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")

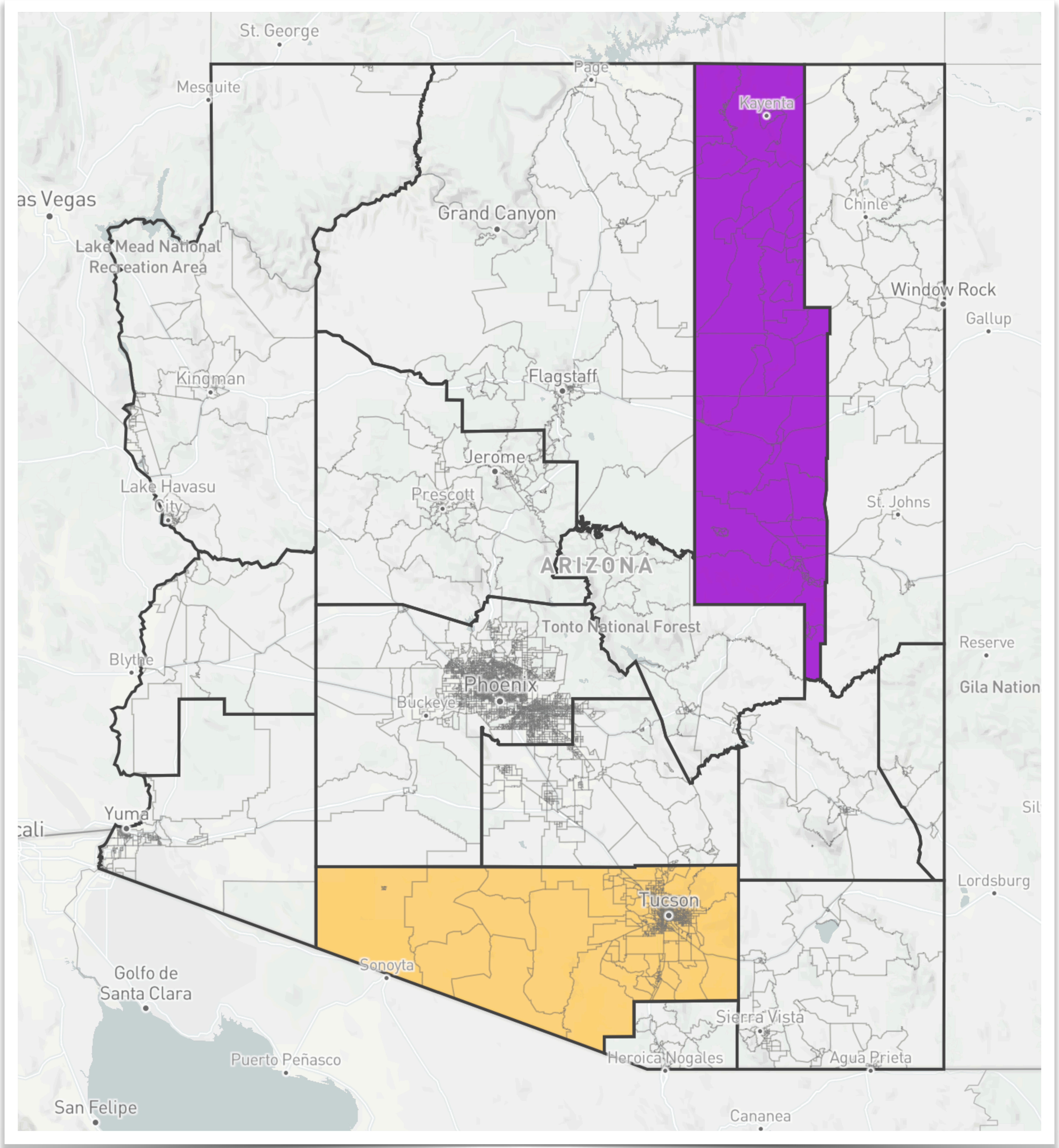
Differential privacy study funded by **Alfred P. Sloan Foundation –** joint work with Aloni Cohen, JN Matthews, and Bhushan Suwal, in collaboration with Mark Hansen, Denis Kazakov, and Peter Wayner

Large districts
(U.S. Congress)

7,151,502/9 ≈ 794,611

Small districts
(Navajo County commission)

107,449/5 ≈ 21,490



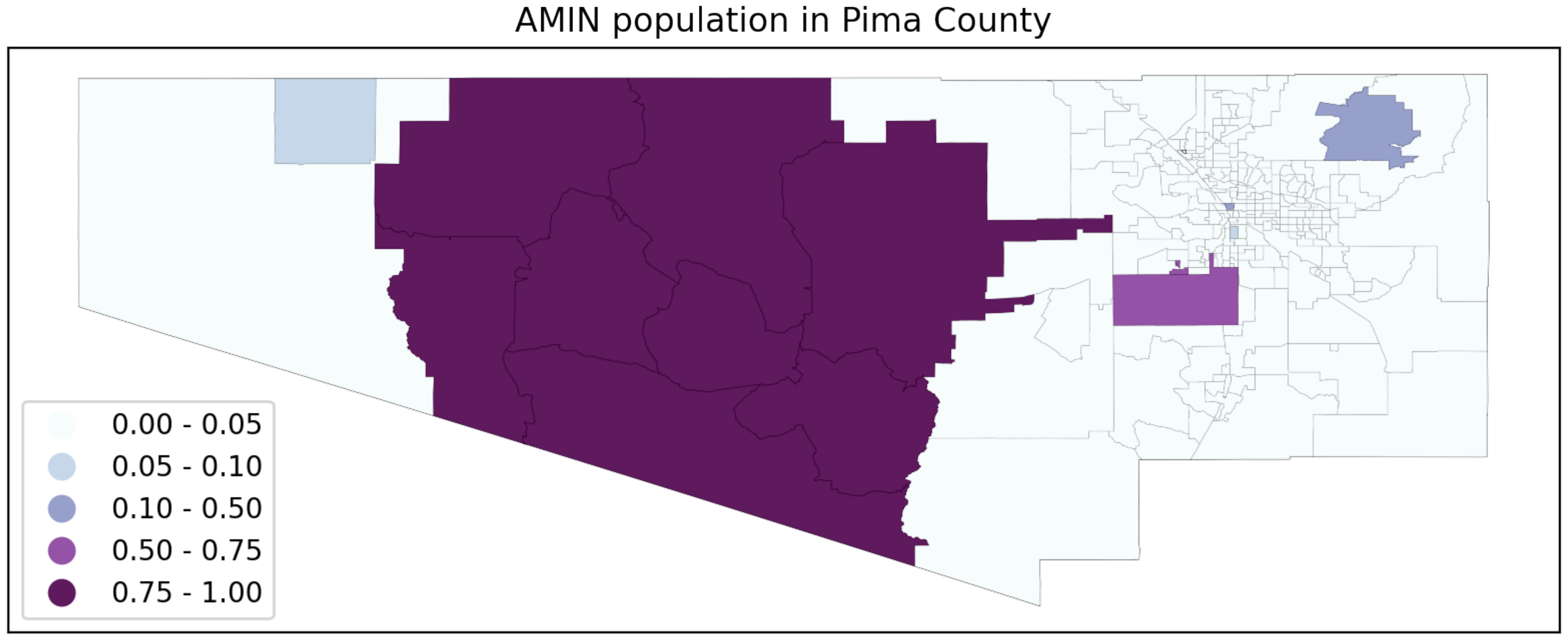
Navajo County,
pop. 107,449

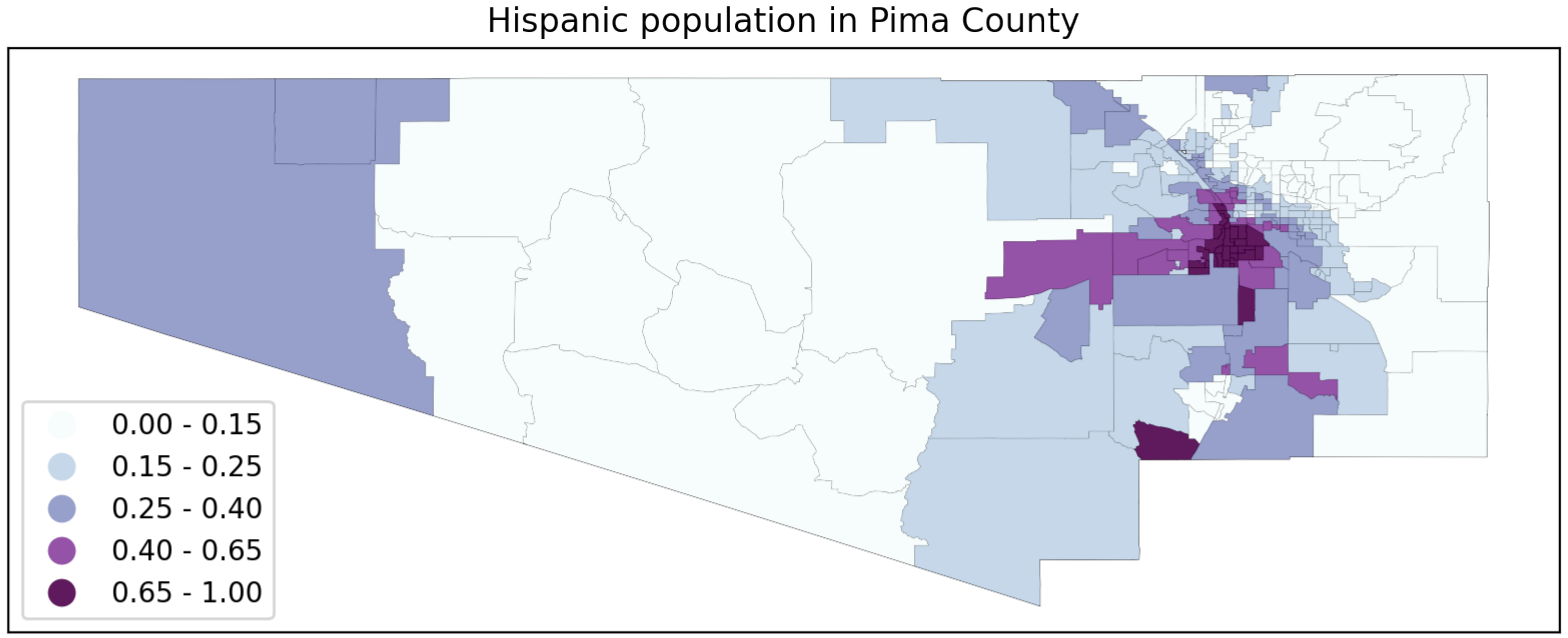
44%W, 11%H,
42%AMIN

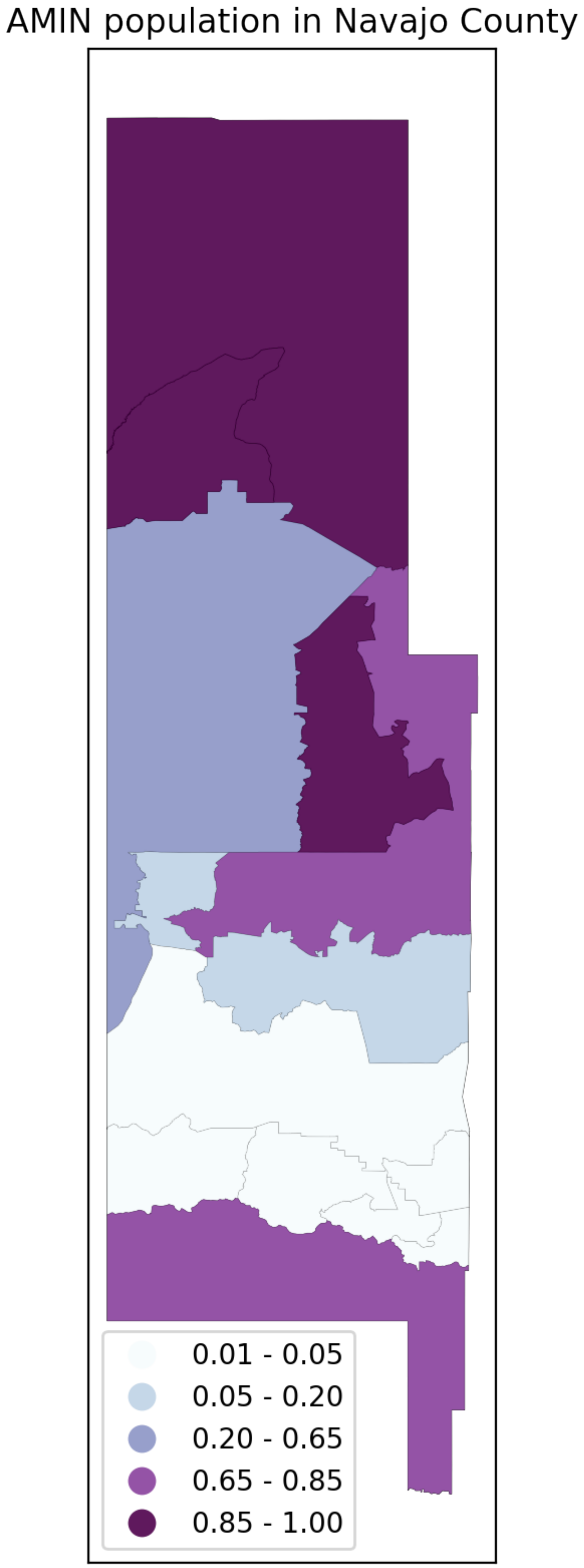
Pima County,
pop. 980,263

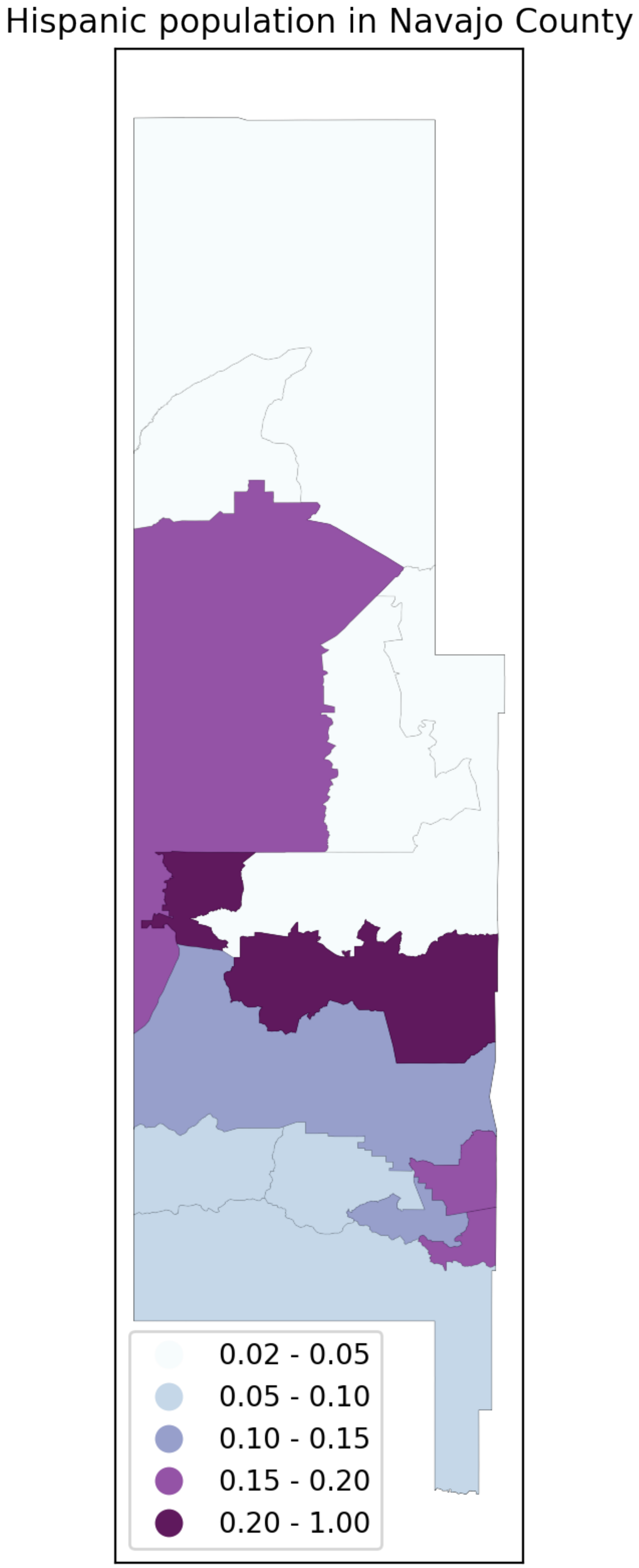
55%W, 35%H,
2.5%AMIN

Both counties have significant diversity









What is the risk?

