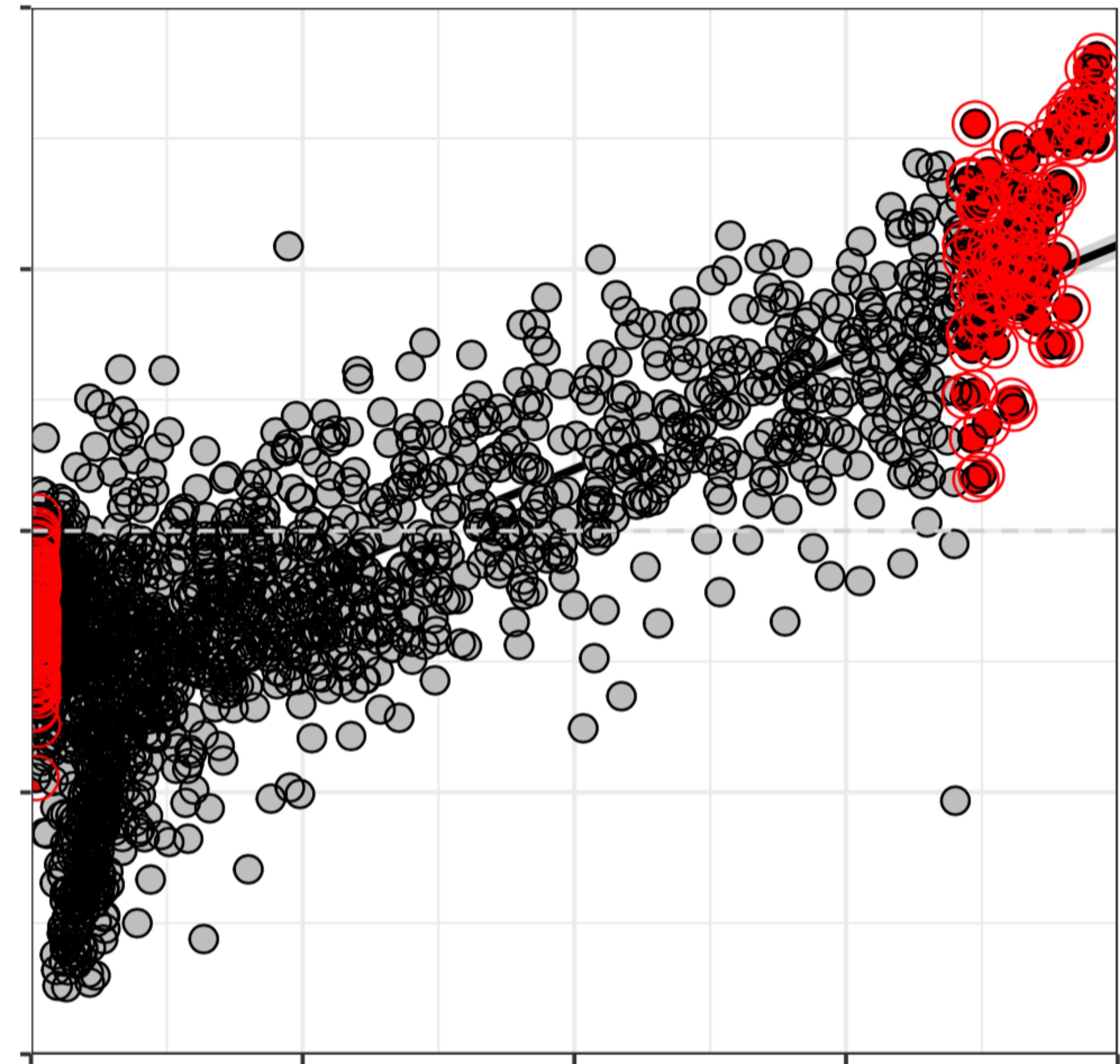




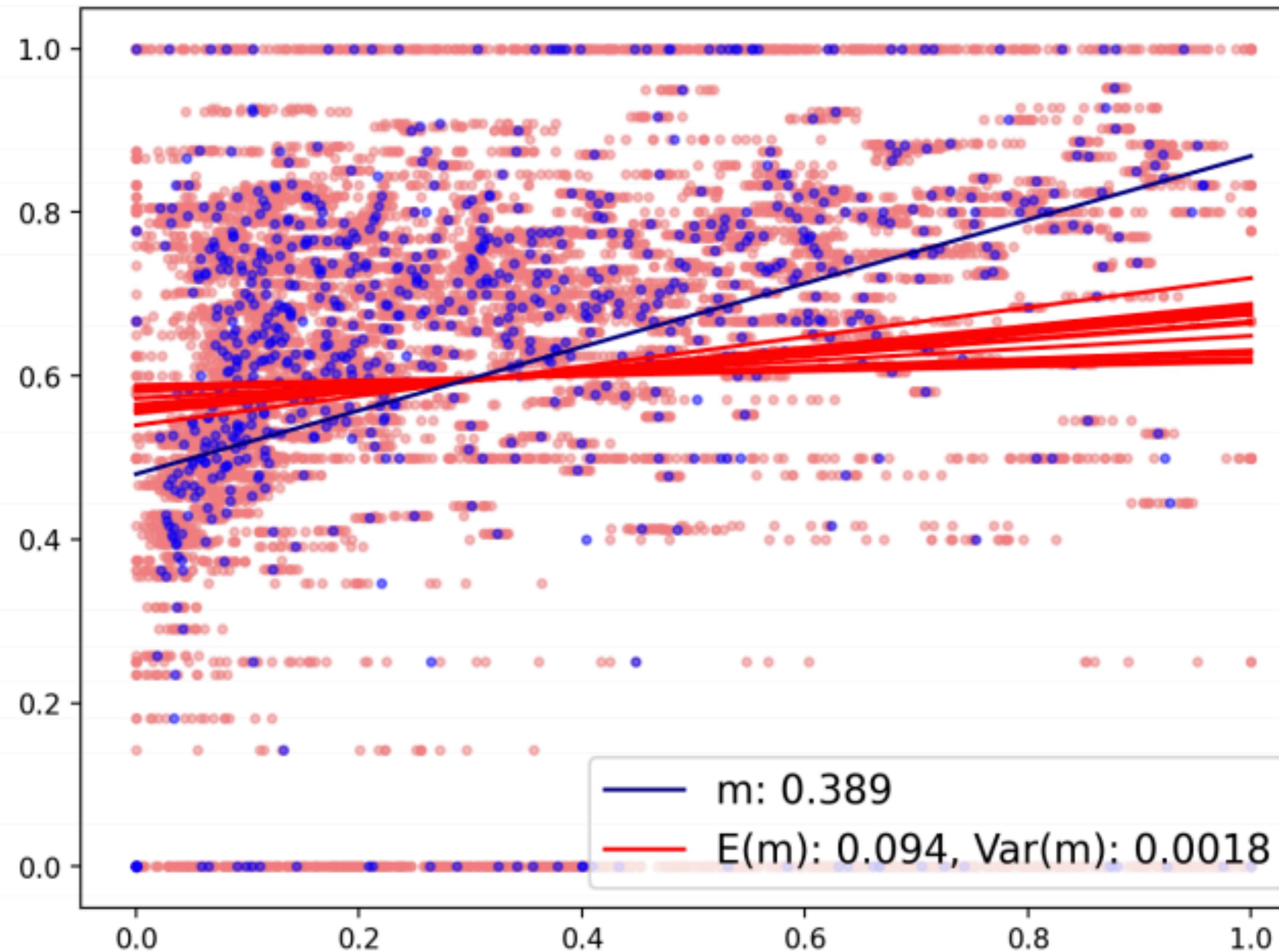
random district #46



Can we identify racially polarized voting?



