Data and the State

PUBPOL 2130 / INFO 3130

Data privacy

Lecture 24, Thursday May 1

Keeping data private

- We've talked about the difference between **microdata** (raw, person-level) and **aggregate data** (counts for a group of people, by type)
- The Census Bureau used to think they could protect data just by anonymizing and aggregating it
- In 1990, 2000, and 2010 Census releases, the Bureau tried to add an additional layer of protection by a practice called **data swapping**, where
- In the 2020 Census, they switched to a new form of disclosure avoidance (privacy protection) using an idea from computer science called **differential privacy**
- Just like a derivative in calculus asks for how much a function's output changes when the input changes by a small amount, differential privacy asks how much your quantitative results can change when *one person is removed from the raw data*

Dwork

- The idea of differential privacy was invented by Cynthia Dwork and collaborators. She's a CS professor at Harvard.

- Dwork is coming to give lectures here at Cornell next week: Talk 1 is Monday May 5 at 3:45pm. (there are two more talks, but they look a lot more technical....)

- You can get extra credit if you go the talk and send me an email afterwards with a one-paragraph summary and response.

Cynthia Dwork

How it works

- Very basic idea: a privacy algorithm is called ε–differentially private if the removal of one person can't change quantitative results by more than $e^{-\varepsilon}$. (This is e like in natural logs, about 2.71828....)

- So if $\varepsilon = 1$, you get a factor of about .37, about one-third. Your answers can change by a factor of 3. This is usually considered a reasonable level of accuracy to trade off against privacy protection. But if $\varepsilon = 10$, you get a factor of one over 22,000, or about 0.000045. This means that the accuracy is very high, since your answers barely changed — so the privacy protection is very low.

- This is how differential privacy works! it's a **tradeoff between accuracy and privacy**, and for a particular policy application, you have to decide where you want to be on that tradeoff.

- Differential privacy was used in the 2020 Census through an algorithm called TopDown. Most scientists expected $\varepsilon = 1$ or $\varepsilon = 2$. But the Bureau got spooked by the freakout over noising their data so they went even farther than $\varepsilon = 10$ and used about $\varepsilon = 19.6$.

- I call this "barely breathing on the data"!!

Helping policymakers understand the science

- The rest of these slides come from a presentation I prepared for policymakers in Arizona who were worried that differential privacy might make their data unusable.

- I had a few goals:
    - explain the science
    - tell them why there was a real problem that the Bureau had to address — to do this, my team actually conducted a privacy attack on their data! (they asked me to.)
    - tell them that the Bureau's differentially private solution was not going to make their work harder to do — to do this, my team checked if racial and ethnic groups would see big changes in their size or if racially polarized voting would get harder to identify

- In the end, we had both suggestions for the Bureau and suggestions for the redistricting commission.