© User:Colin / Wikimedia Commons

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

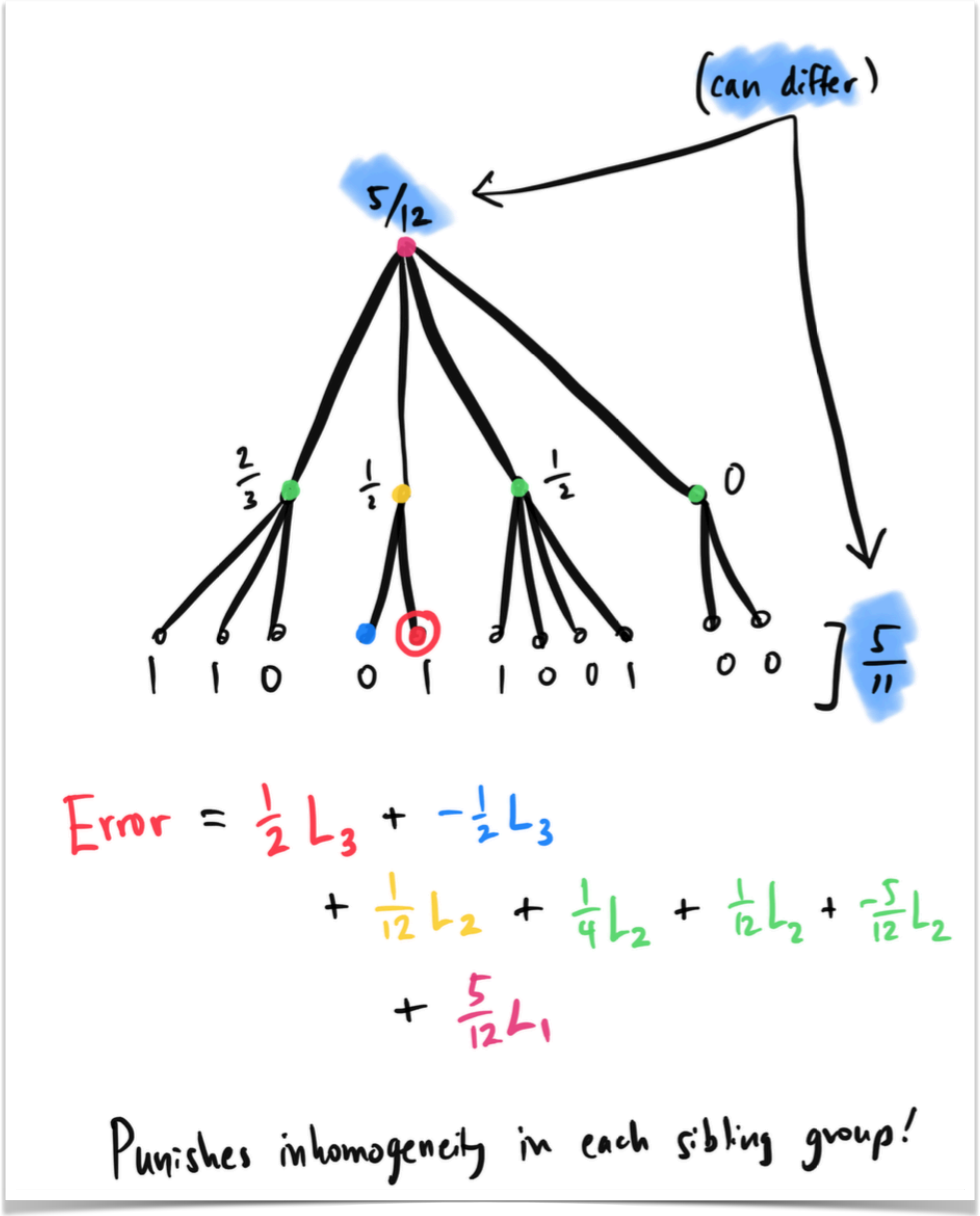
can get whole state in a few days

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

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

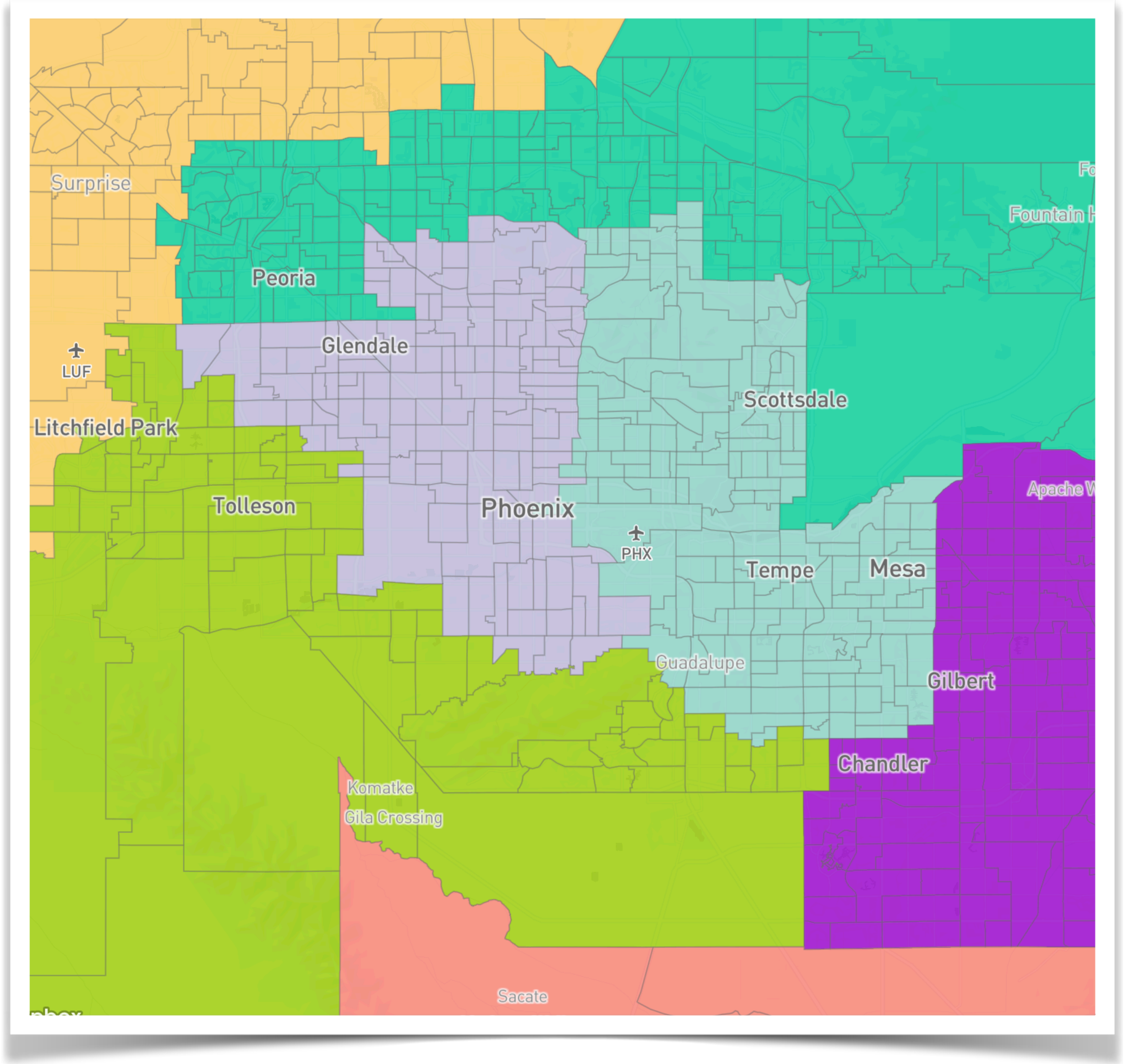
pairs with easily obtained commercial data to get full **reidentification**

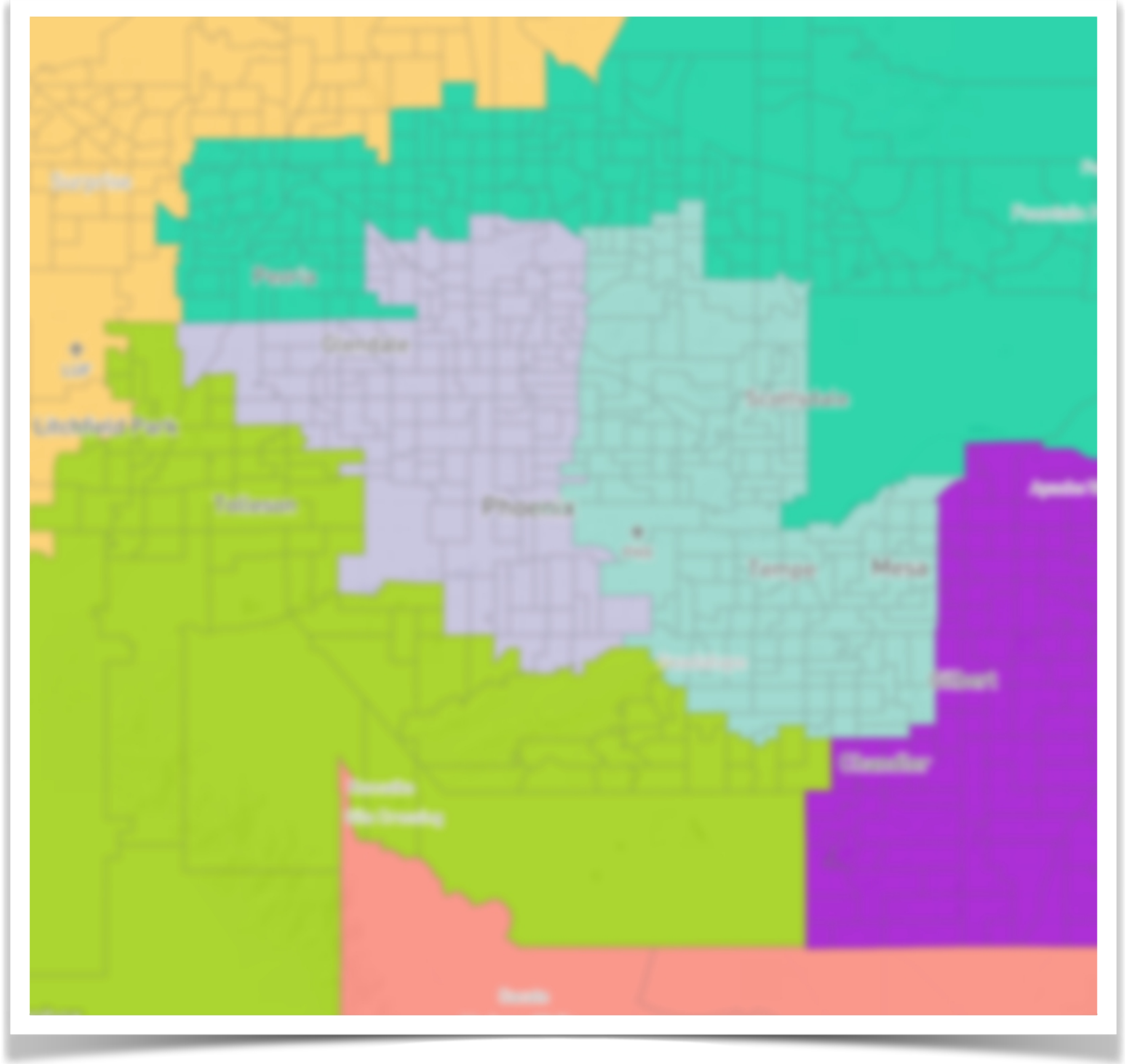
What is differential privacy?

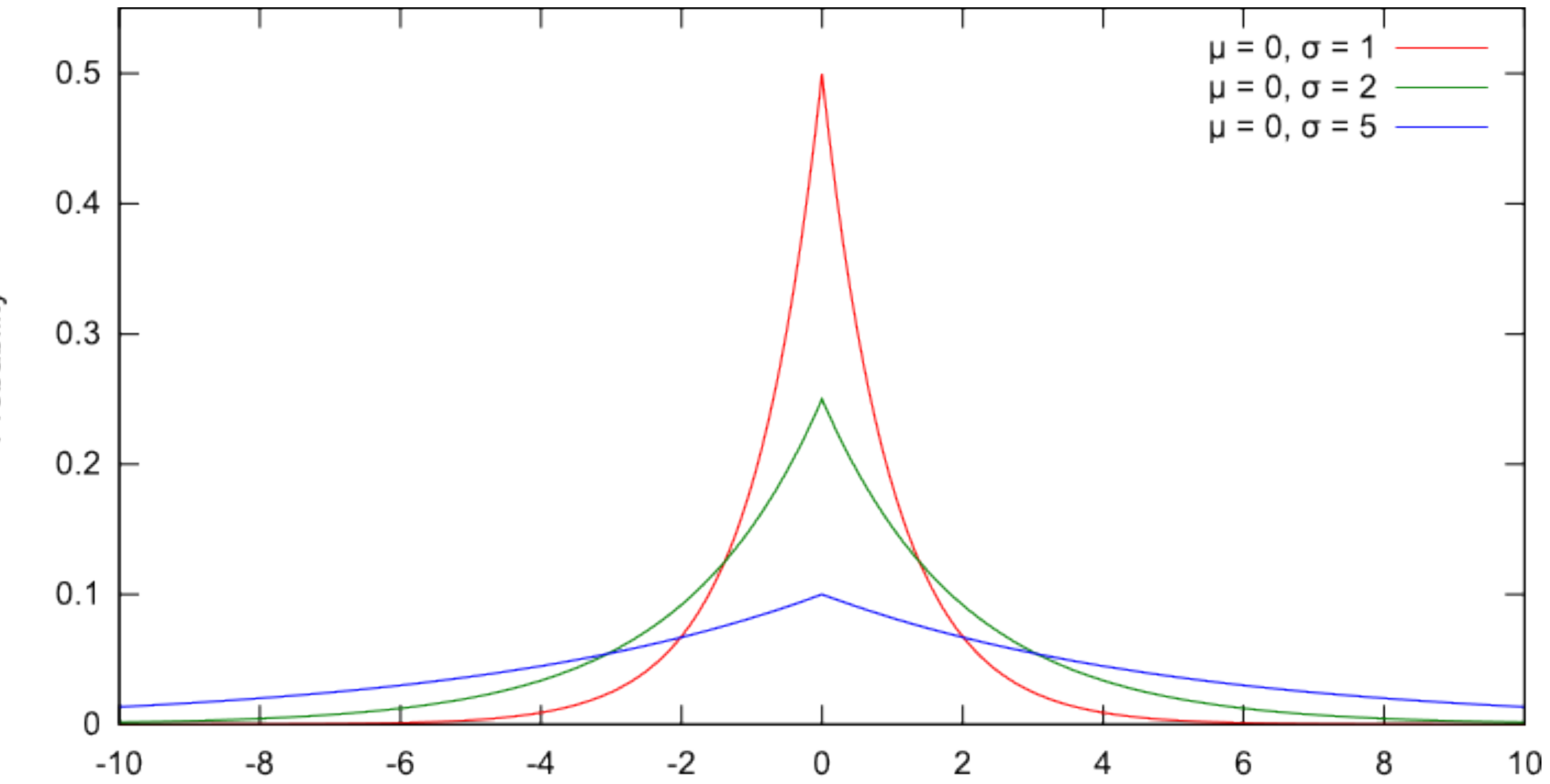


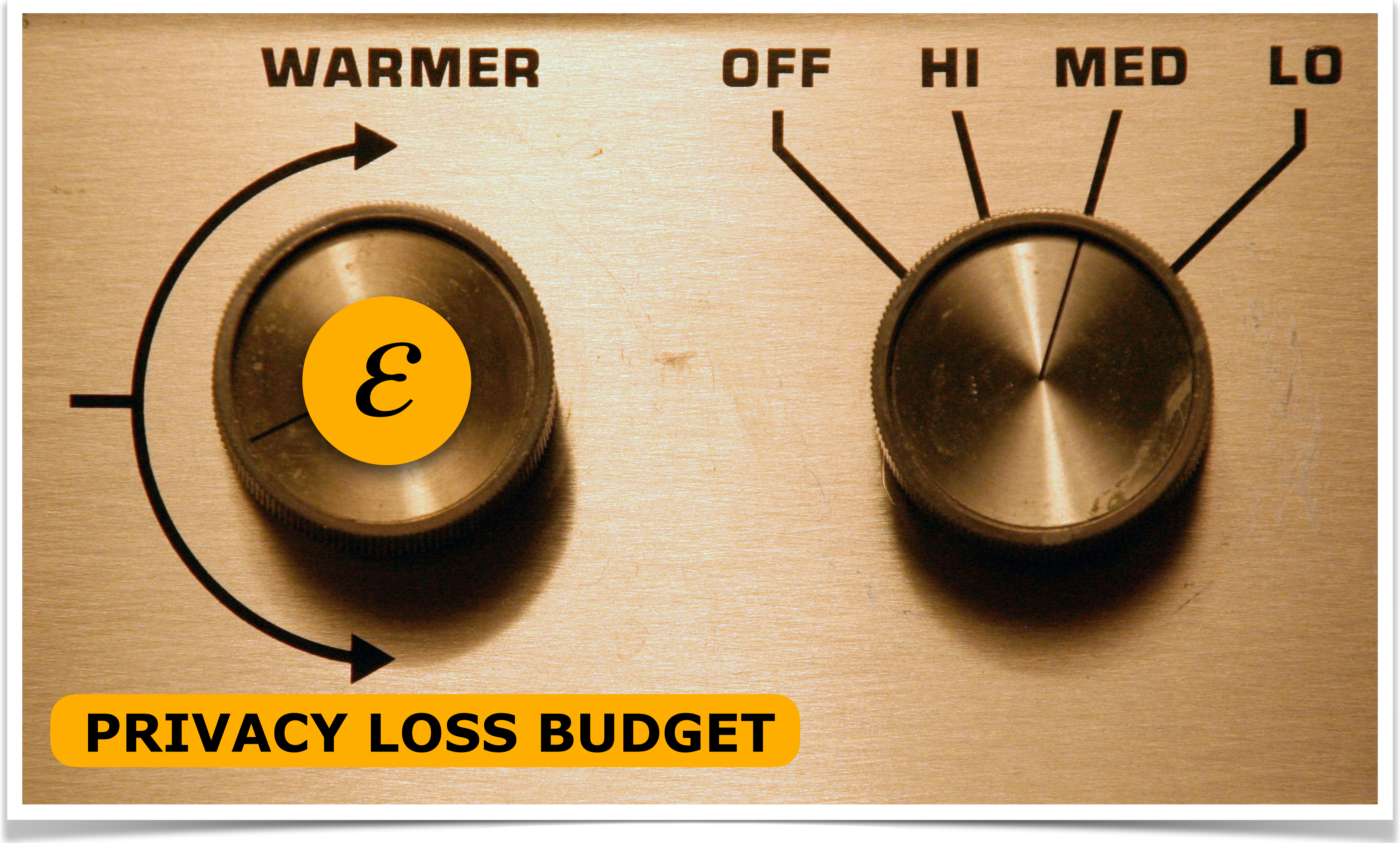
Idea: for **privacy**, add **noise**

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)

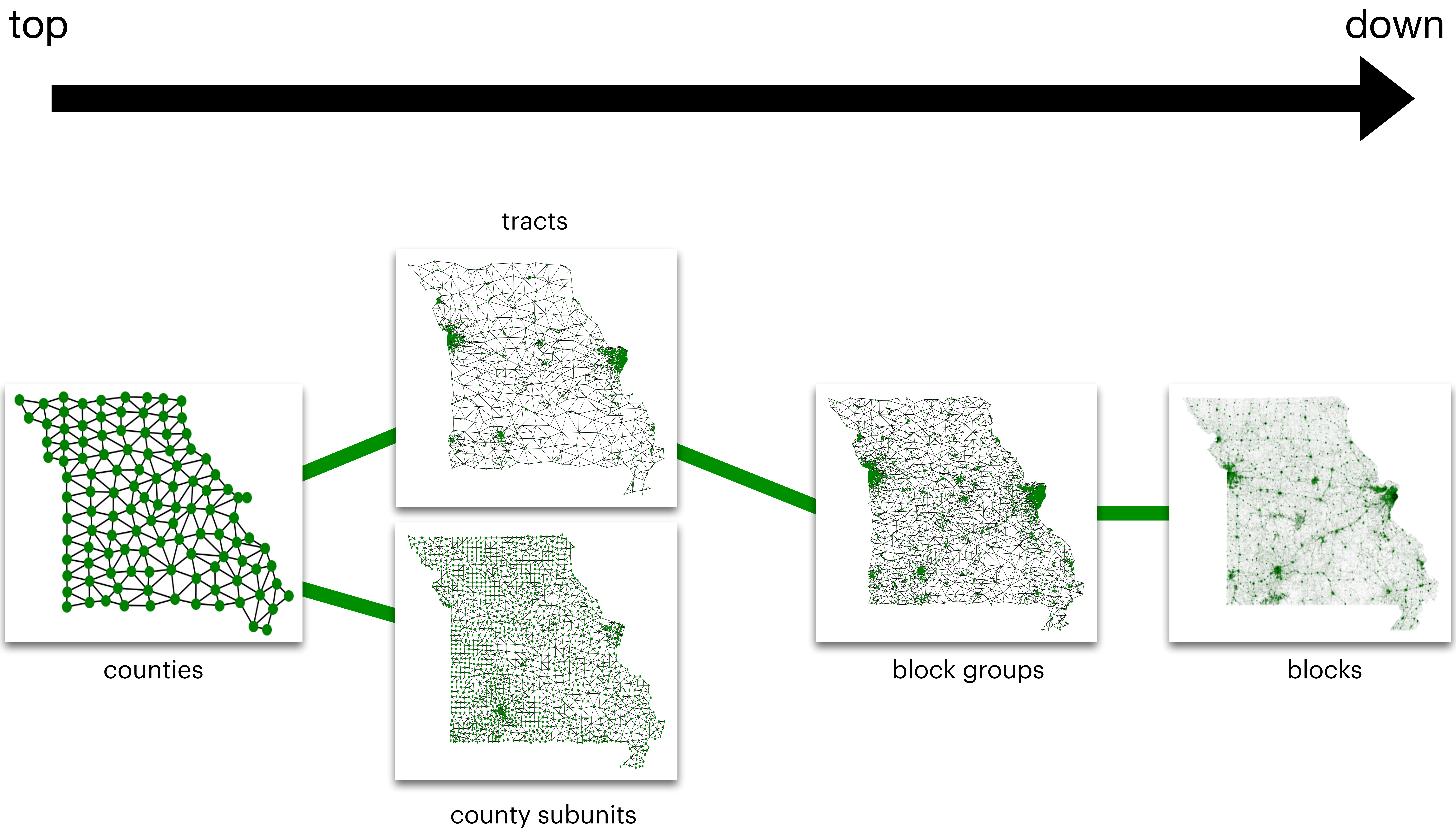
"differential privacy" essentially means that you have control over the knobs – can **calibrate** the tradeoff between privacy and accuracy