blue: un-noised
pink dots: noisy data
red lines: lines fit to noisy data

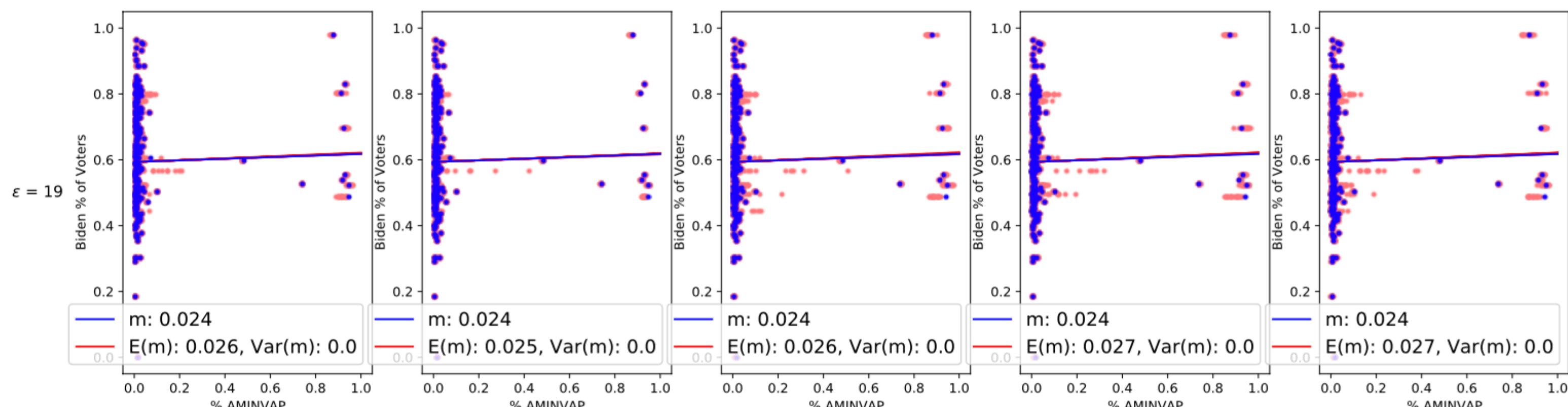
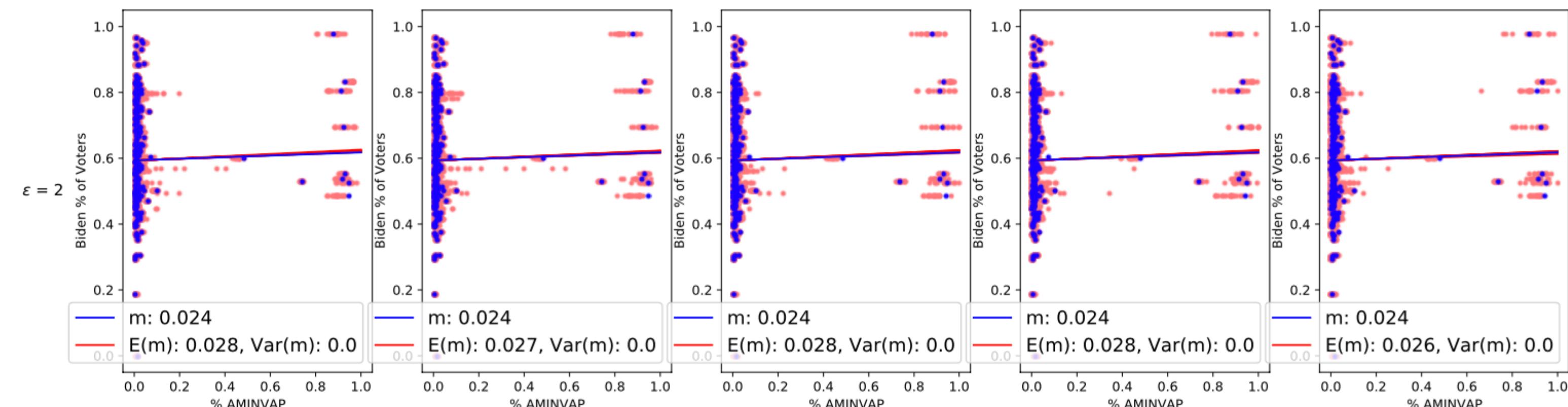
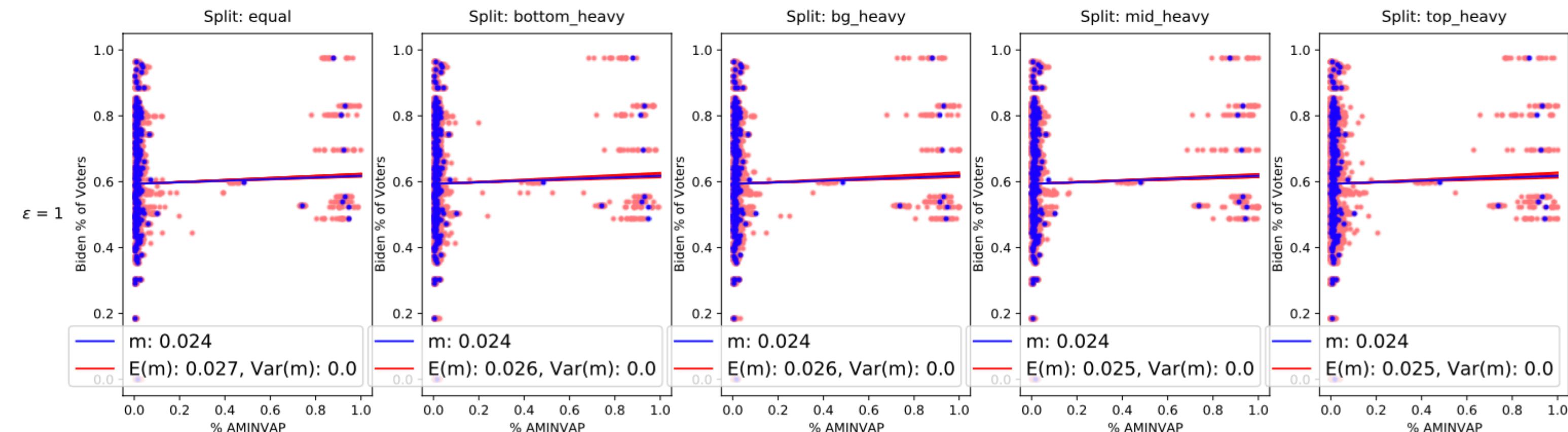
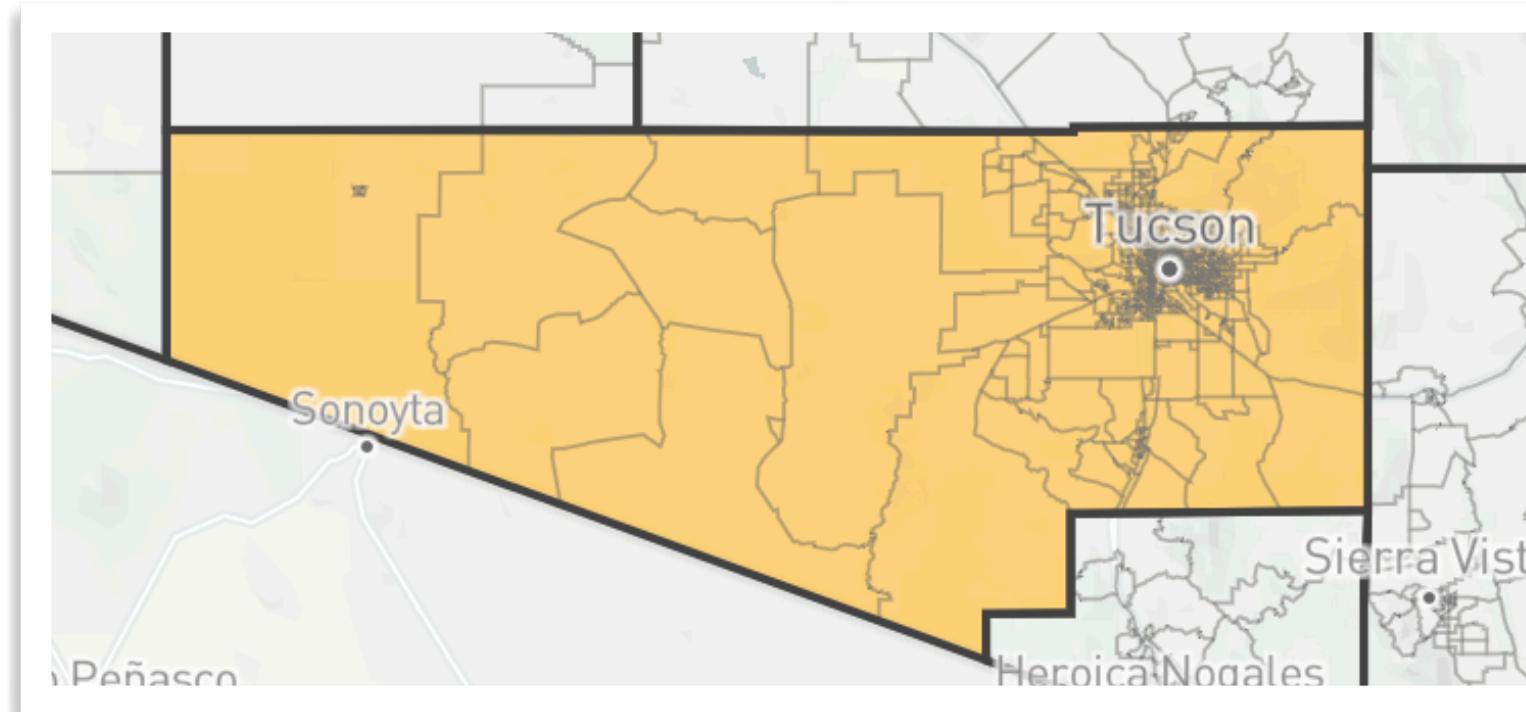