Privacy, Census Data, and Arizona Redistricting

an overview with experiments

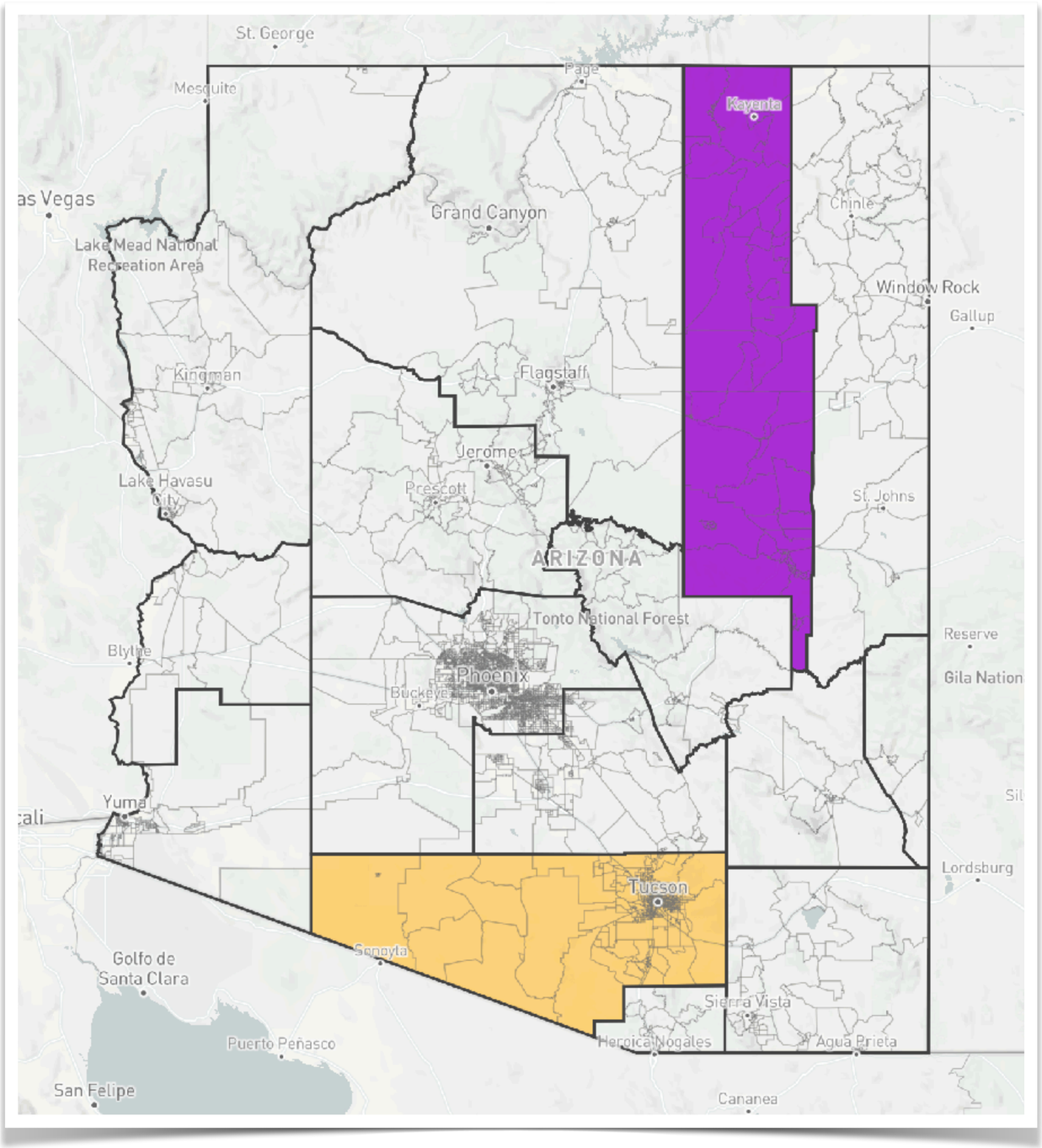
a presentation for the Arizona Independent Redistricting Commission from July 2022