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

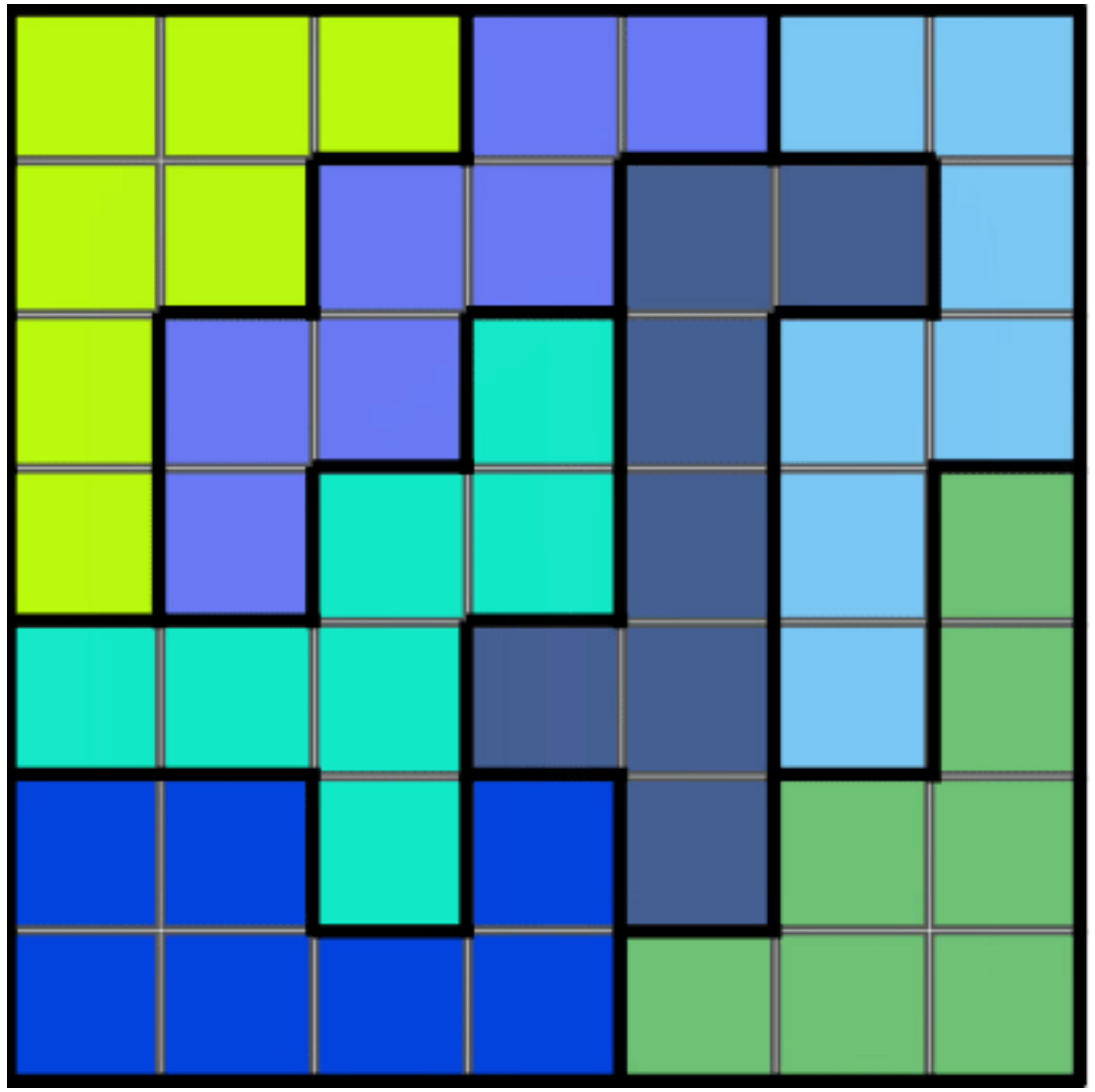

top
down
tracts
counties
county subunits
block groups
blocks

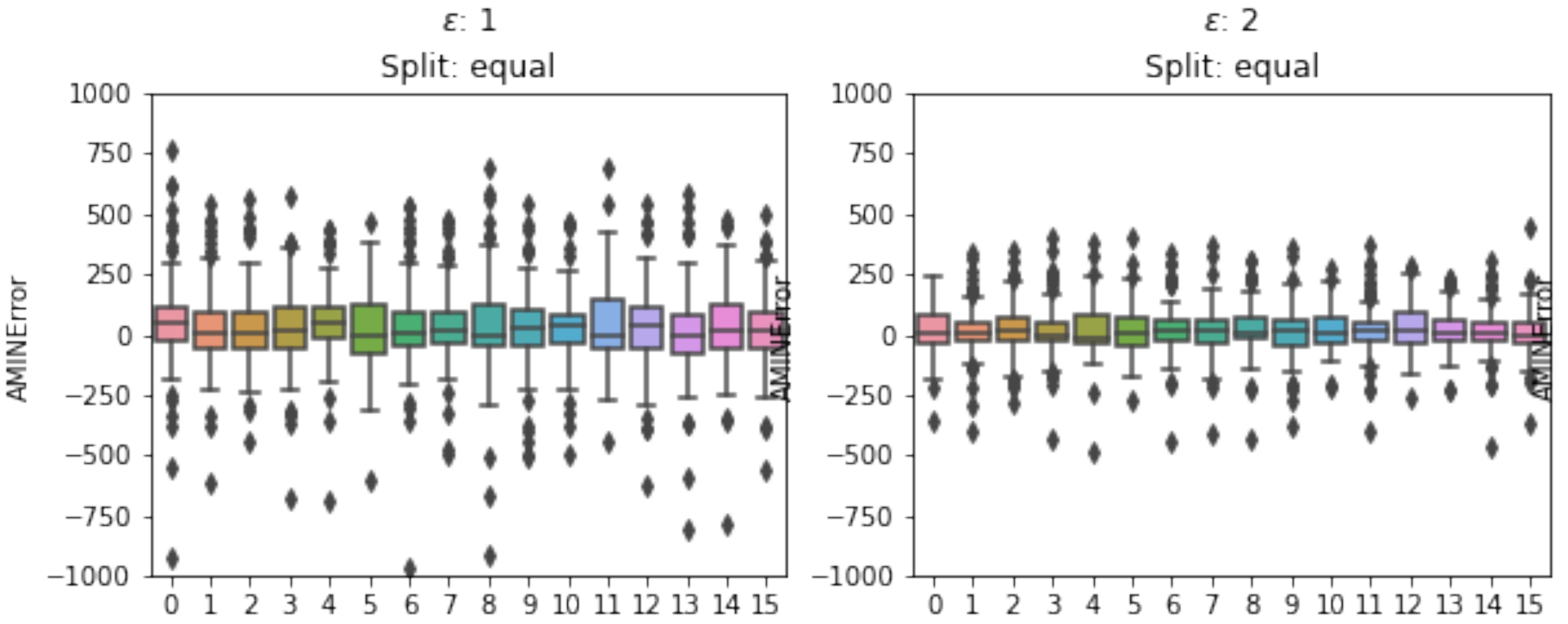
let's see some

experiments

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

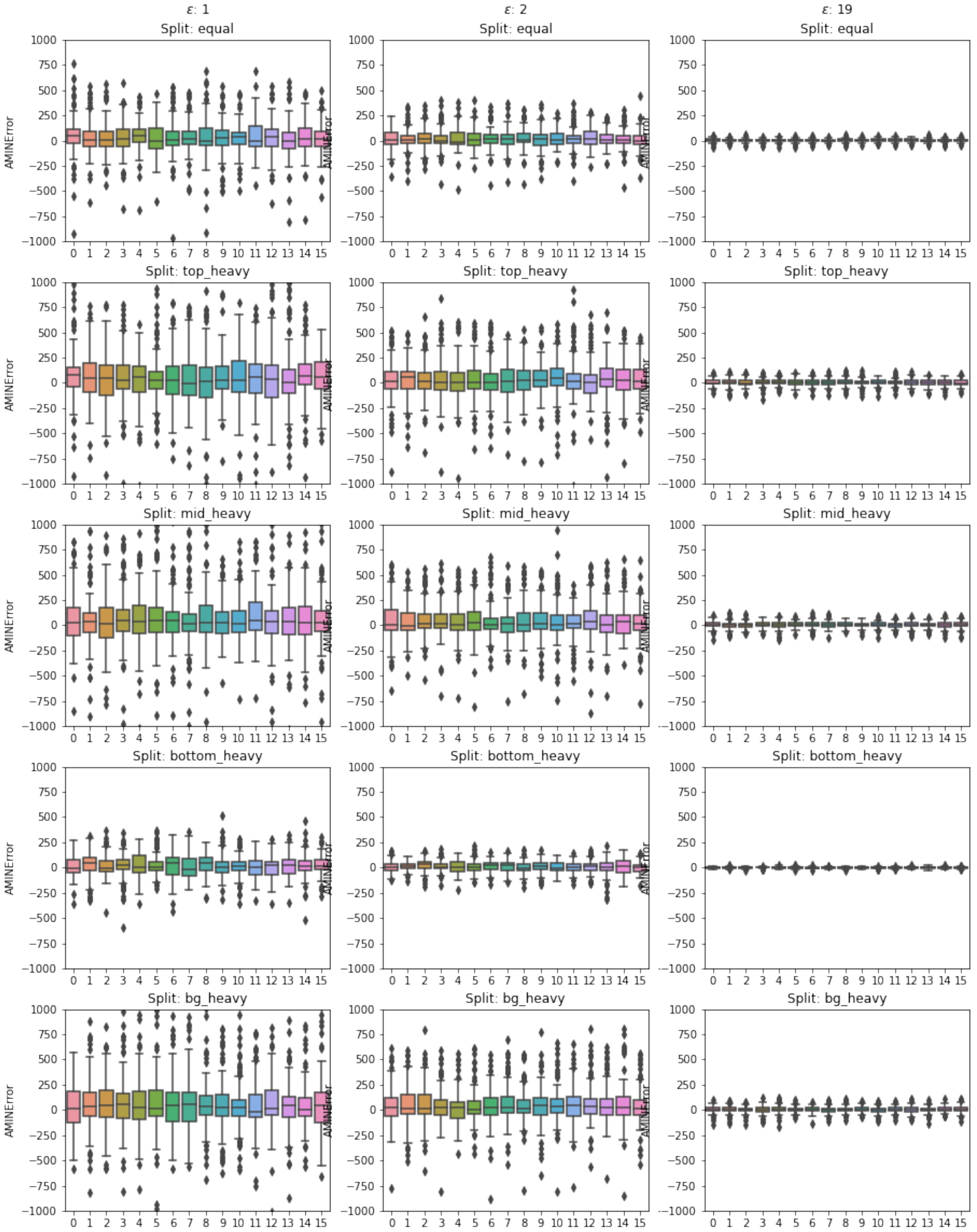
Do districts lose Native population?





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

…truly tiny at $\varepsilon = 19$



$\varepsilon = 1, 2, 19$

Navajo County

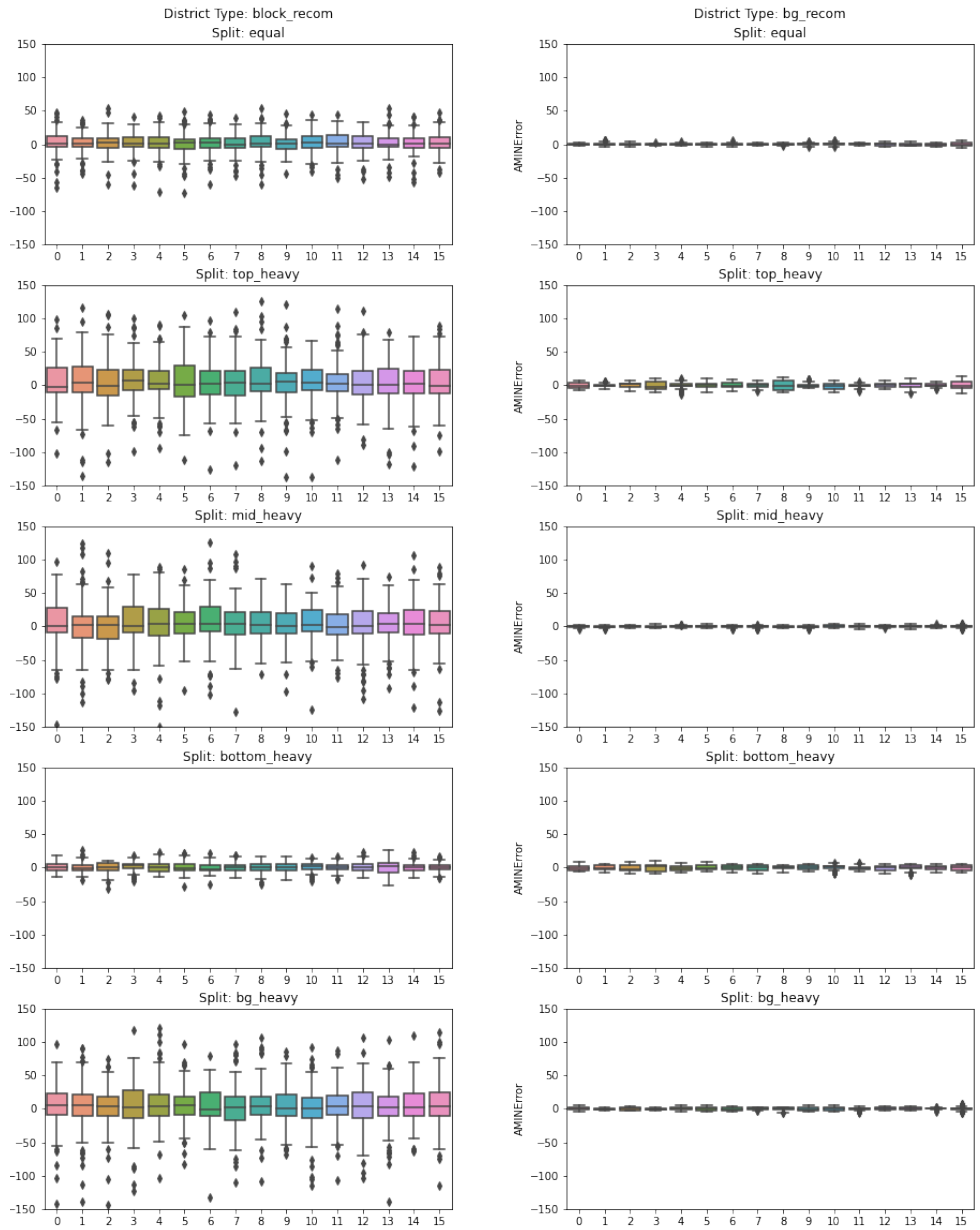
k=5 districts, population 20K

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 $\varepsilon = 1$, the typical discrepancy is under 500

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



Navajo County

built from blocks vs. block groups

k=5 districts, population 20K

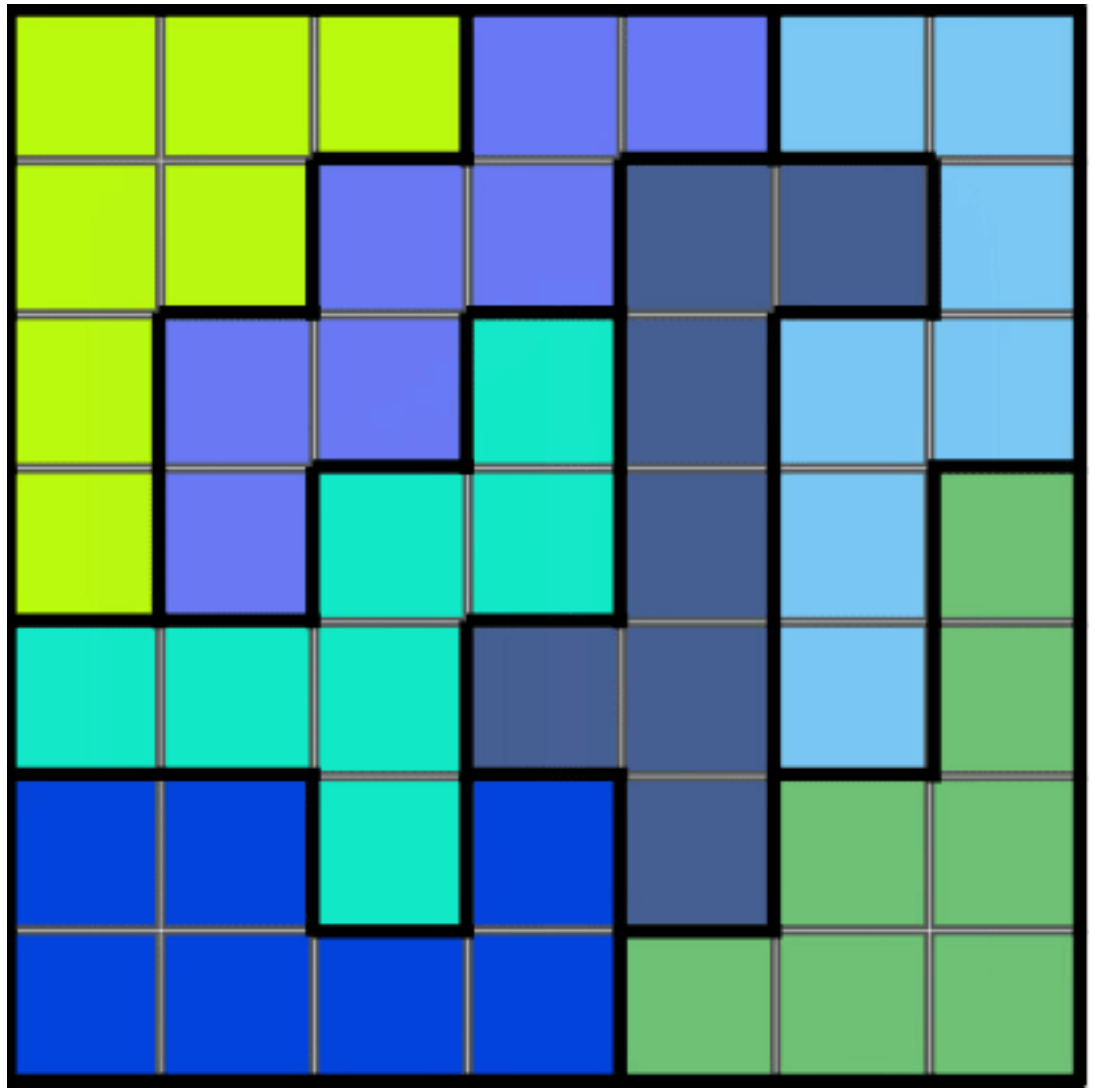
these plots show the discrepancy introduced by top-down style differential privacy