the nightmare scenario
adding noise loses the signal
of racially polarized voting
might be unable to test merit
of VRA claims



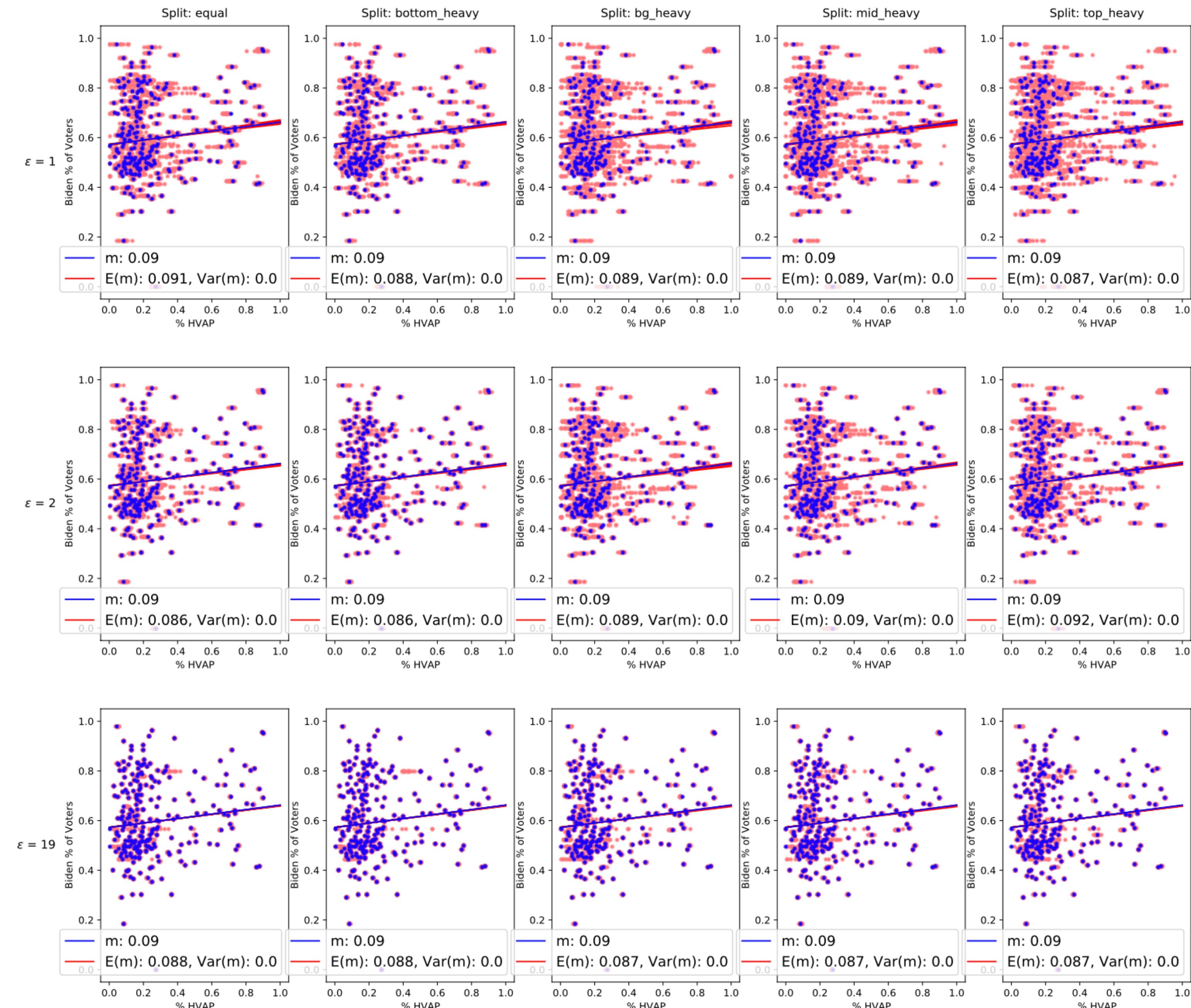
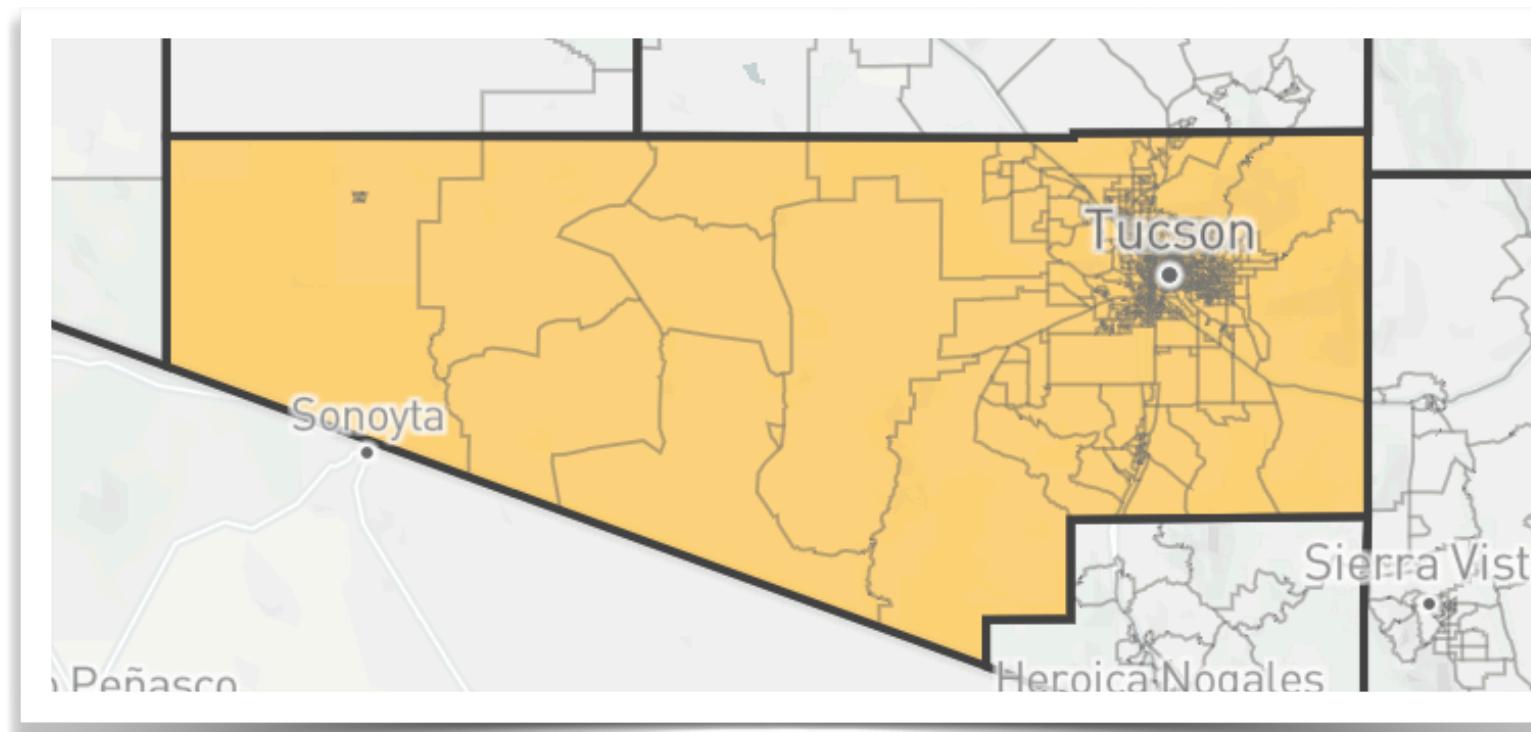
Pima County