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

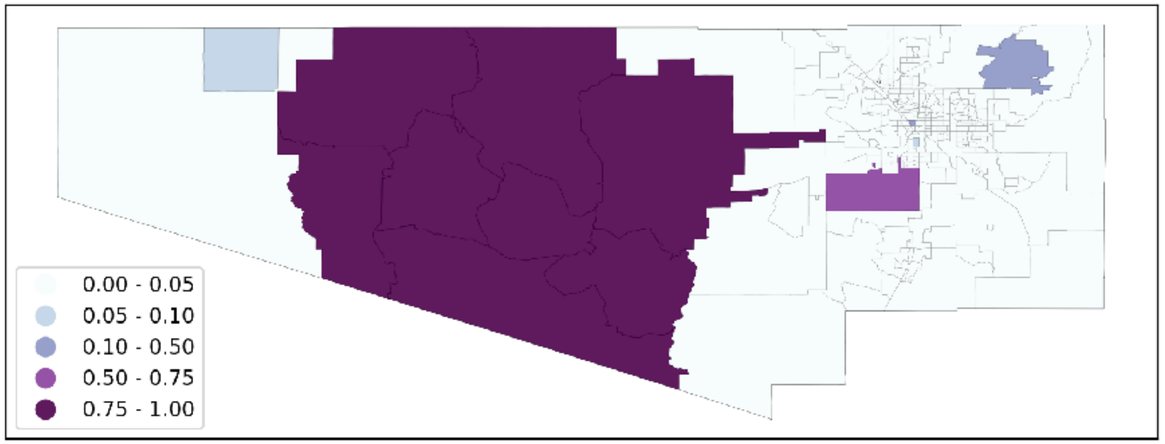
44%W, 11%H, 42%AMIN

Pima County, pop. 980,263

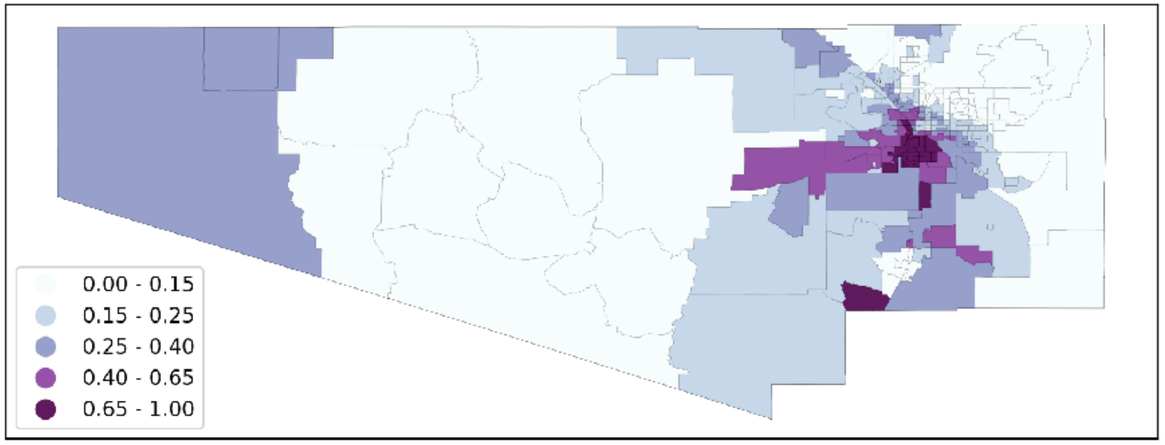
55%W, 35%H, 2.5%AMIN

Both counties have significant diversity

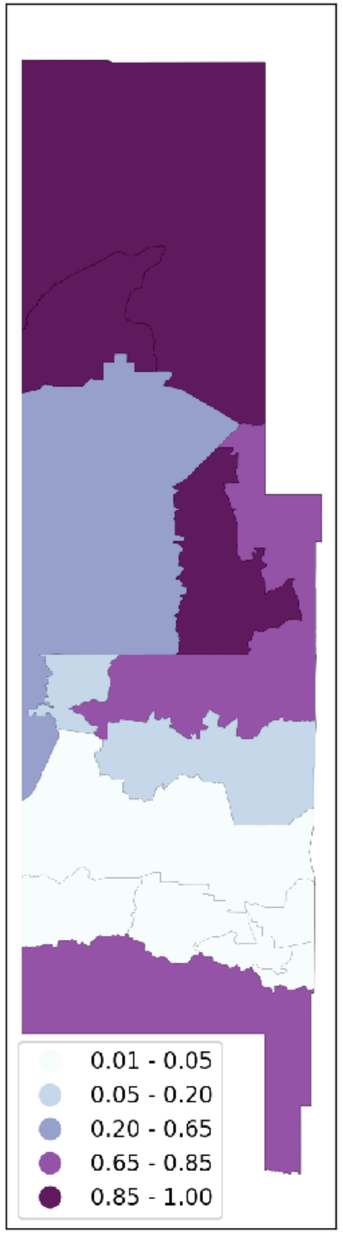