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

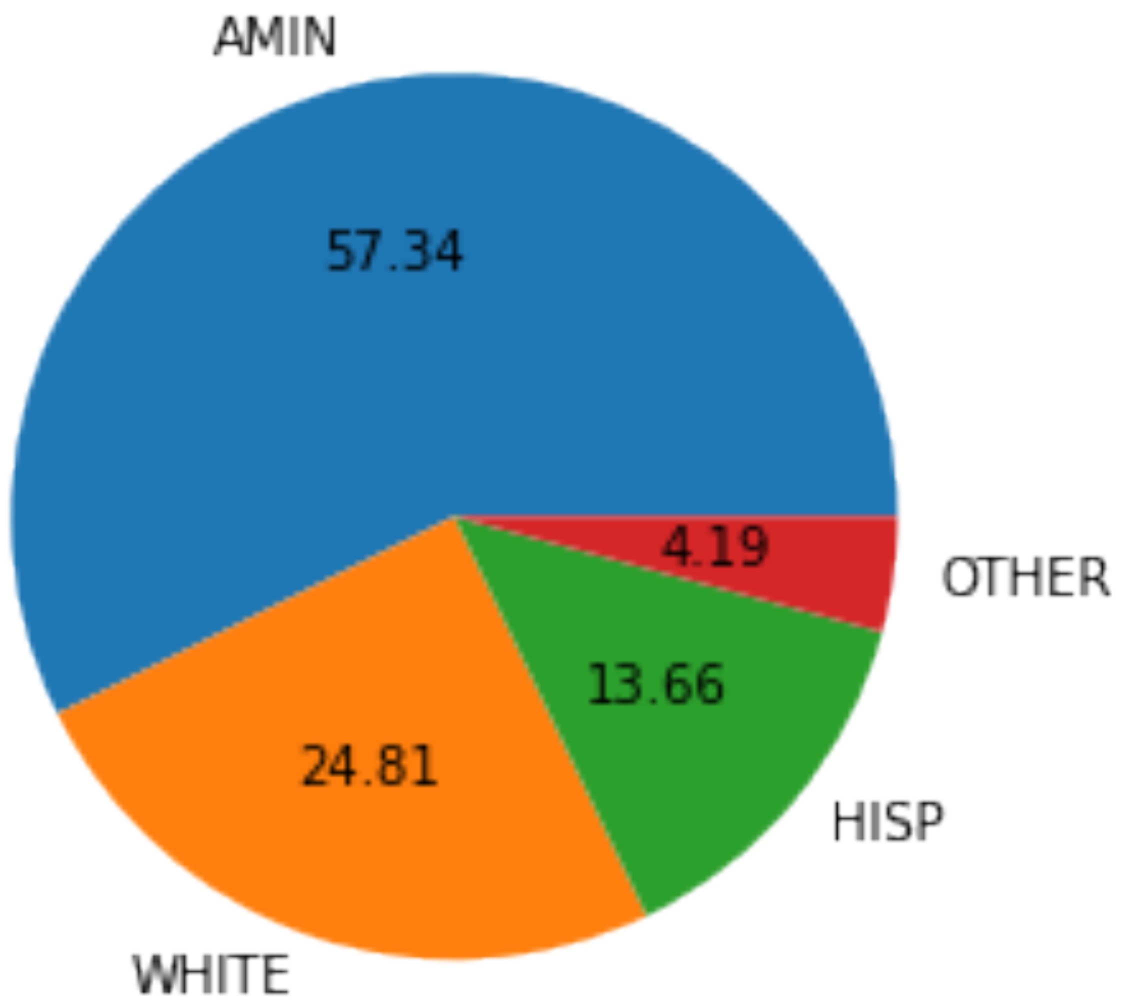
construction matters!

far better accuracy on districts built from larger pieces

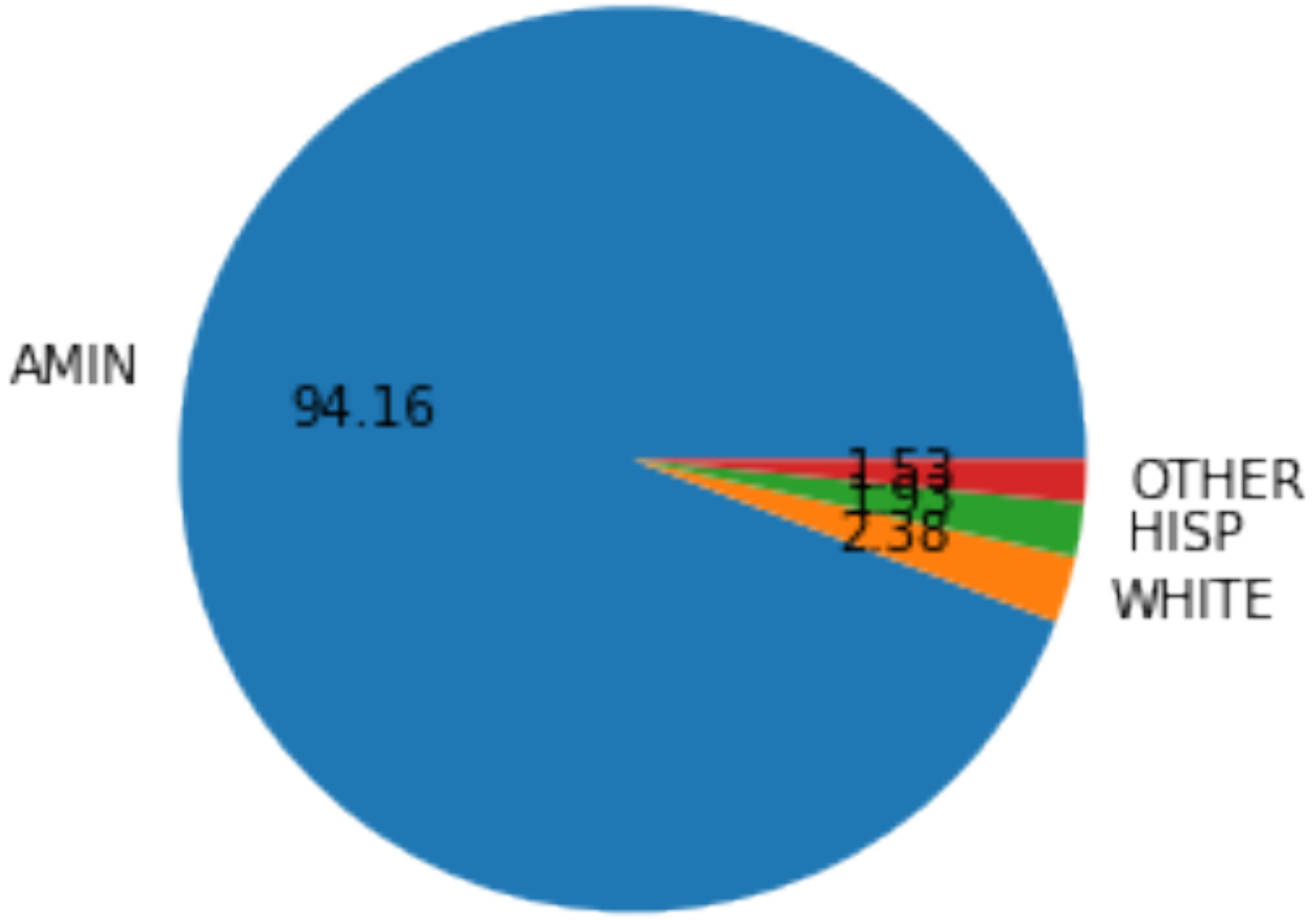
Do districts change their overall racial composition?



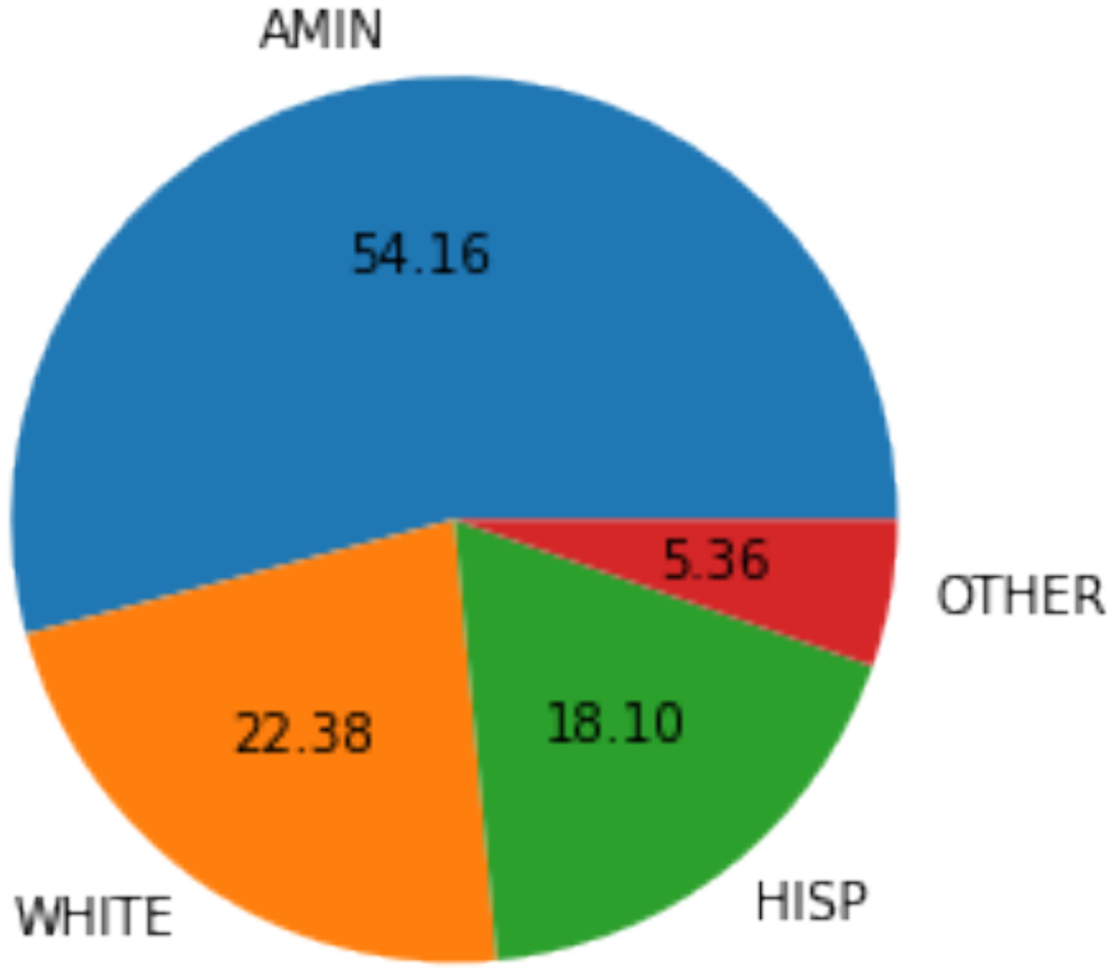
we will noise these 16 times with $\varepsilon = 2$ and equal allocation over the geographical levels