AMIN support for Biden



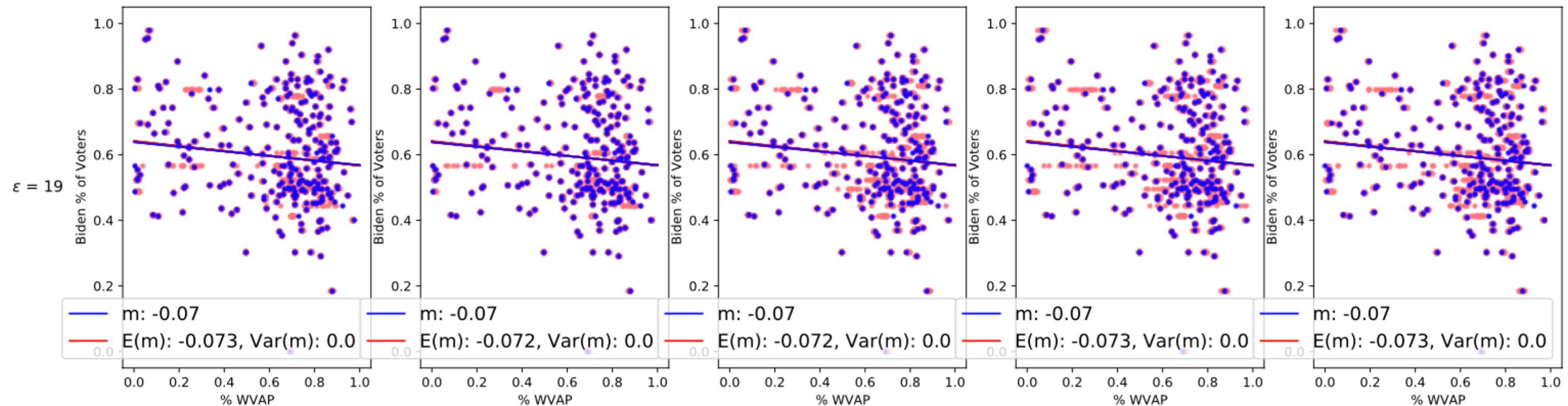
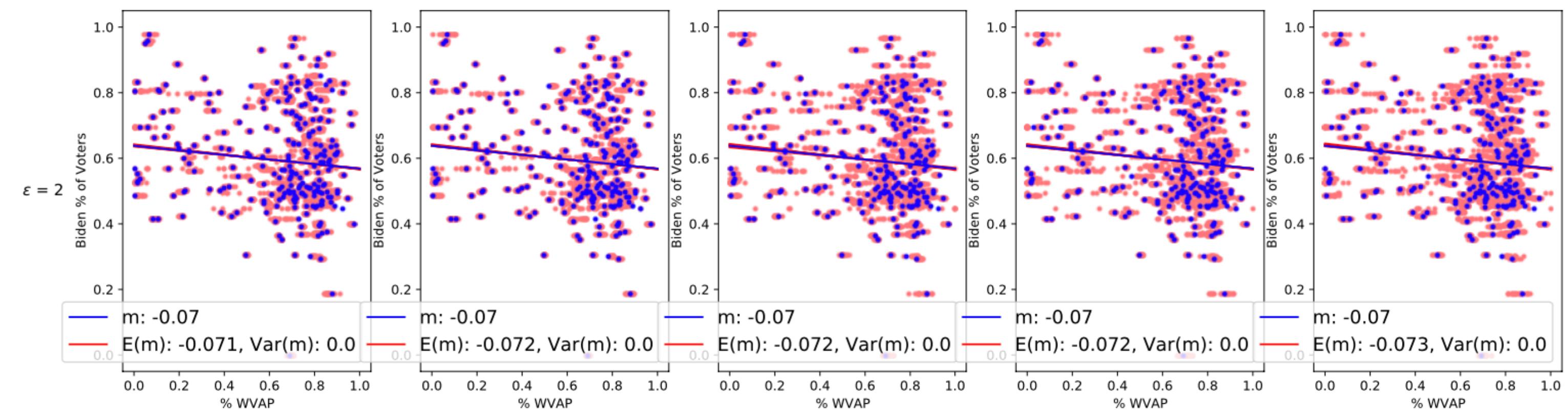
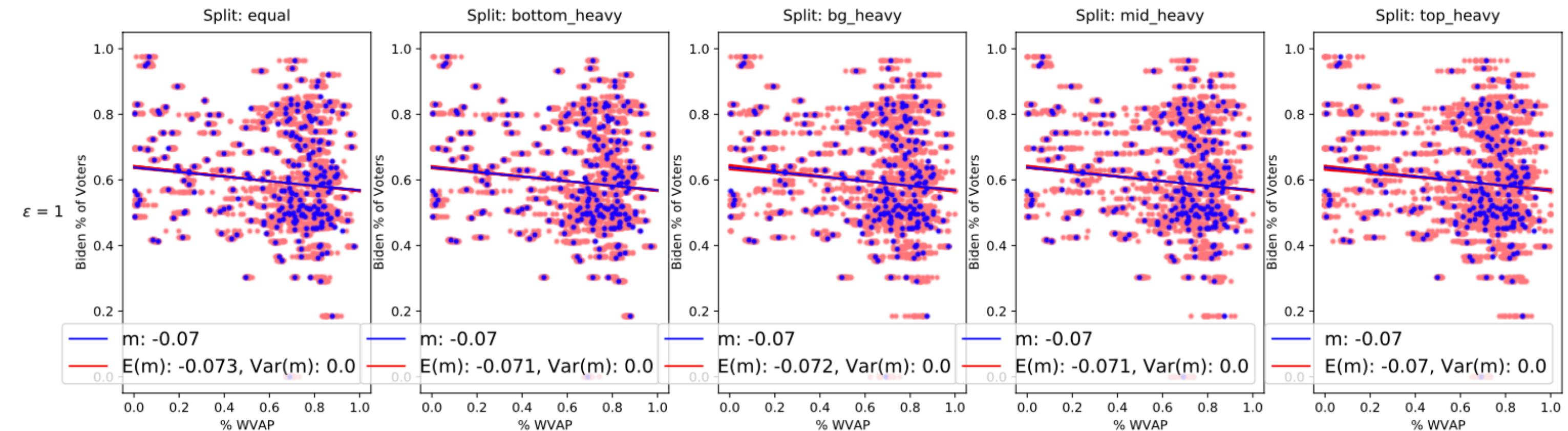
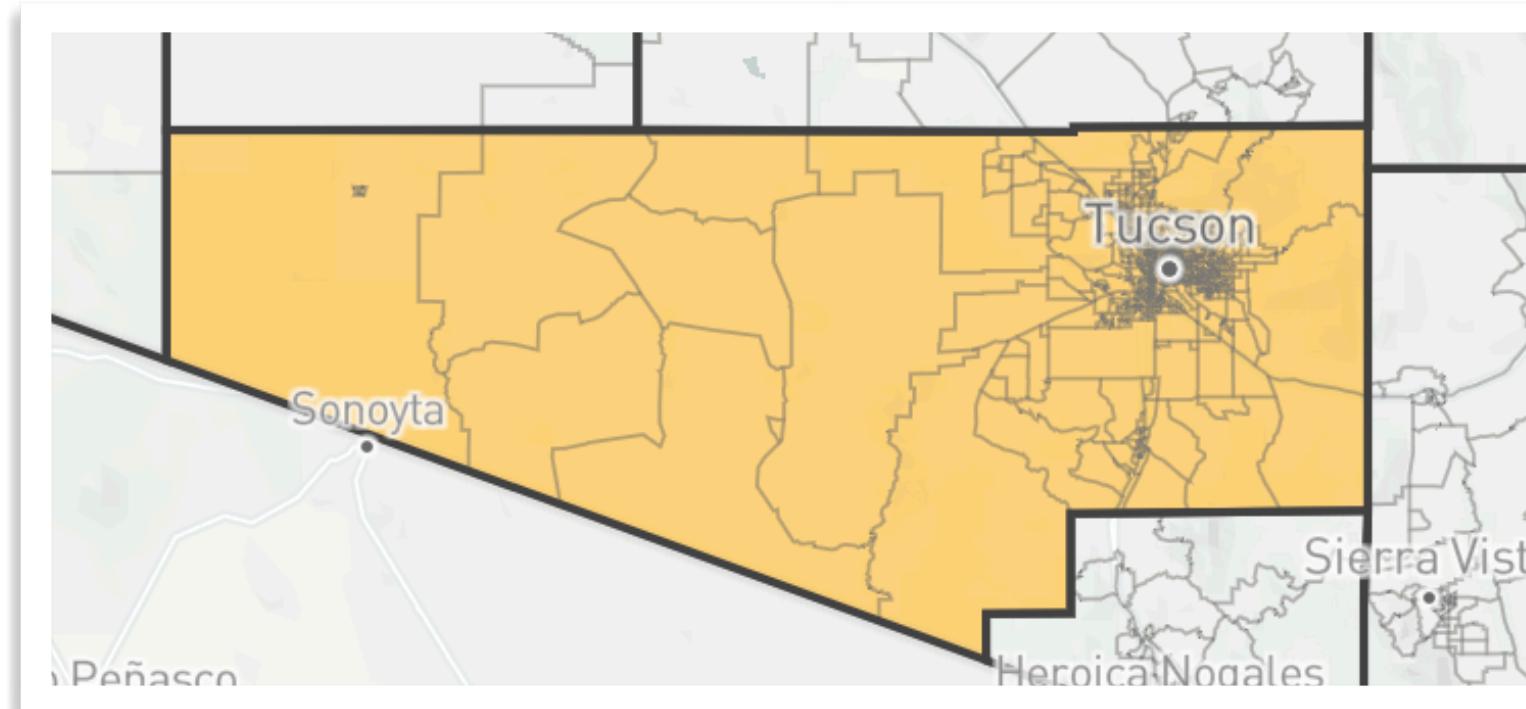
Pima County