AMIN population in Pima County



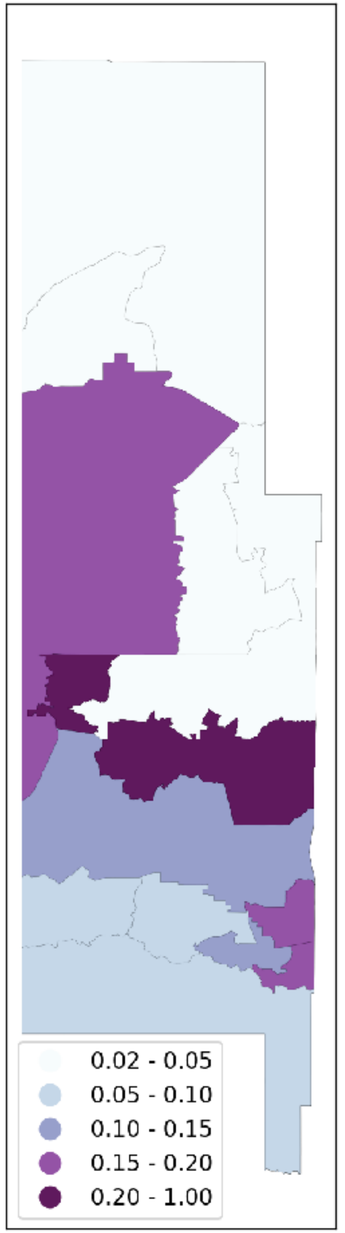
Hispanic population in Pima County



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

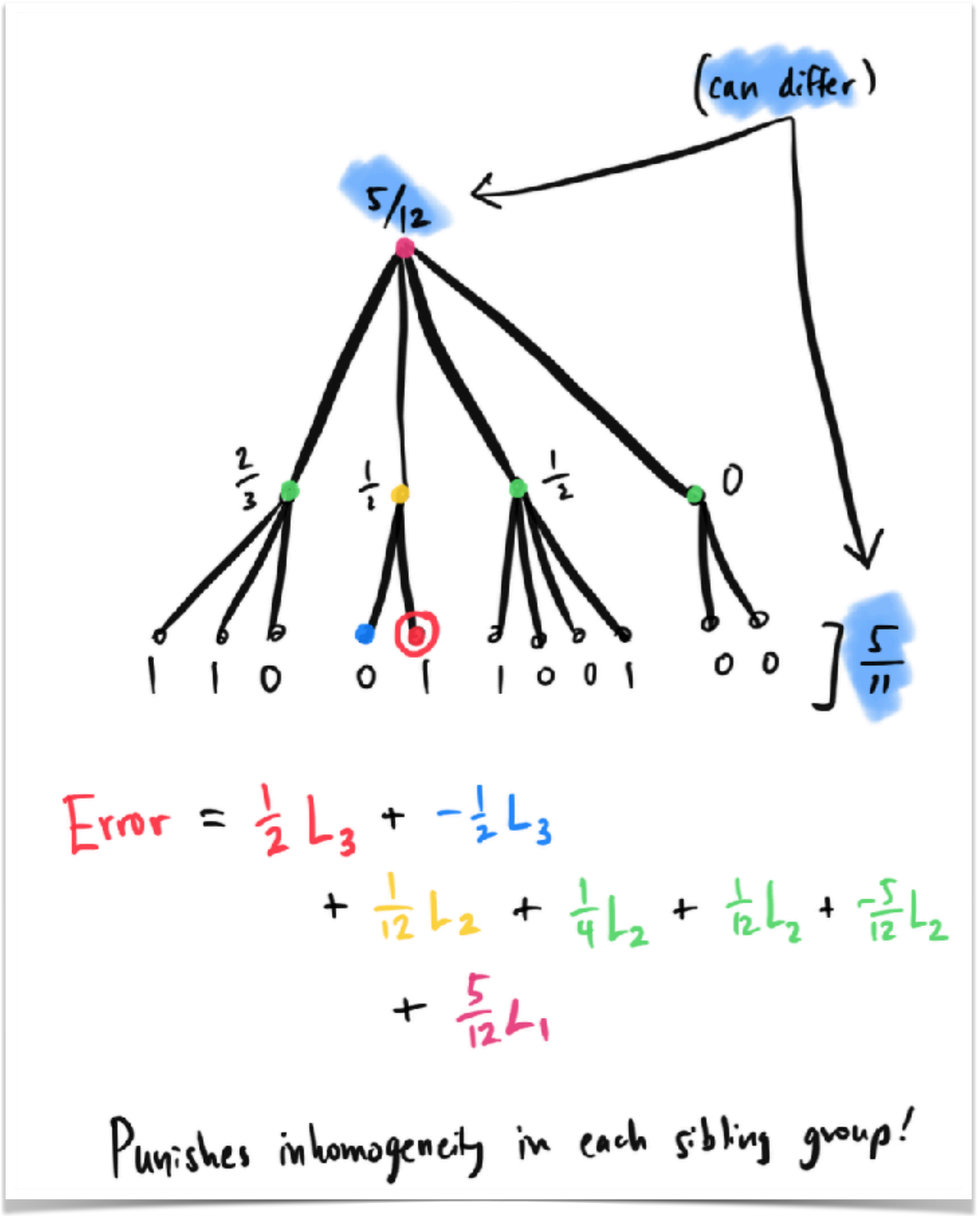
can get whole state in a few days

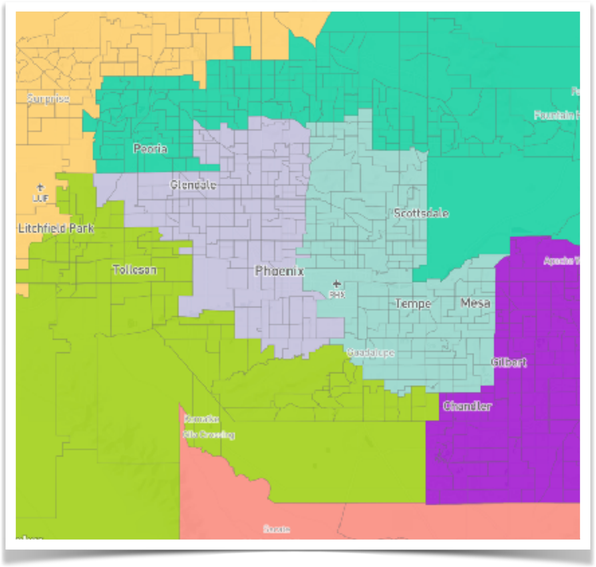
our table is **100% consistent** with the aggregate numbers released by the Census

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

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

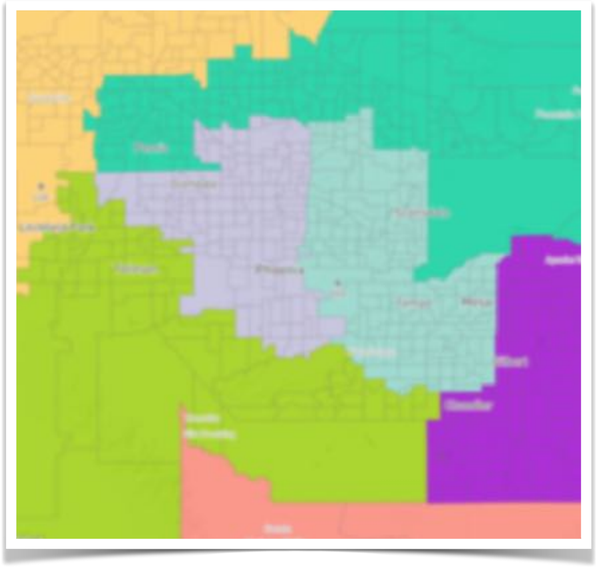
What is differential privacy?