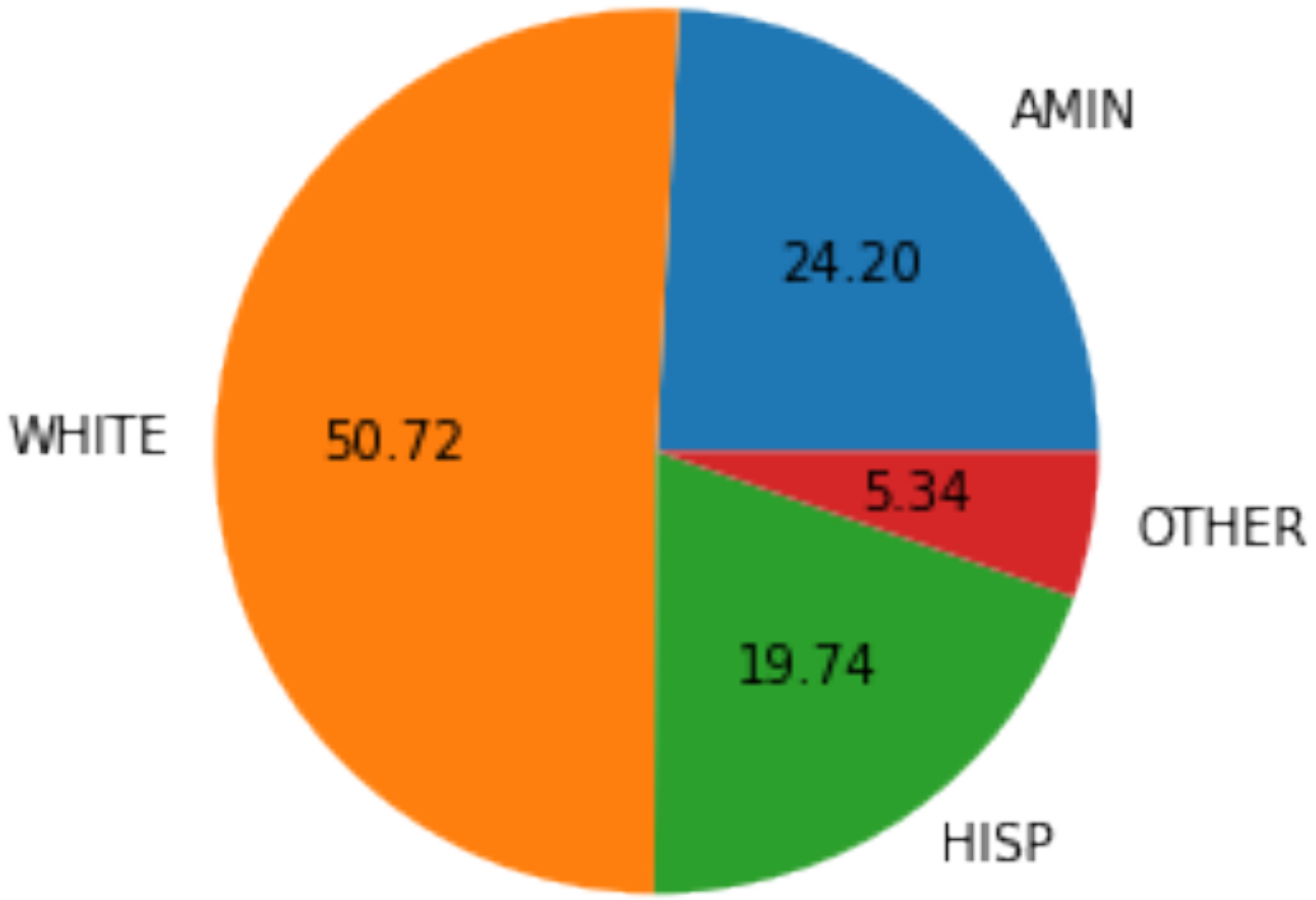
random district #2



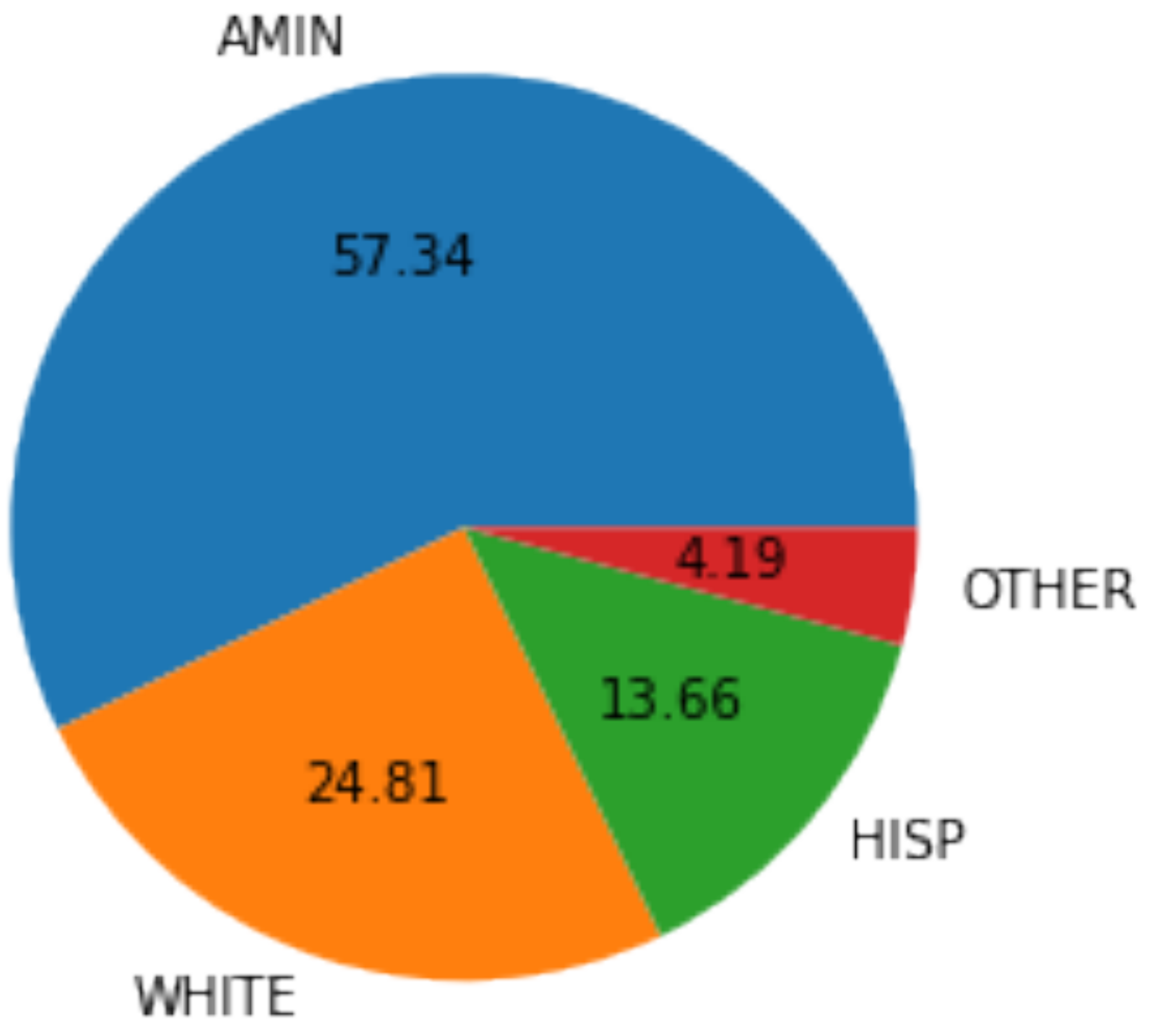
random district #9



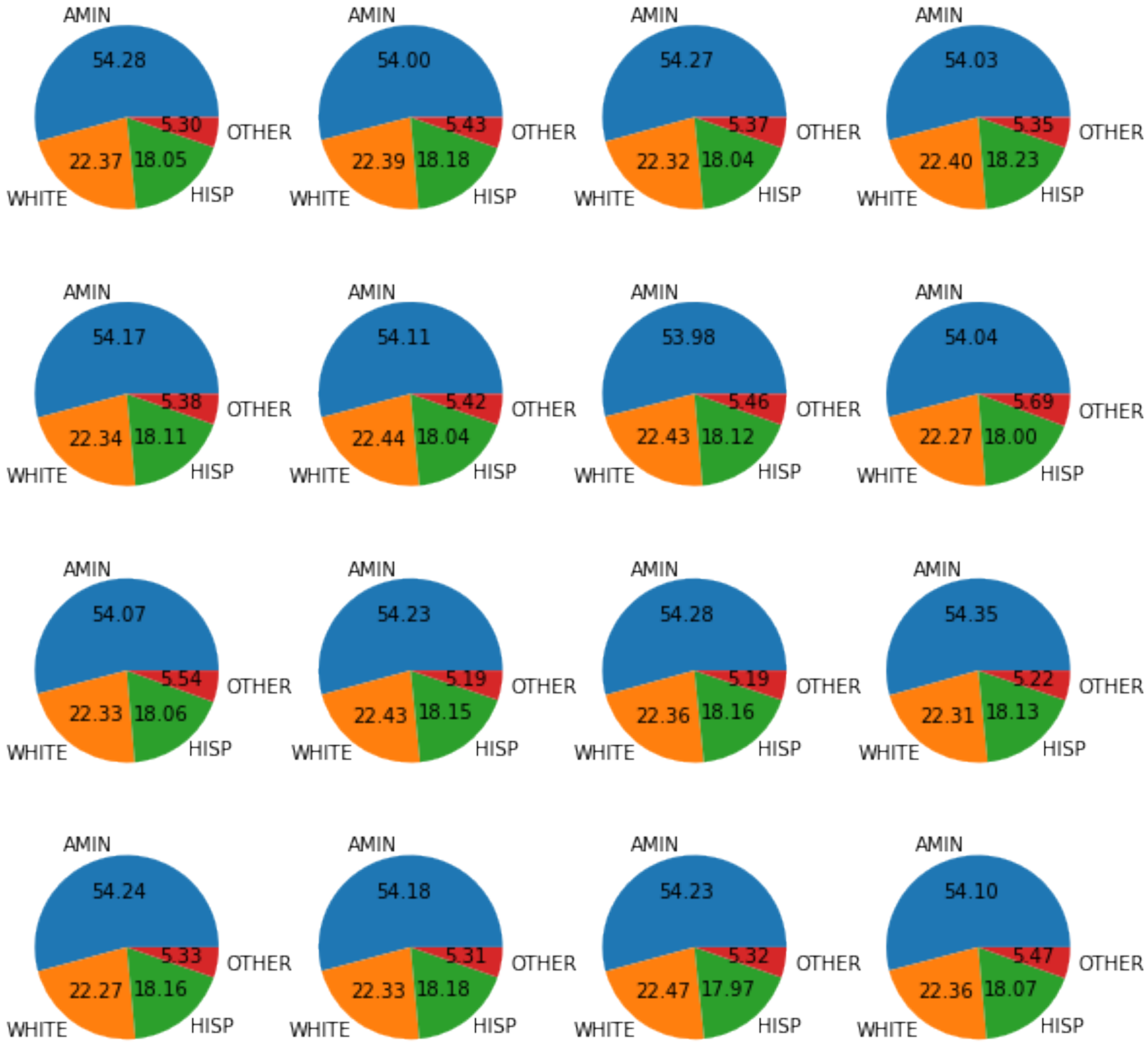
random district #13

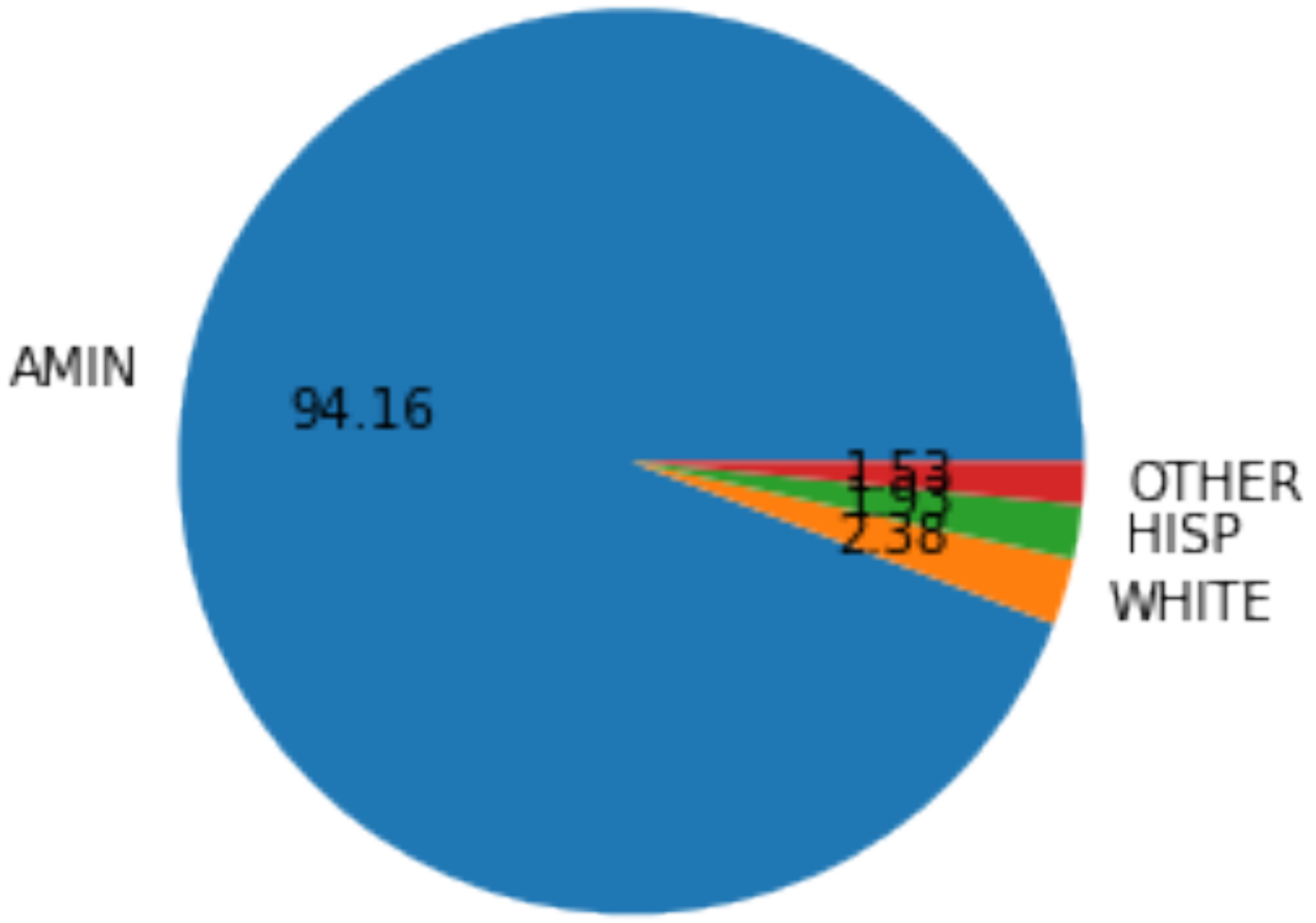


random district #46

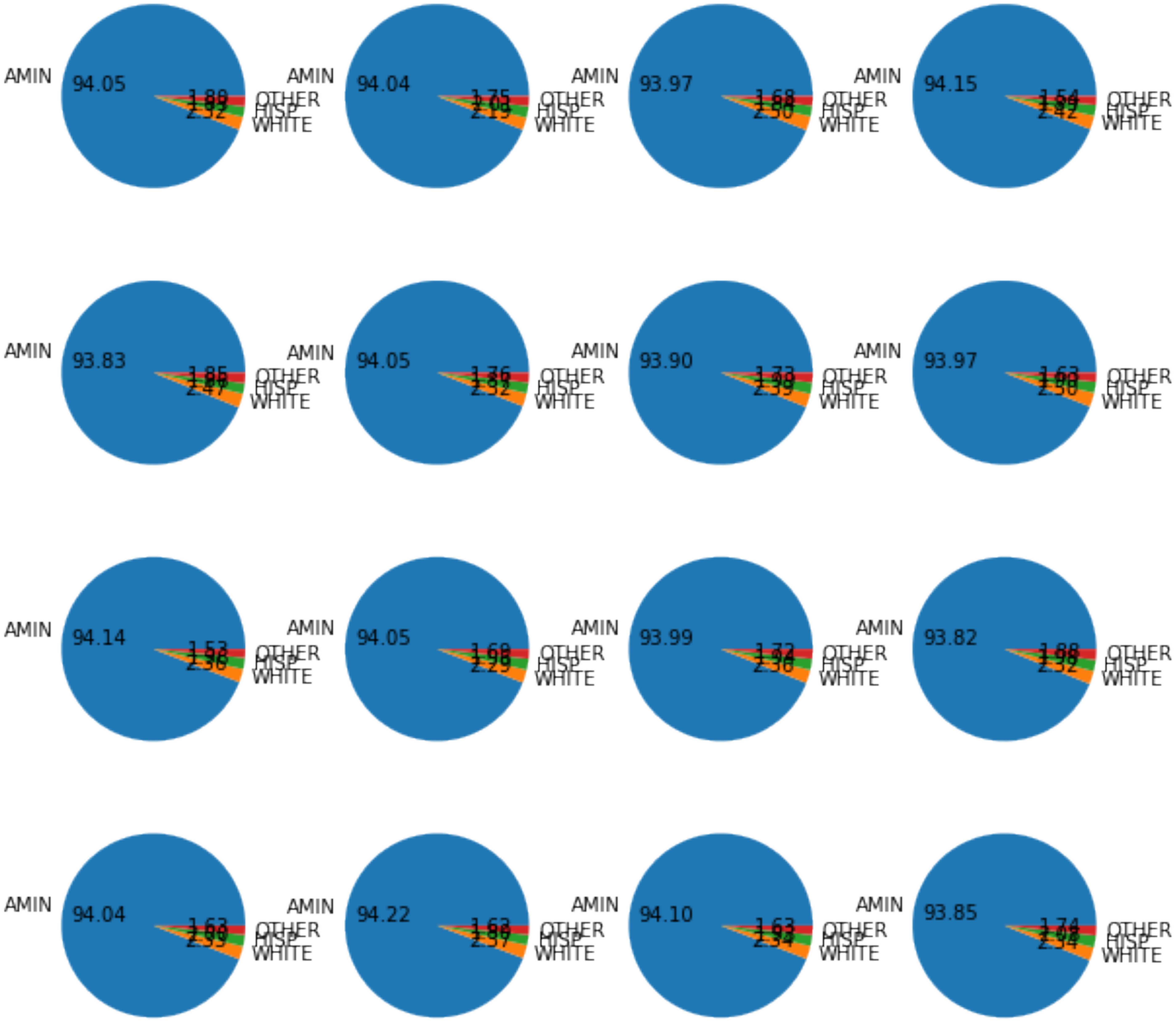


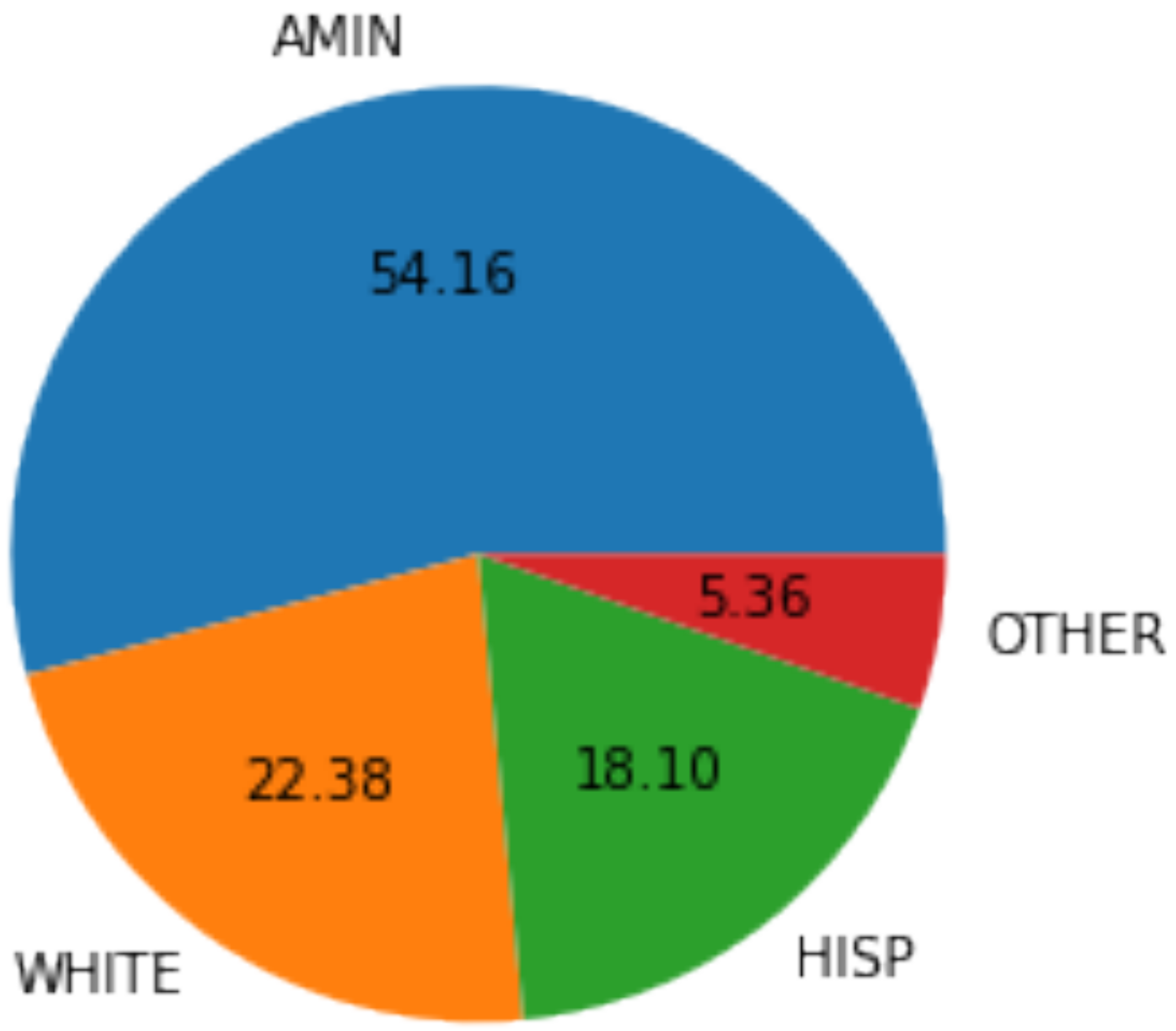
random district #2



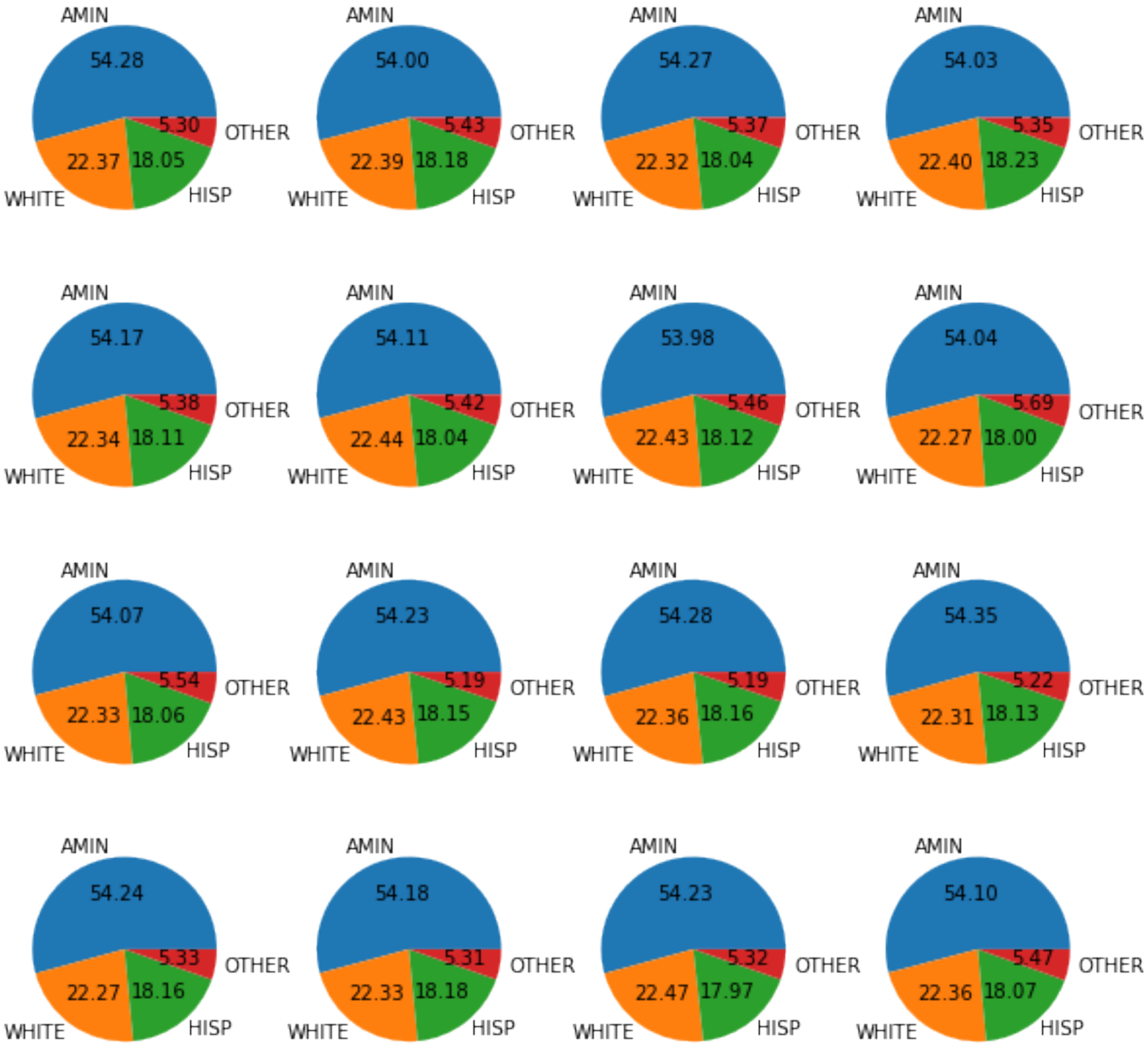


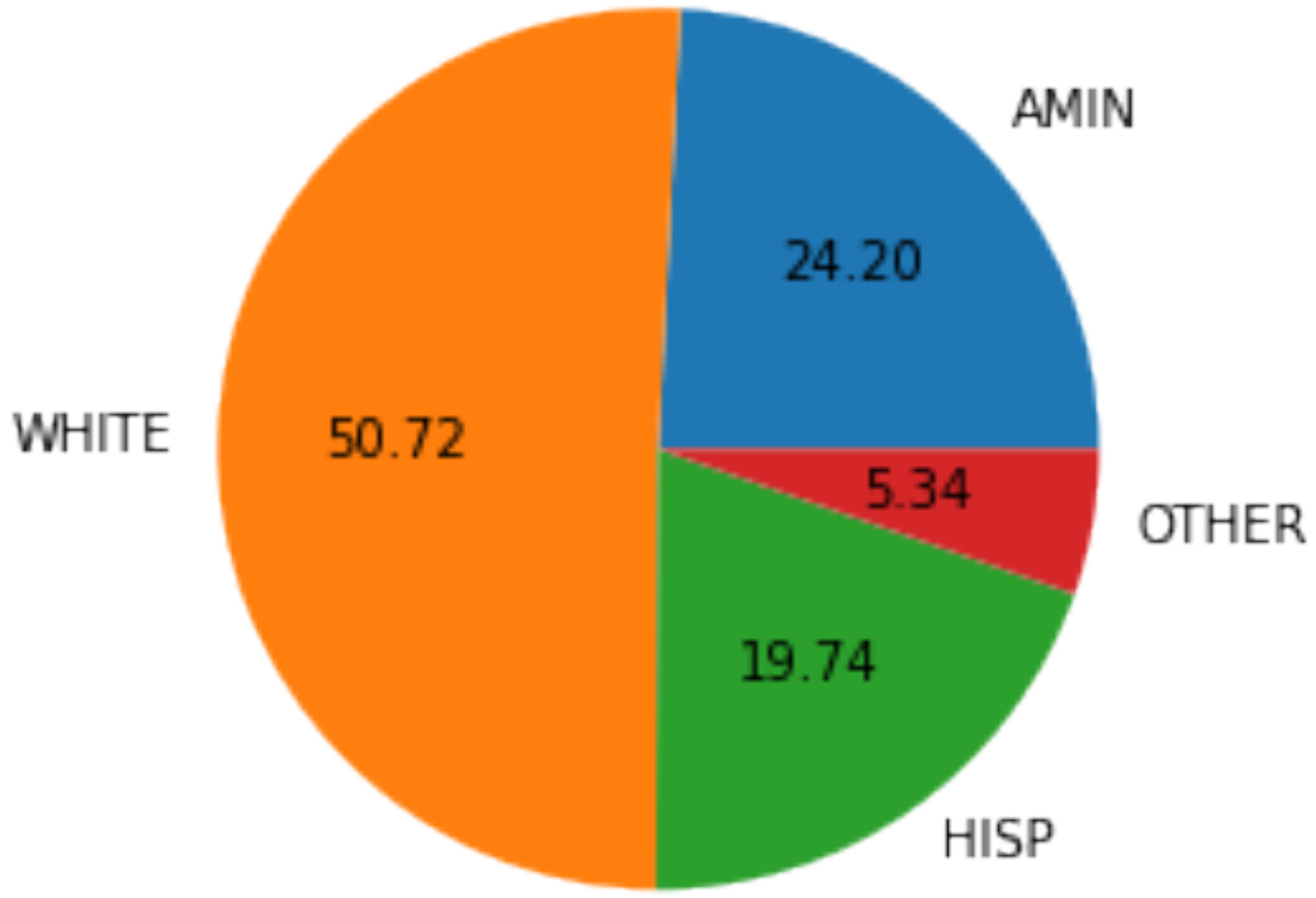