HISP support for Biden



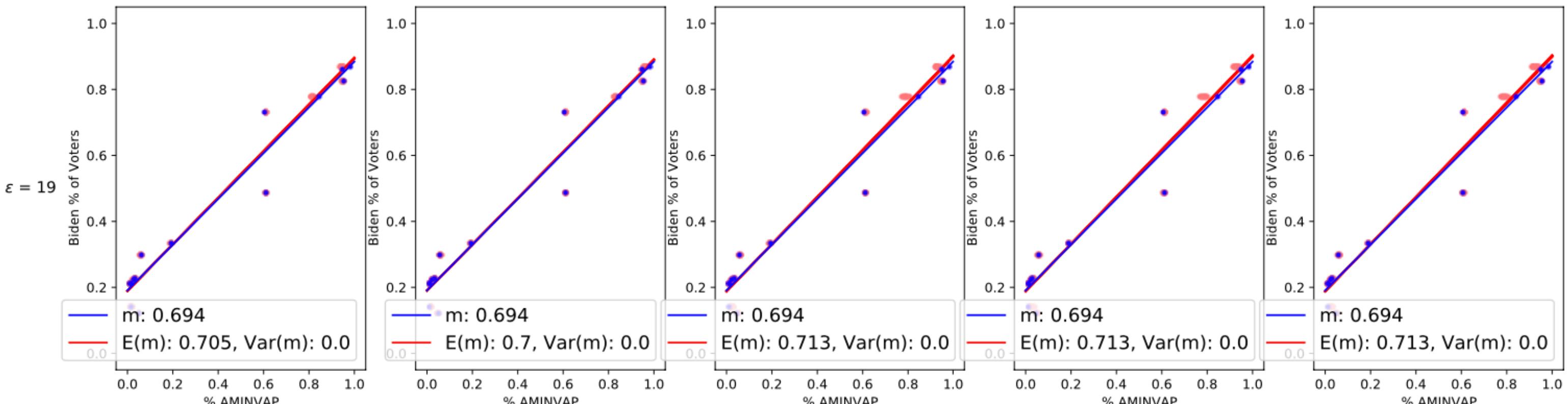
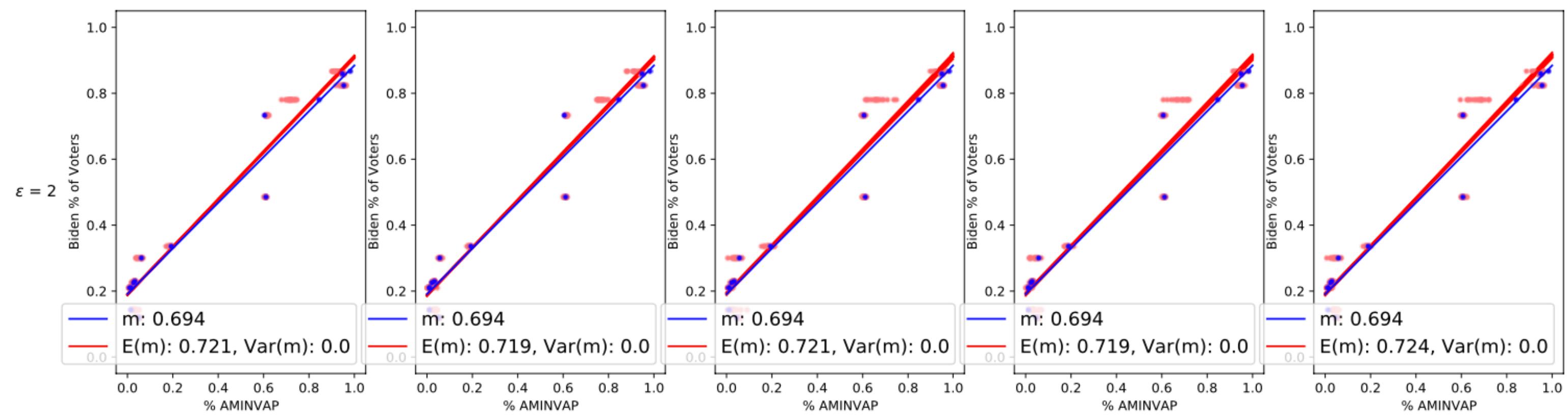
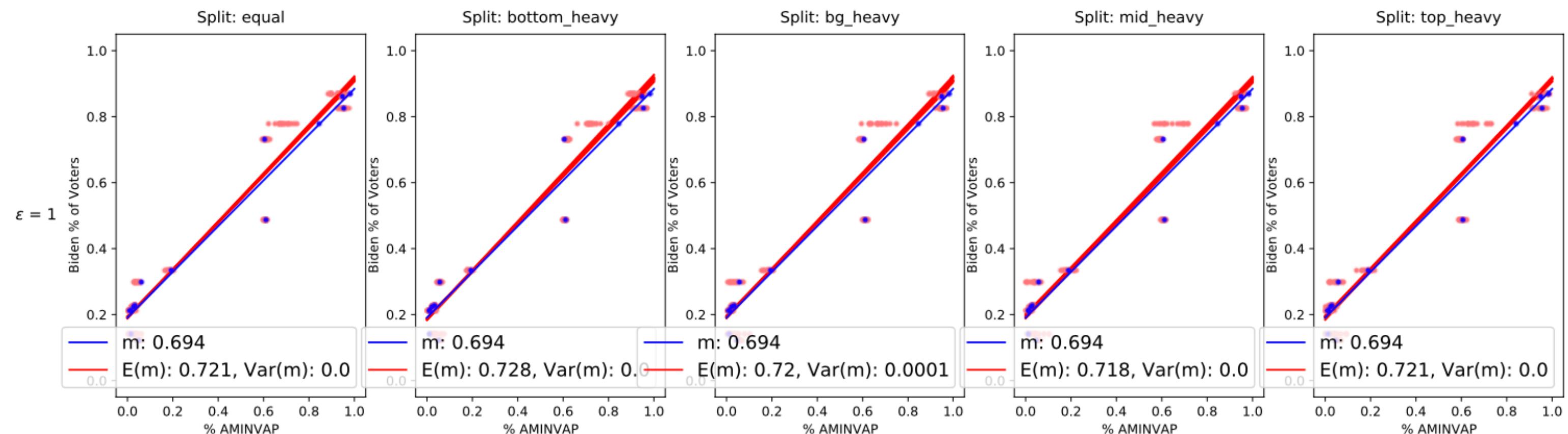
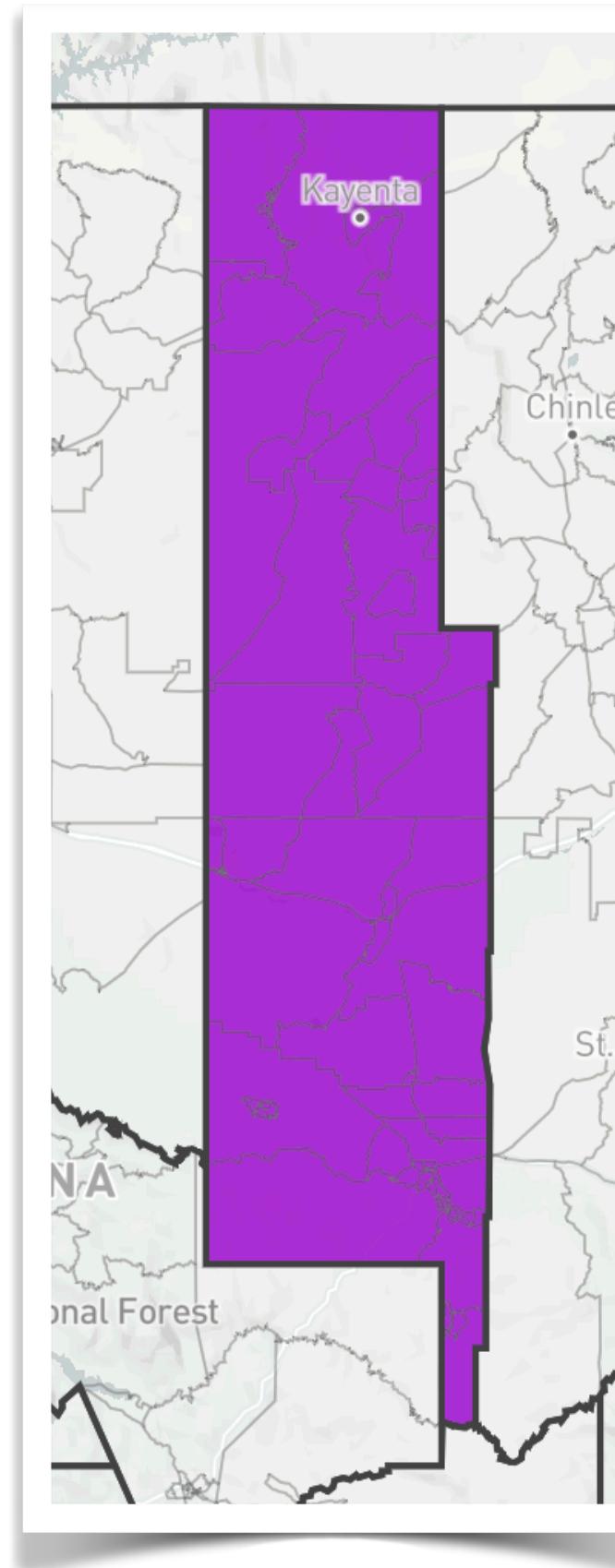
Pima County