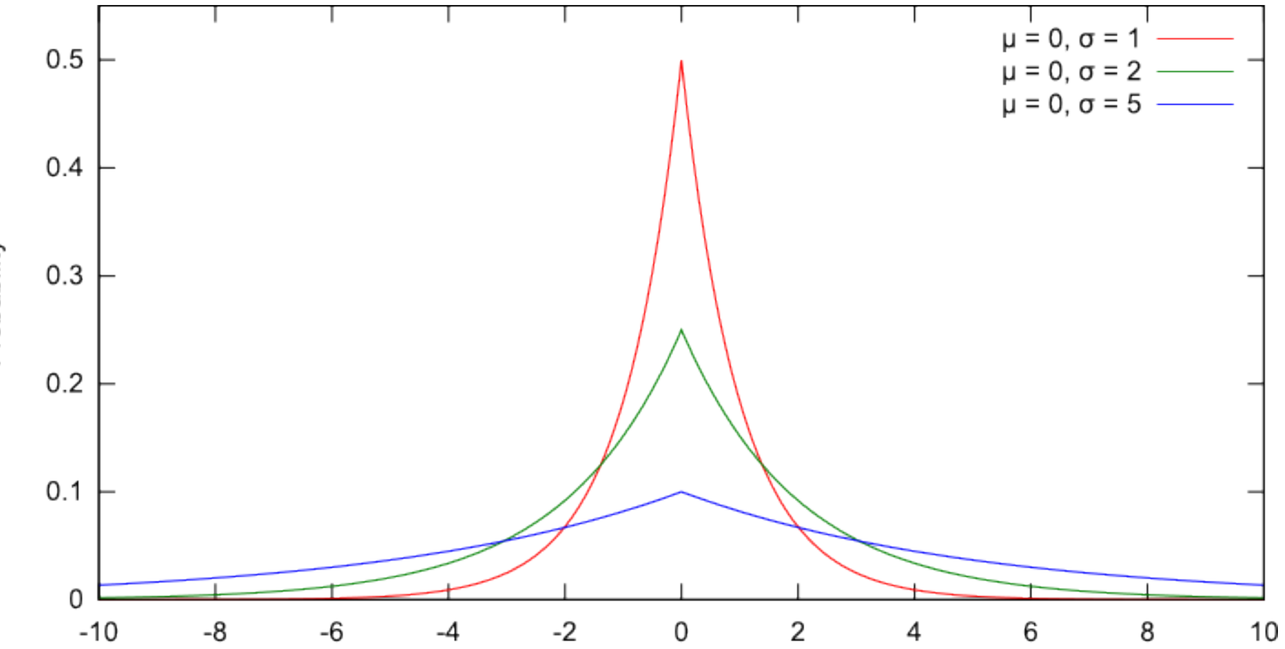


Idea: for **privacy**, add **noise**

make the numbers fuzzier so
exact reconstruction is impossible



…but data still aggregates up
with high accuracy!



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

What is the worry?