random district #9



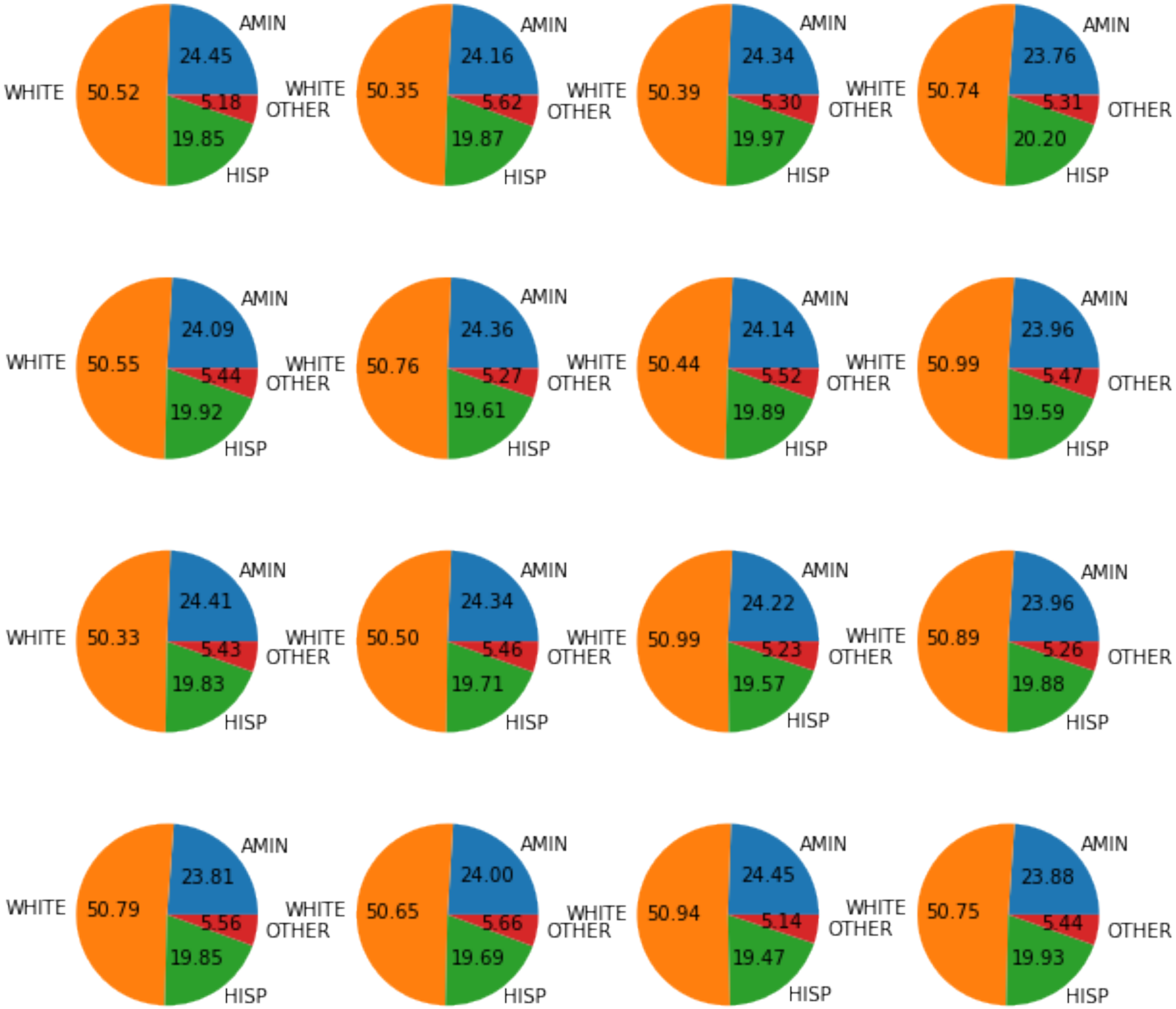


random district #13

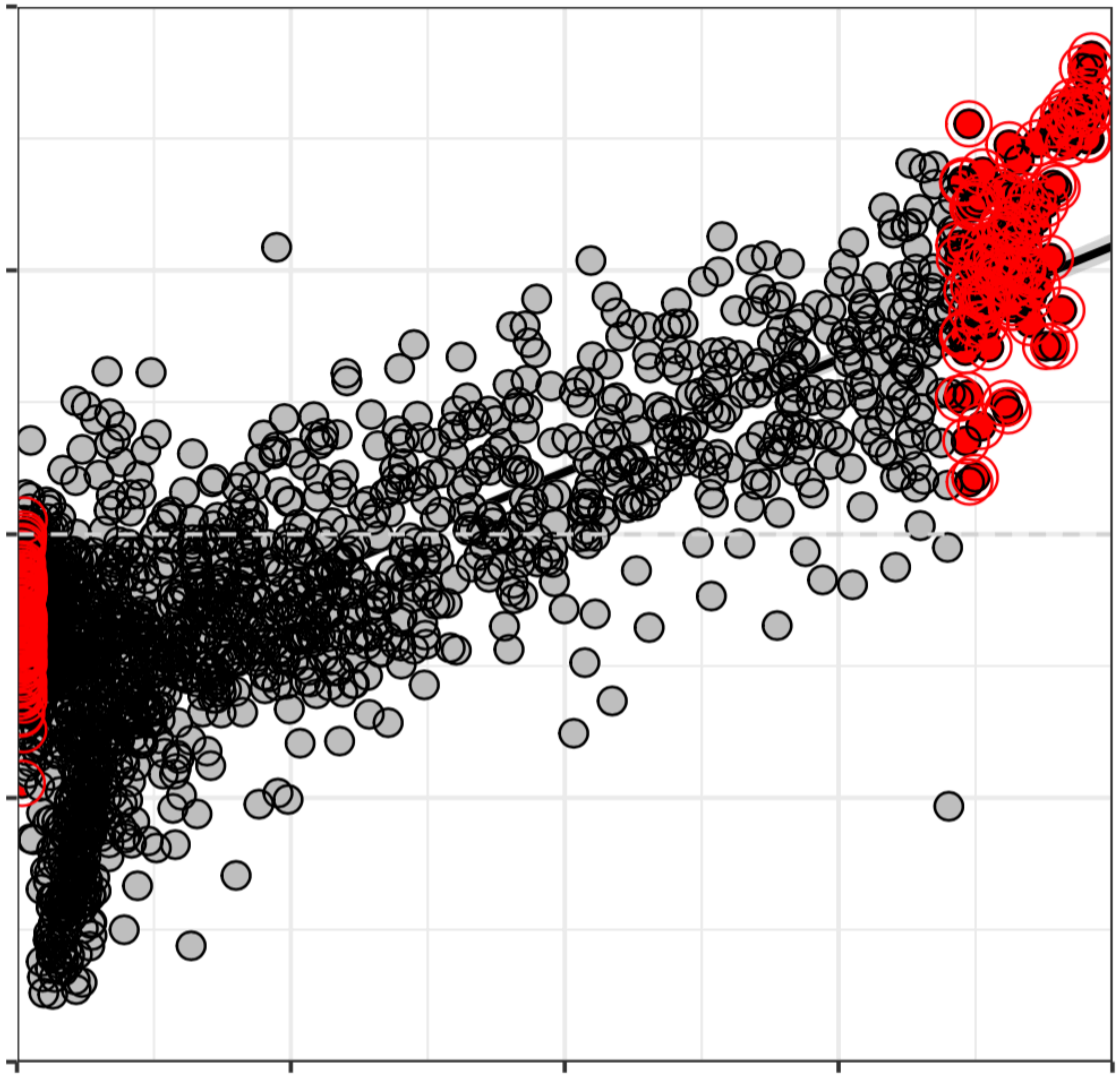




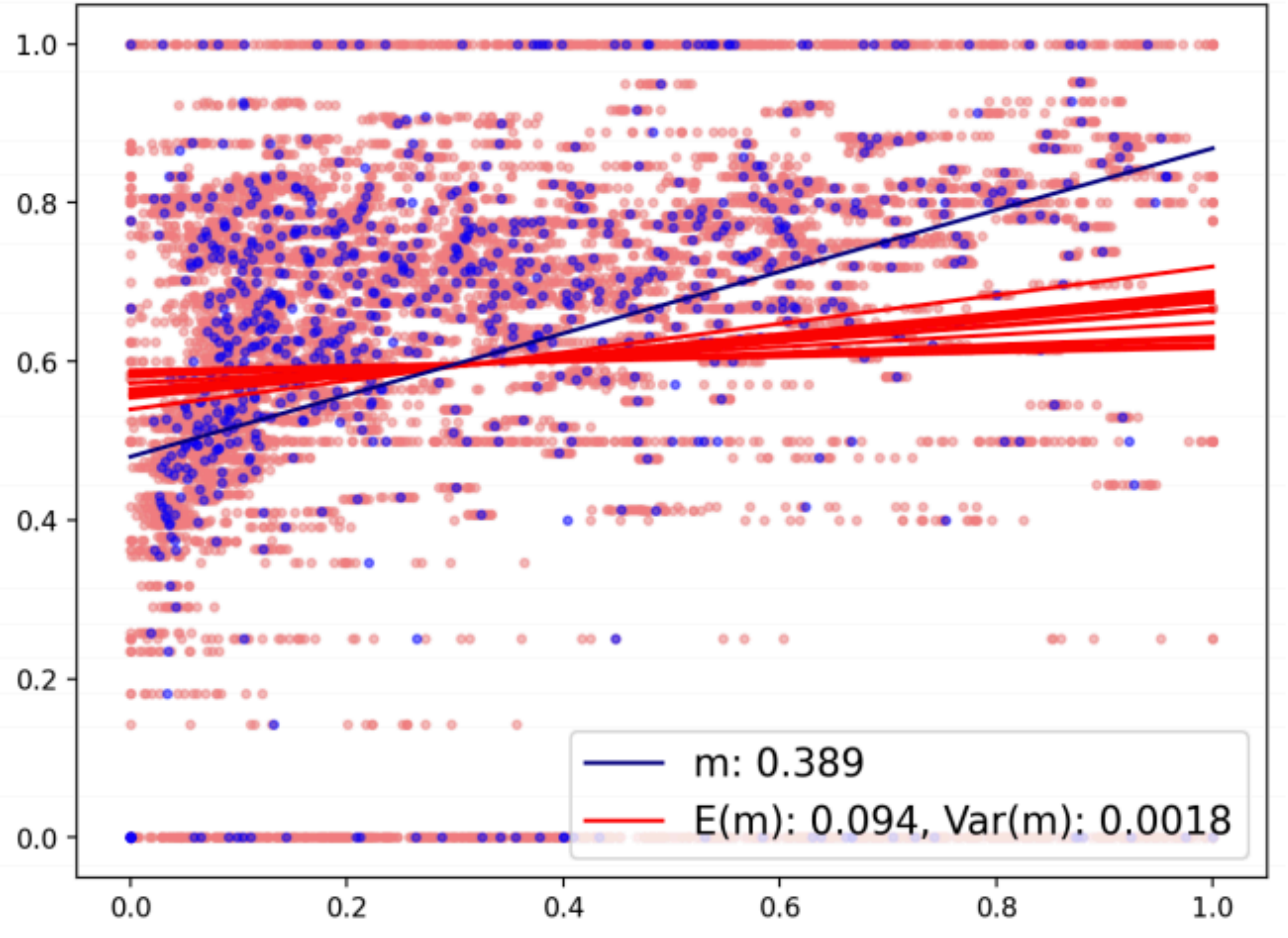
random district #46



Can we identify racially polarized voting?



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



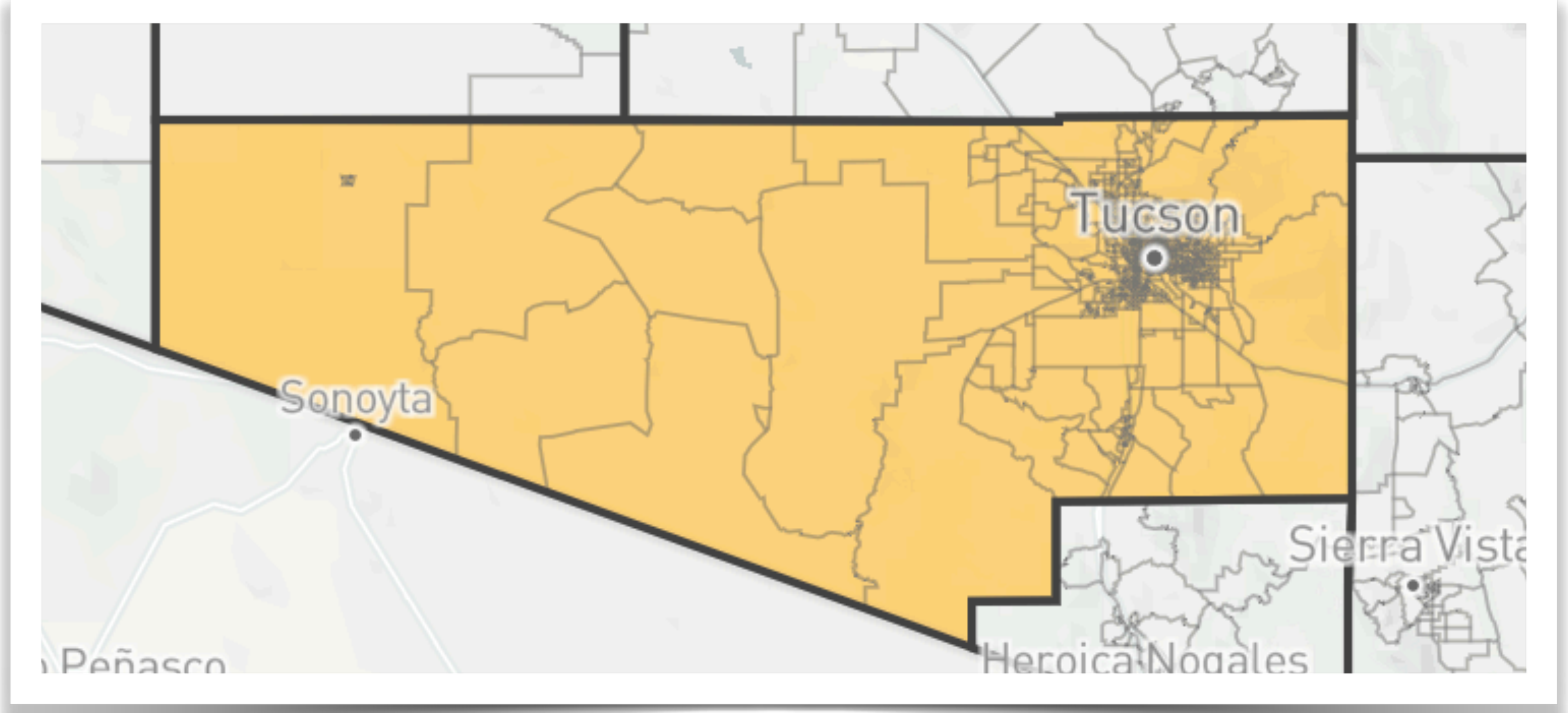
the nightmare scenario

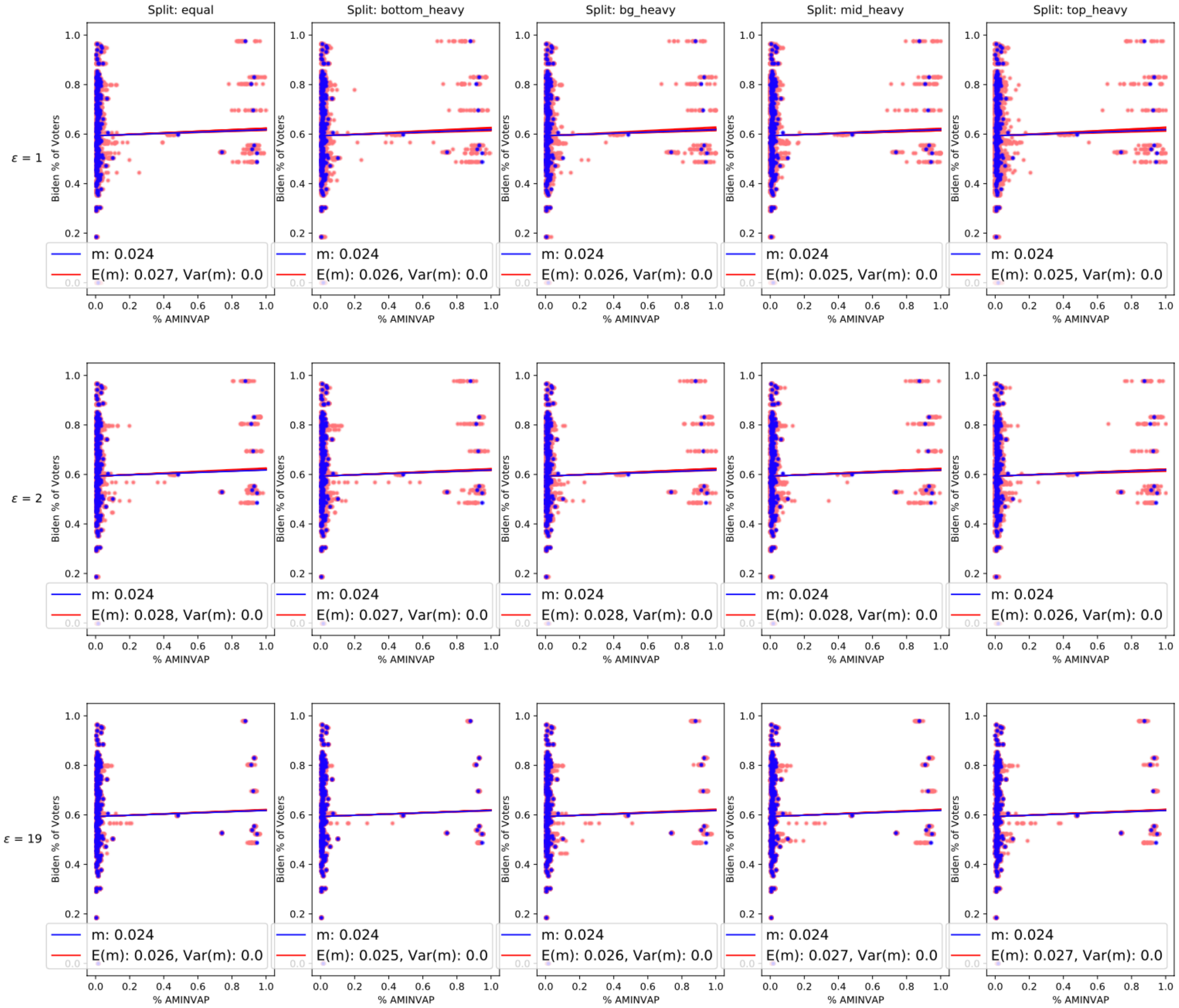
adding noise loses the signal of racially polarized voting