W support for Biden



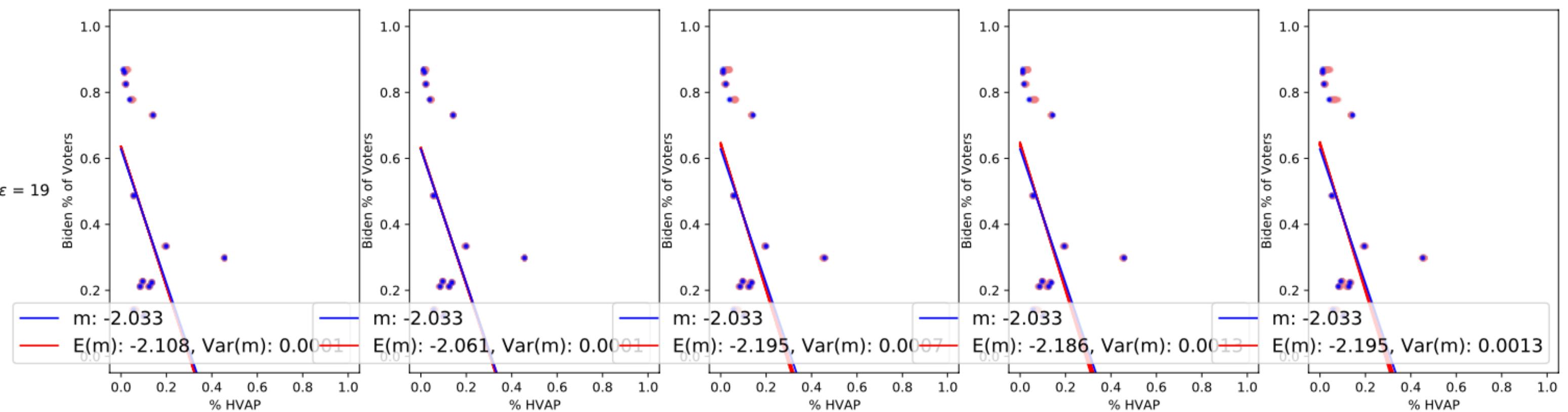
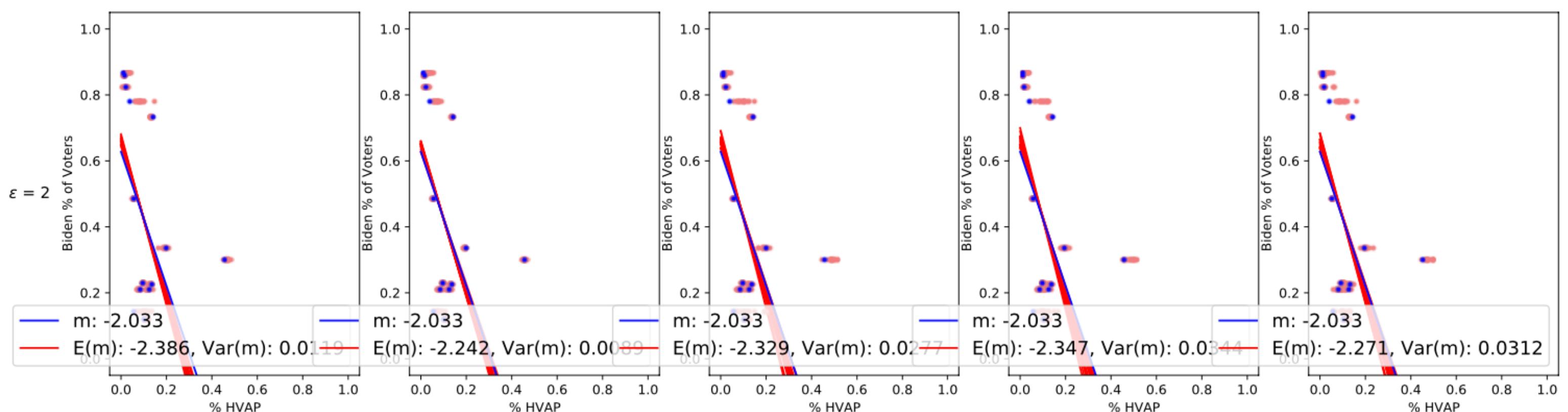
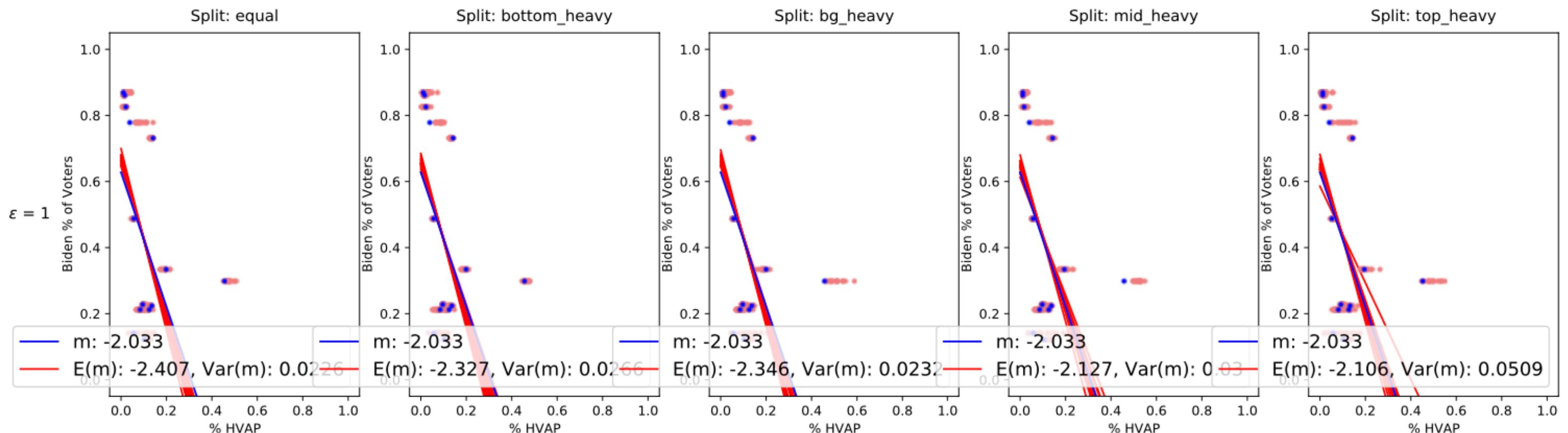
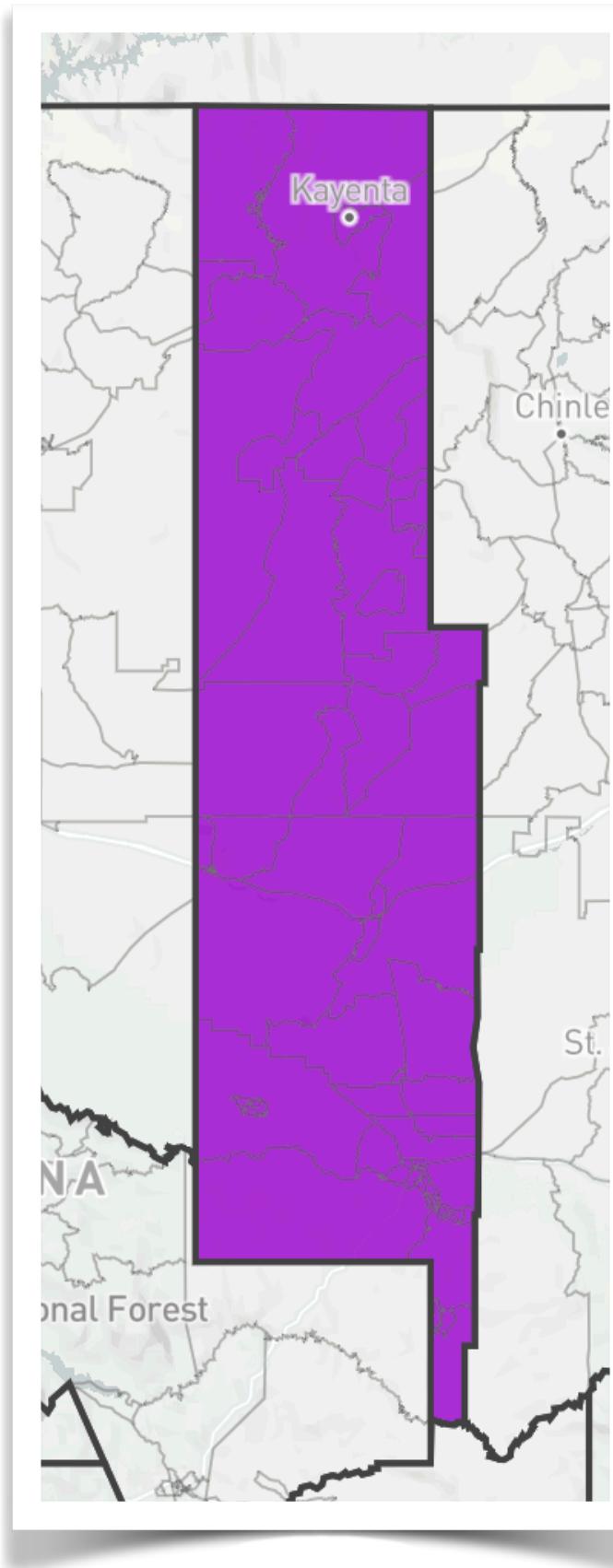
Navajo County