might be unable to test merit of VRA claims

Pima County

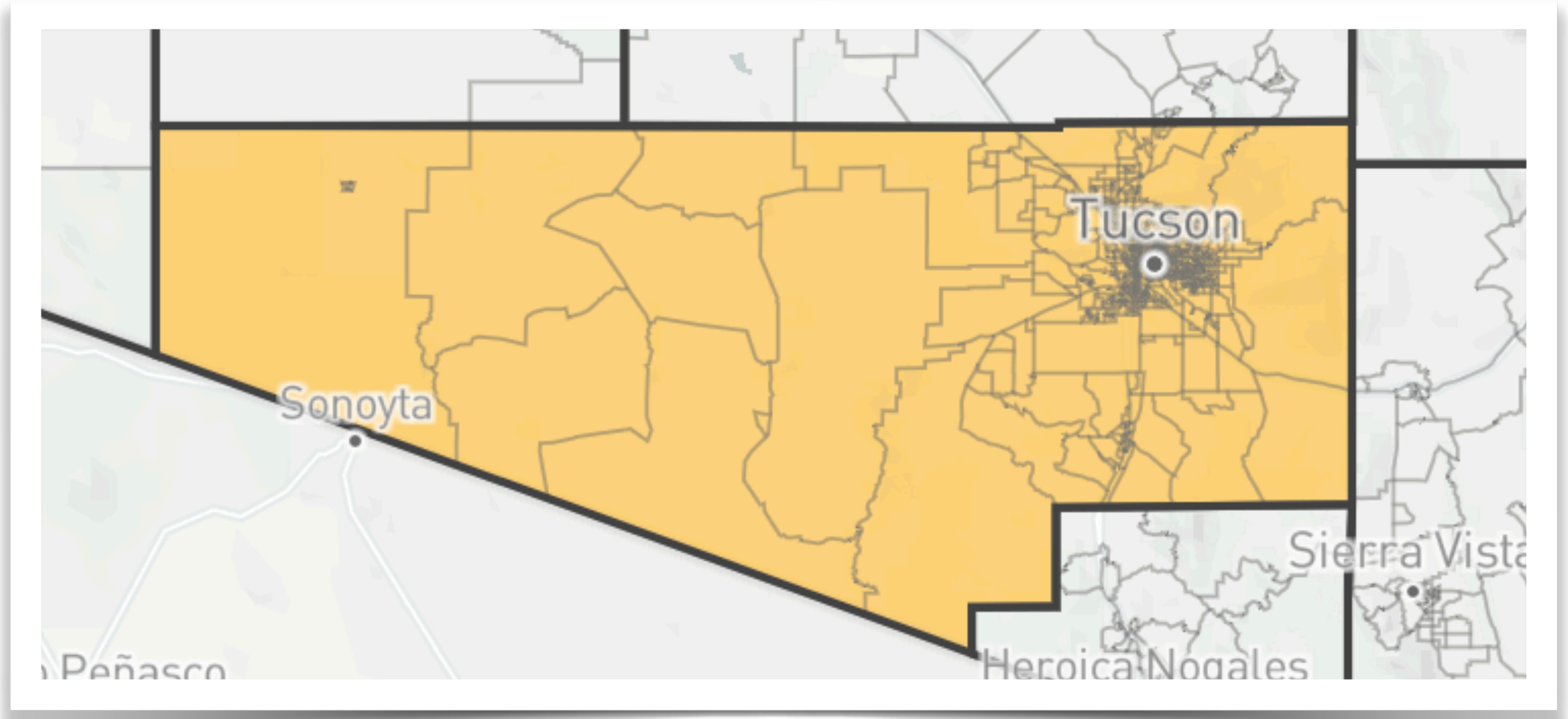
AMIN support for Biden

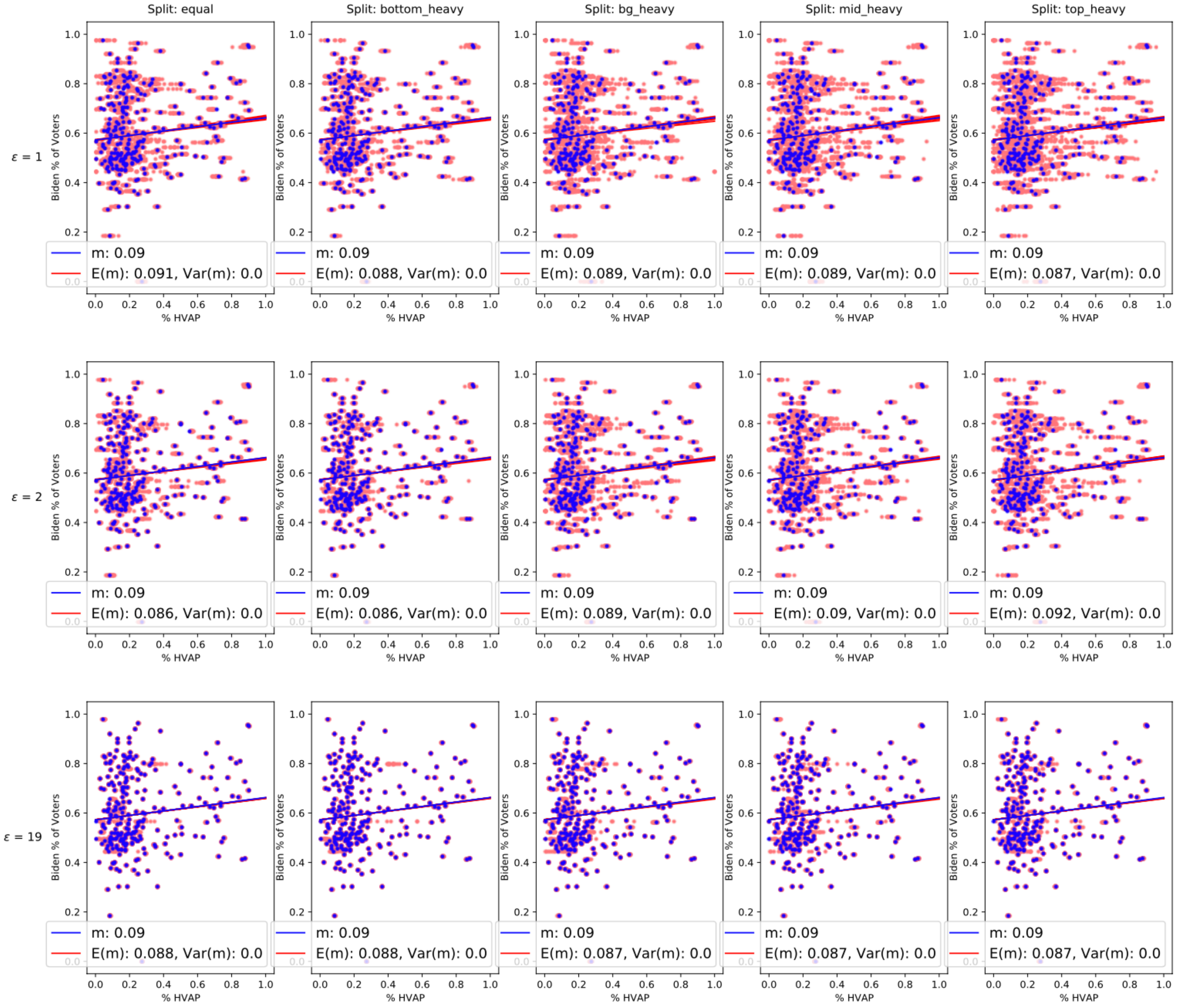




Pima County

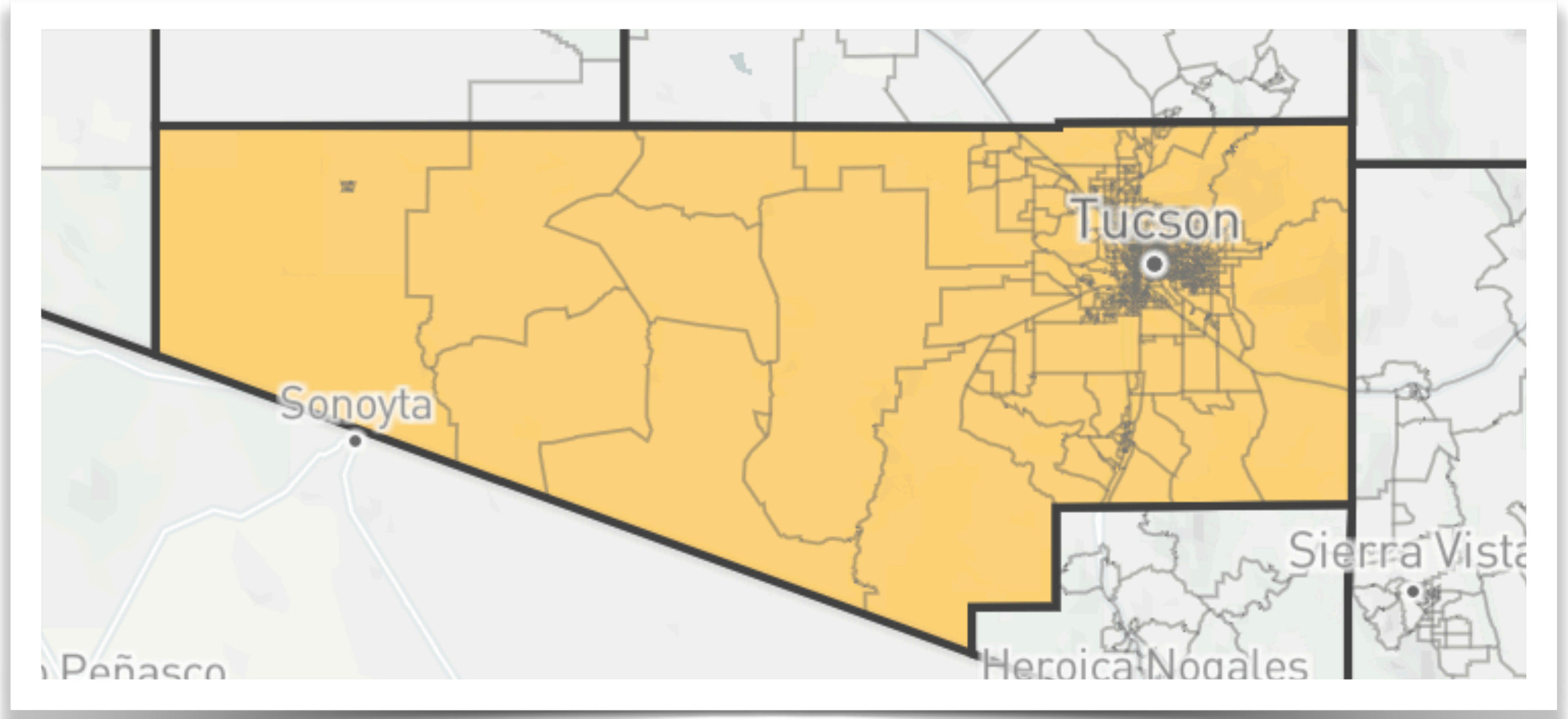
HISP support for Biden

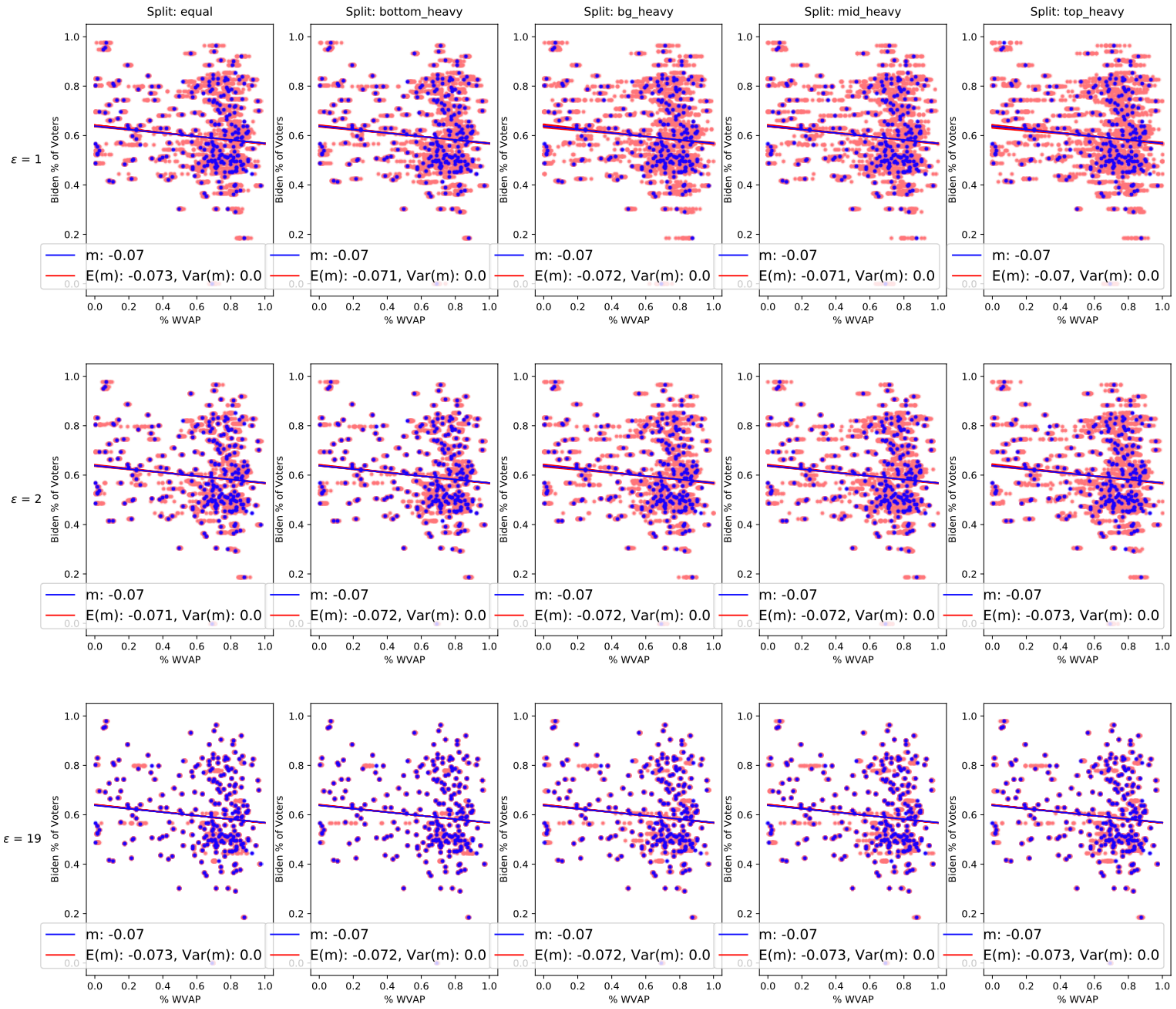




Pima County

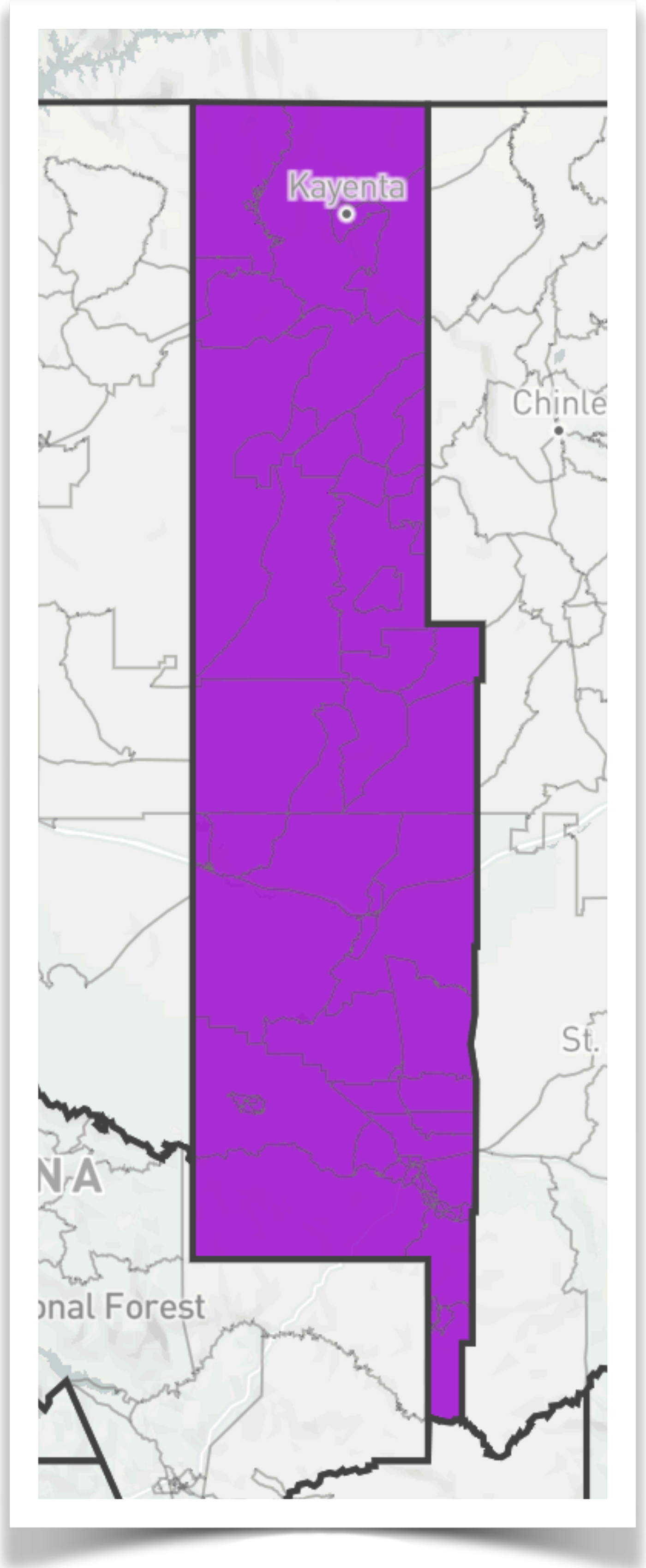
W support for Biden

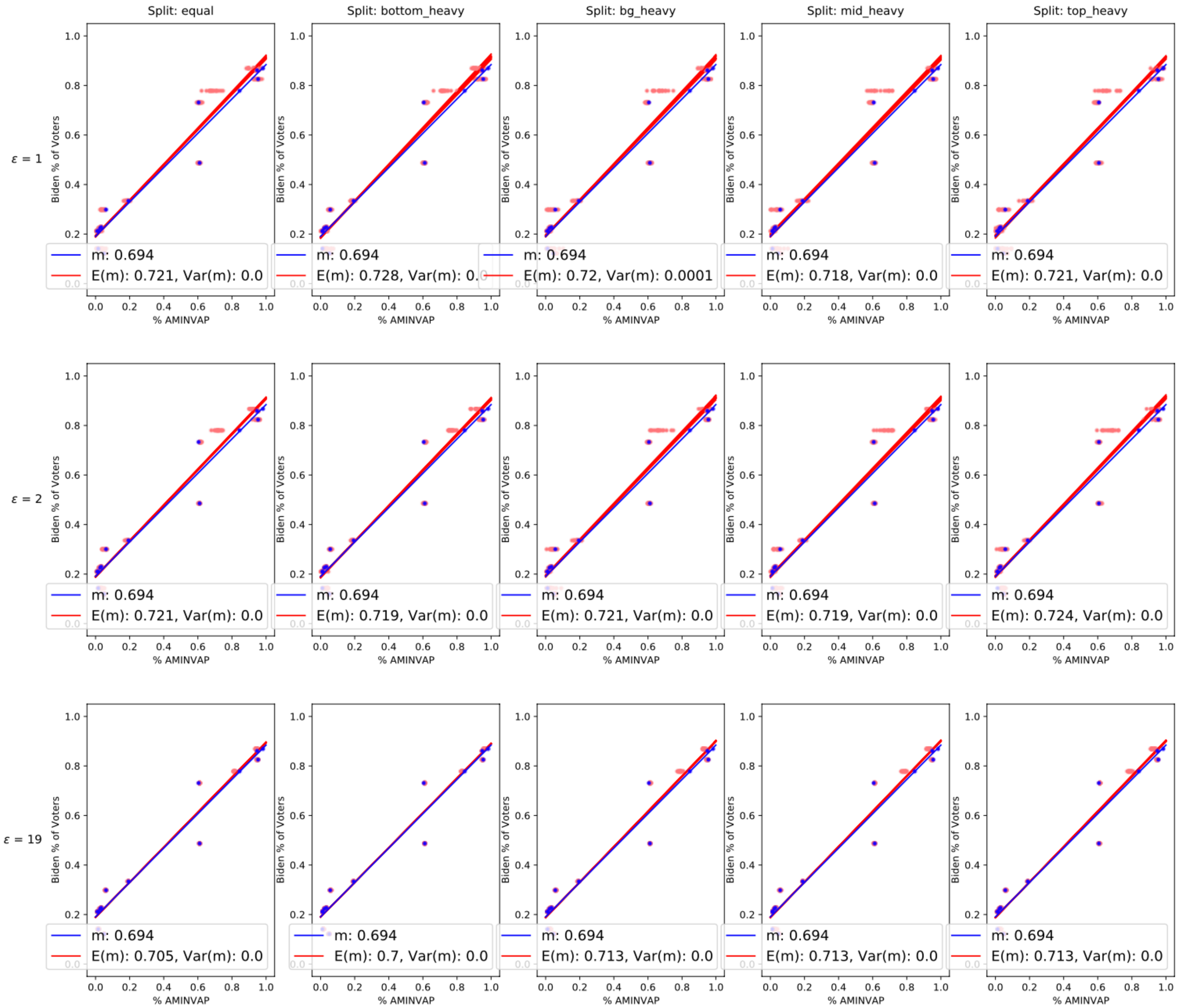




Navajo County

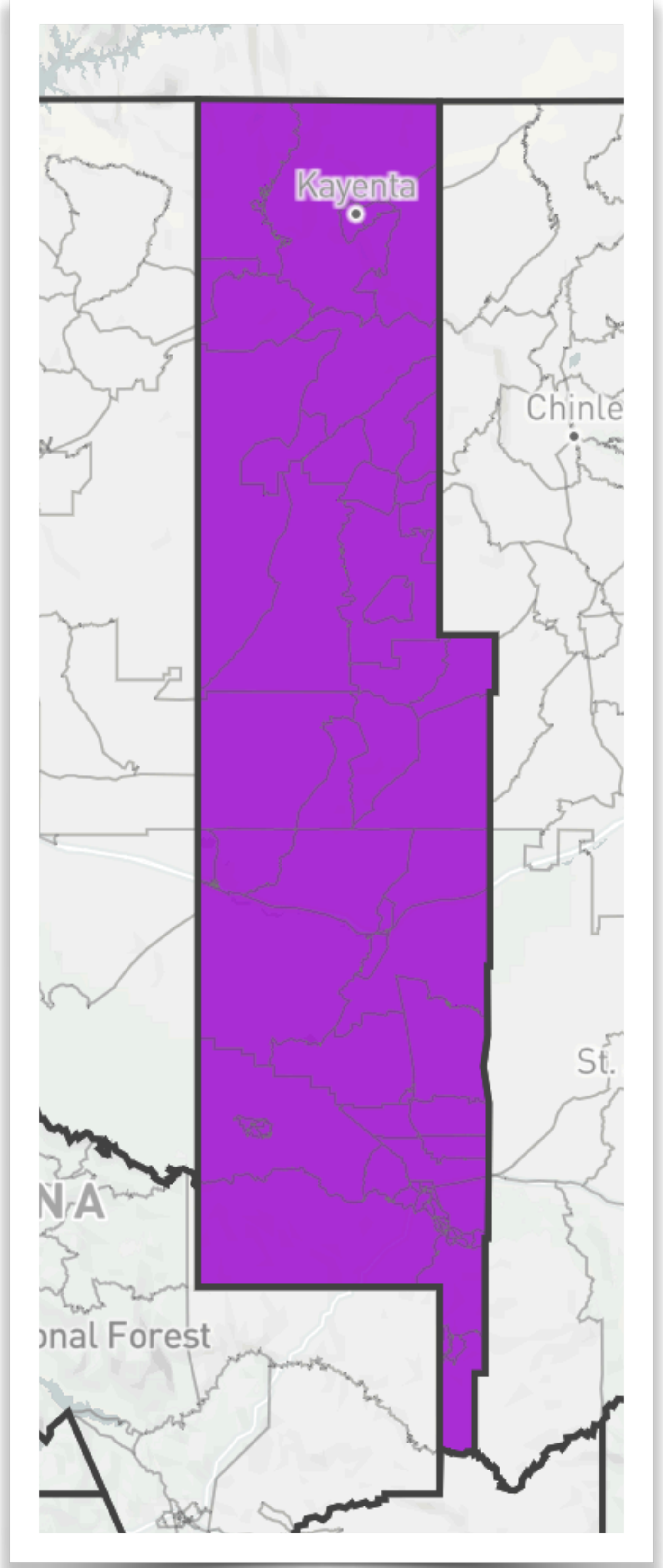
AMIN support for Biden

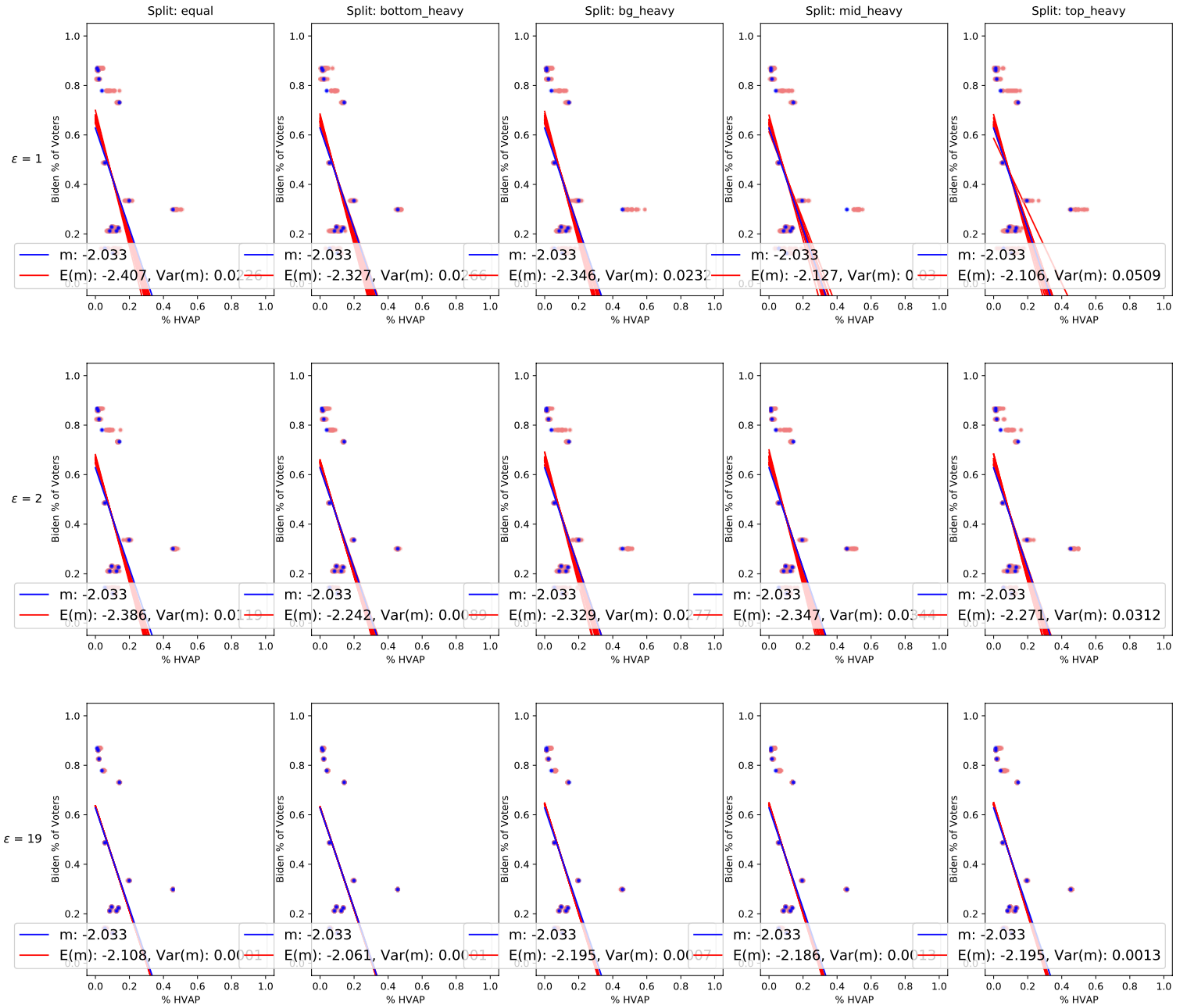




Navajo County

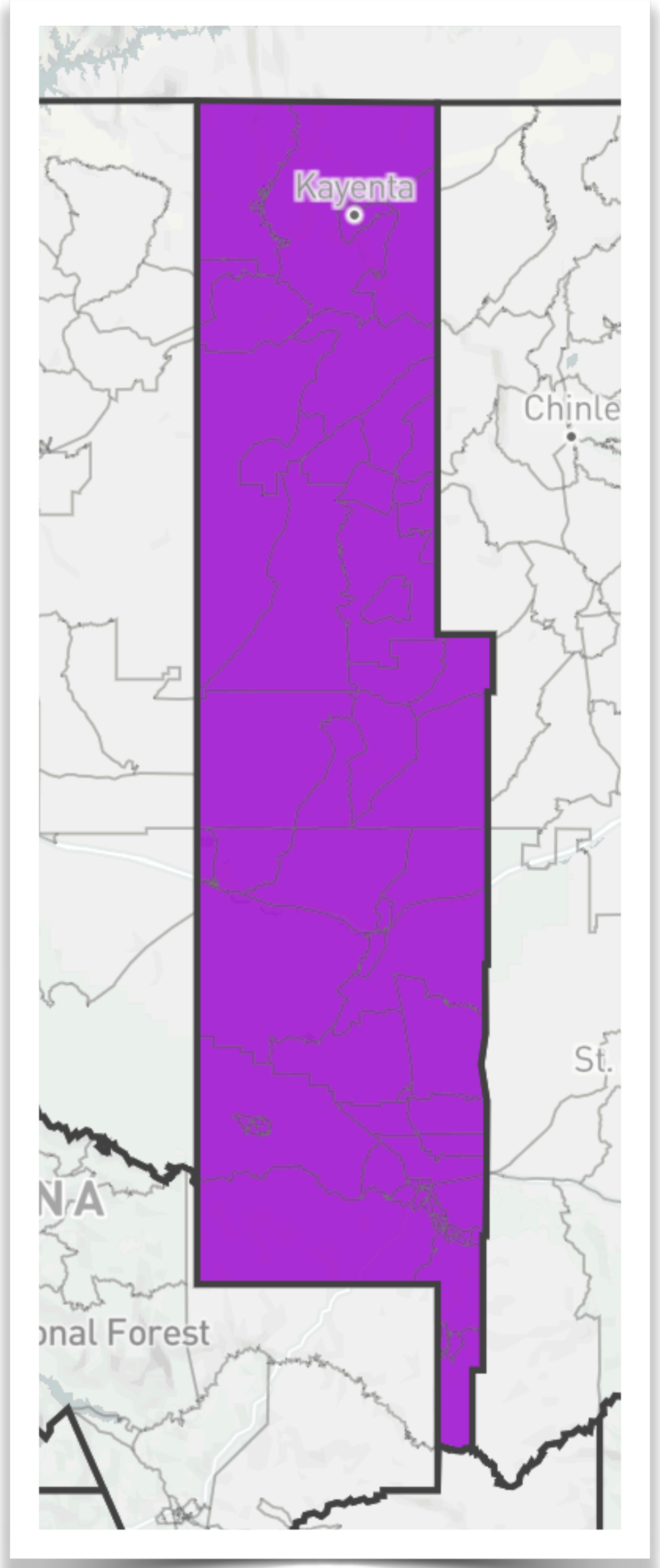
HISP support for Biden

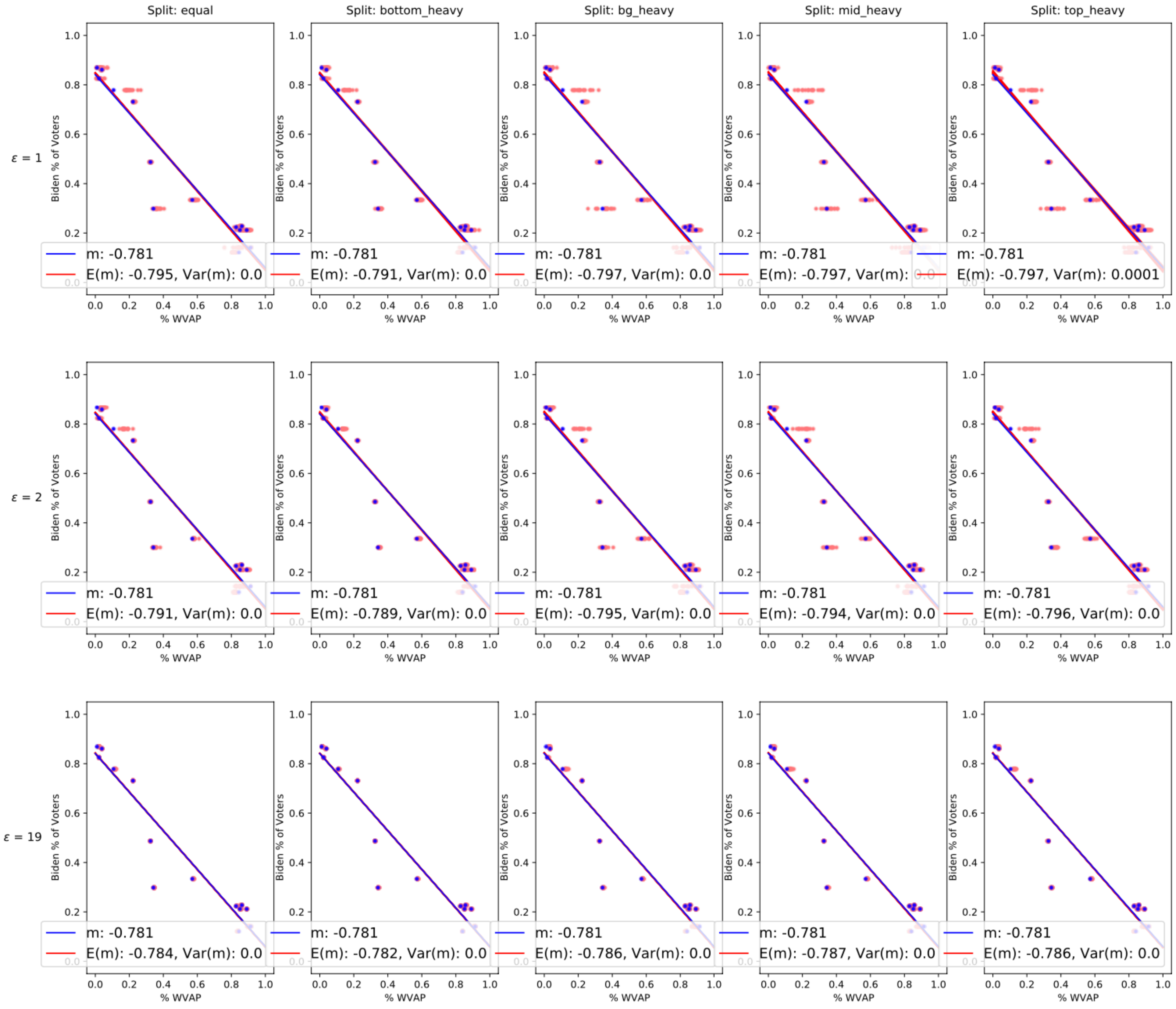




Navajo County

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