AMIN support for Biden



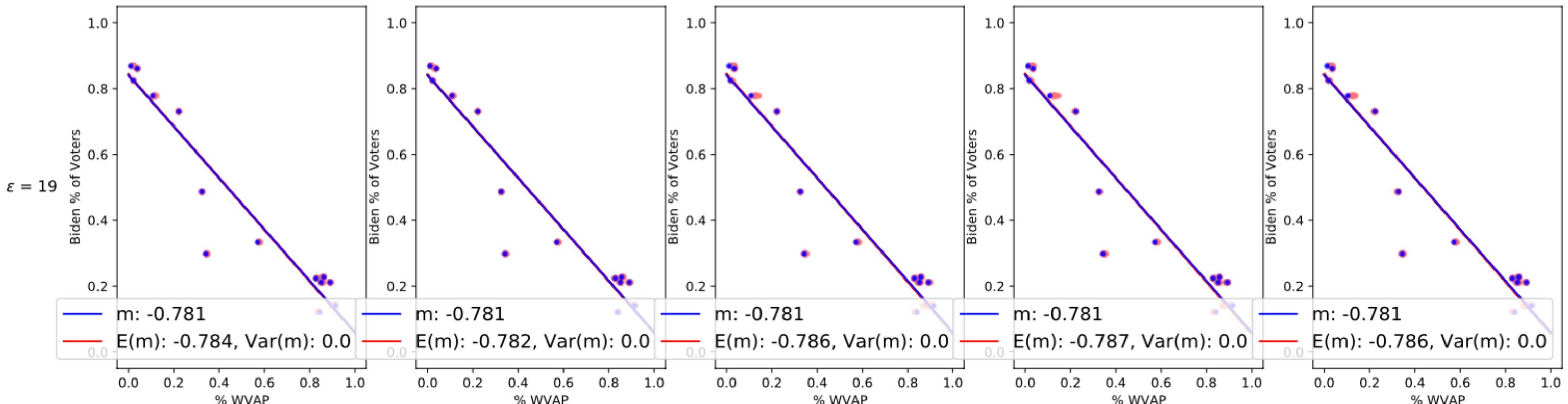
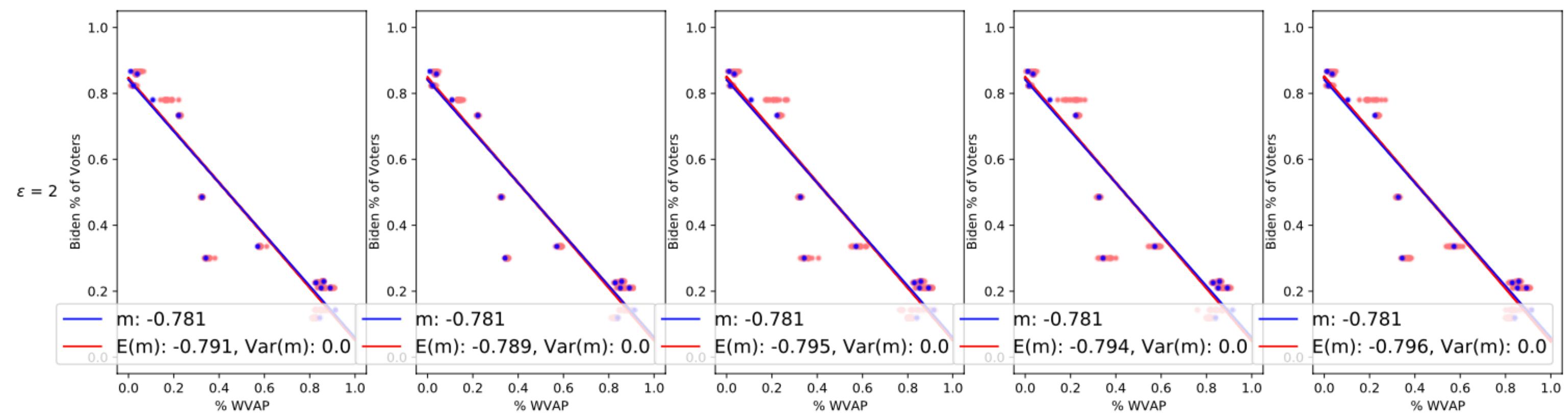
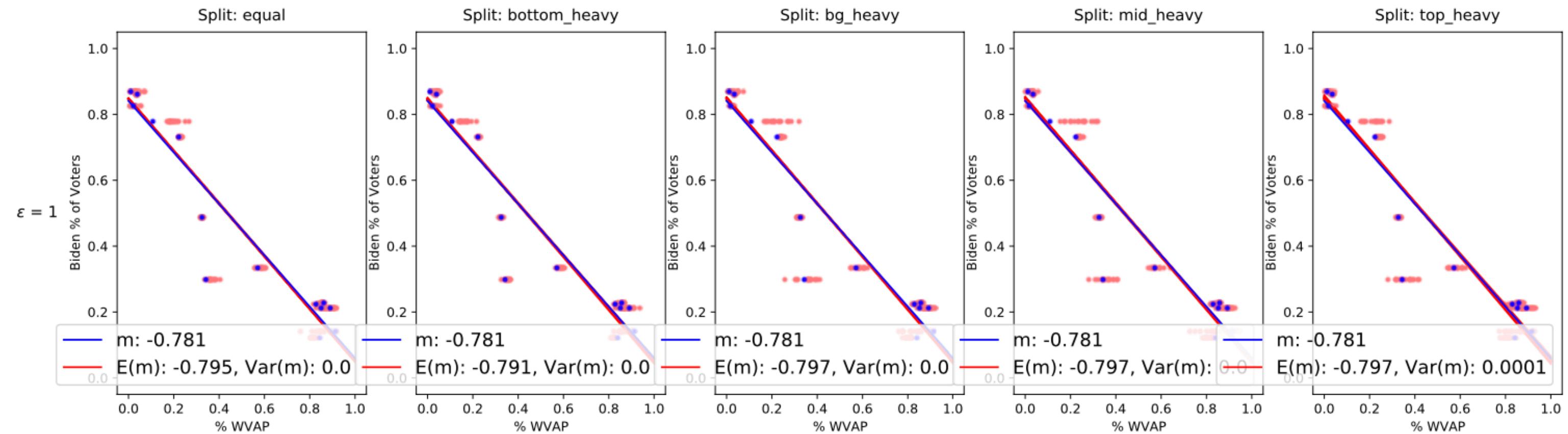
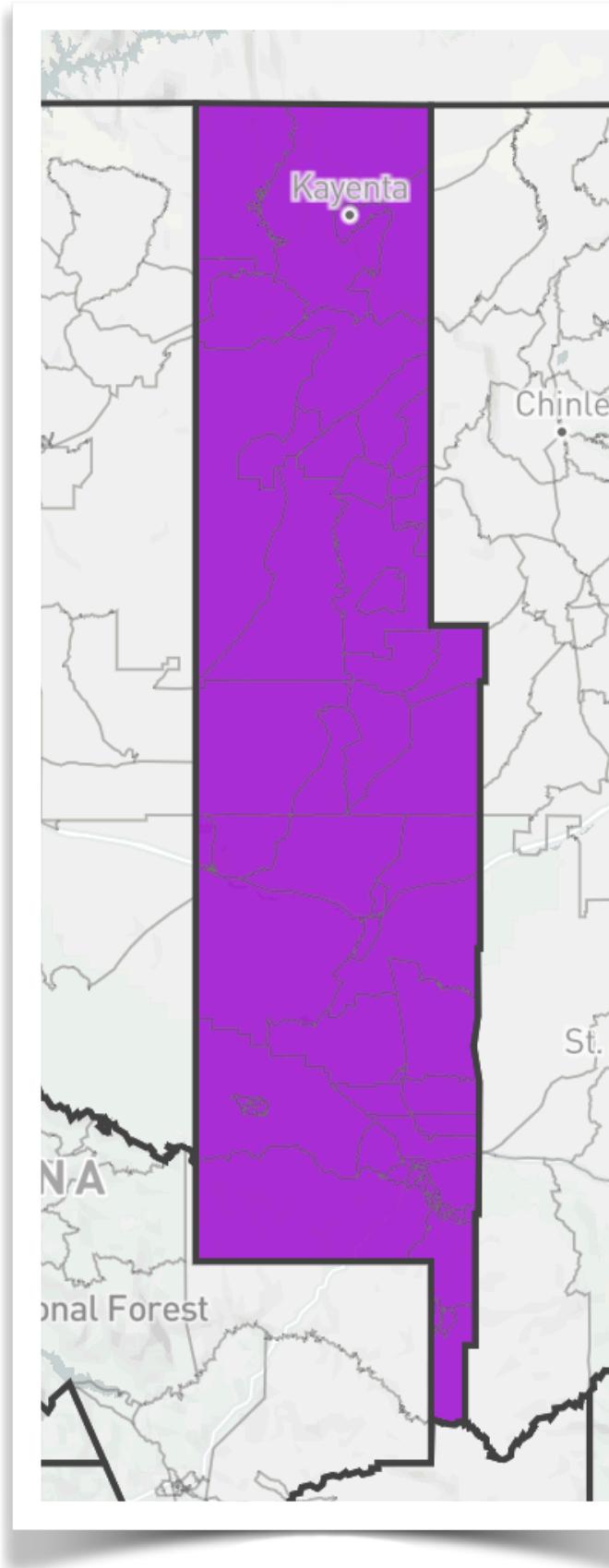
Navajo County

HISP support for Biden



Navajo County

W support for Biden



noised 16 times with
 $\varepsilon = 2$ and equal
allocation over the
geographical levels

PIMA	Hispanic for Biden	non-Hisp for Biden
un-noised	66.3%	57.2%
lowest of 16 noisy trials	65.3%	57.2%
highest of 16 noisy trials	66.3%	57.5%

noised 16 times with
 $\varepsilon = 2$ and equal
allocation over the
geographical levels

NAVAJO	AMIN for Biden	non-AMIN for Biden
un-noised	88.4%	17.0%
lowest of 16 noisy trials	88.7%	16.7%
highest of 16 noisy trials	89.2%	17.0%

How realistic are these experiments?

We studied DP for a year using Census code from July 2019

Since then, Bureau has announced many details/changes, some in response to end-user pushback

- **TopDown** instead of **ToyDown** – *more accurate overall*
- Gaussian vs Laplace noise – *noise has thinner “tails”*
- “Optimized block groups” – *will fit cities/towns better*
- Tuned workload and invariants – *leverages household, other structure*

All of these make discrepancies substantially **smaller!**

Takehome messages

The privacy risks are real

The previous disclosure avoidance methods (e.g., “swapping”) are opaque, ad hoc, and underpowered

For each geography we considered, the Census data will clearly be completely adequate for every redistricting application we studied

We find no threat to VRA enforcement or to reasonable population balance

Our study suggests some updated best practices for redistricting

- Build from bigger units
- Weight your regressions
- Time to break zero-balance habit?



thanks!

mduchin@mggg.org