Census “TopDown” algorithm

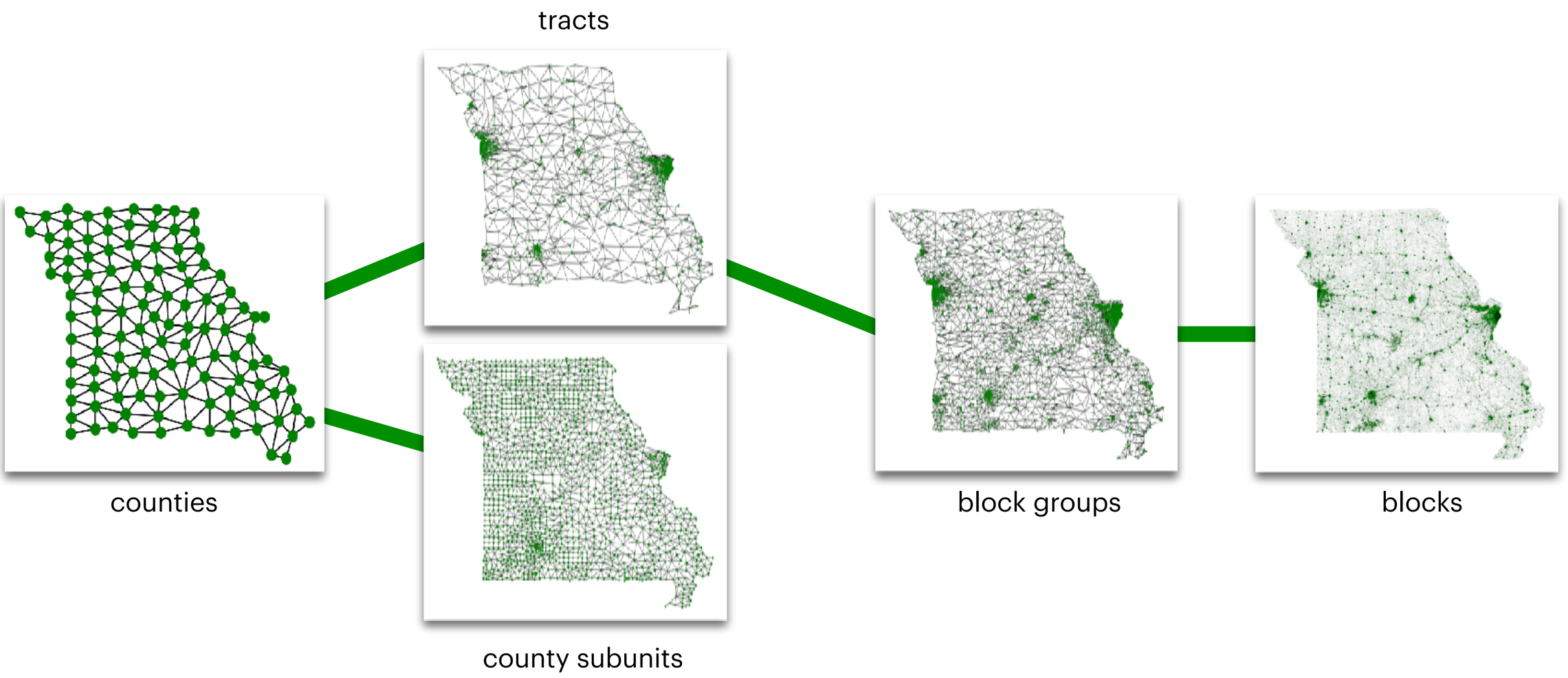
two main things to know:

(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

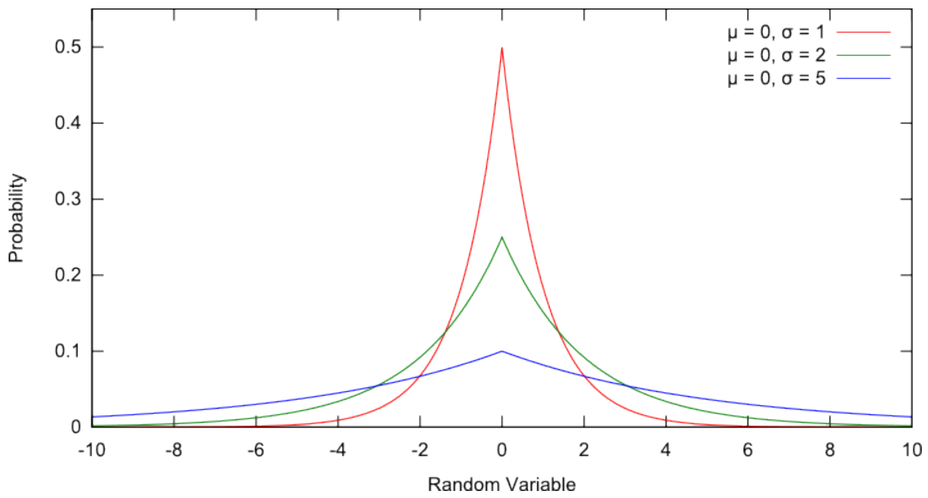
top

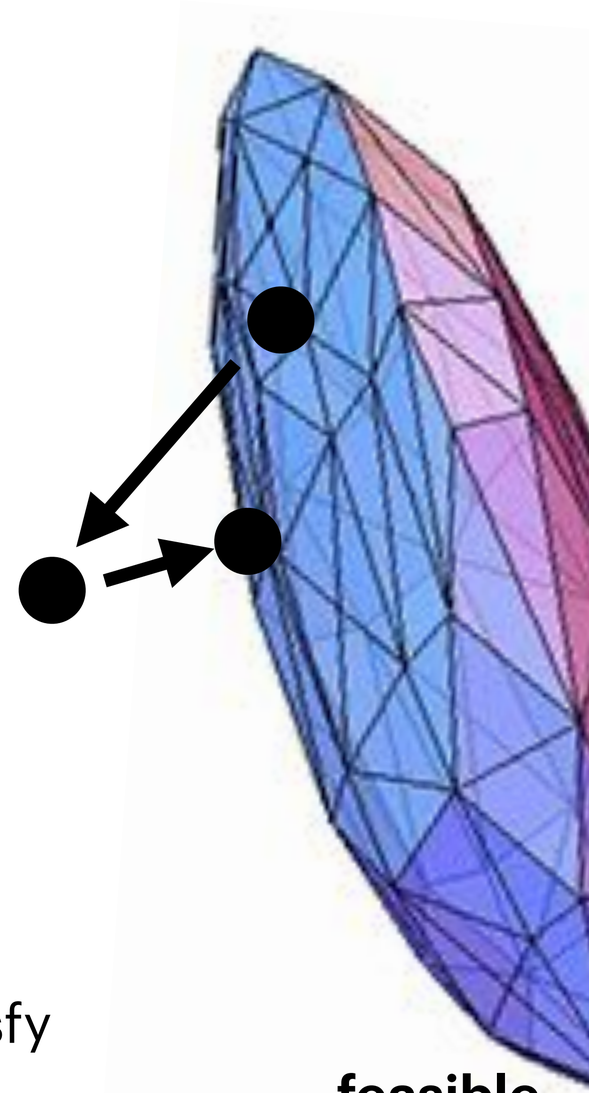
down



For each histogram bin, additive noise drawn i.i.d. from

Laplace distribution: $L_j \sim \mathrm{Lap}\left(\frac{1}{\varepsilon_j}\right)$ for $\varepsilon = \varepsilon_1 + \ldots + \varepsilon_d$.





Then **post-processed** to satisfy plausibility invariants.

feasible region

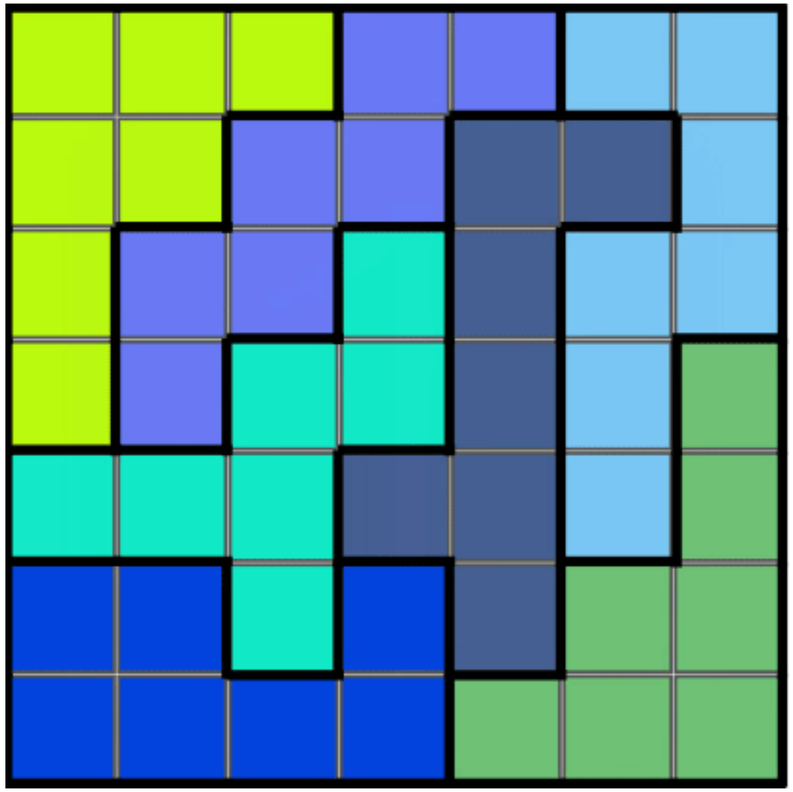
examples: state totals unchanged, no negative numbers, no households with 40 children and no adults, ...

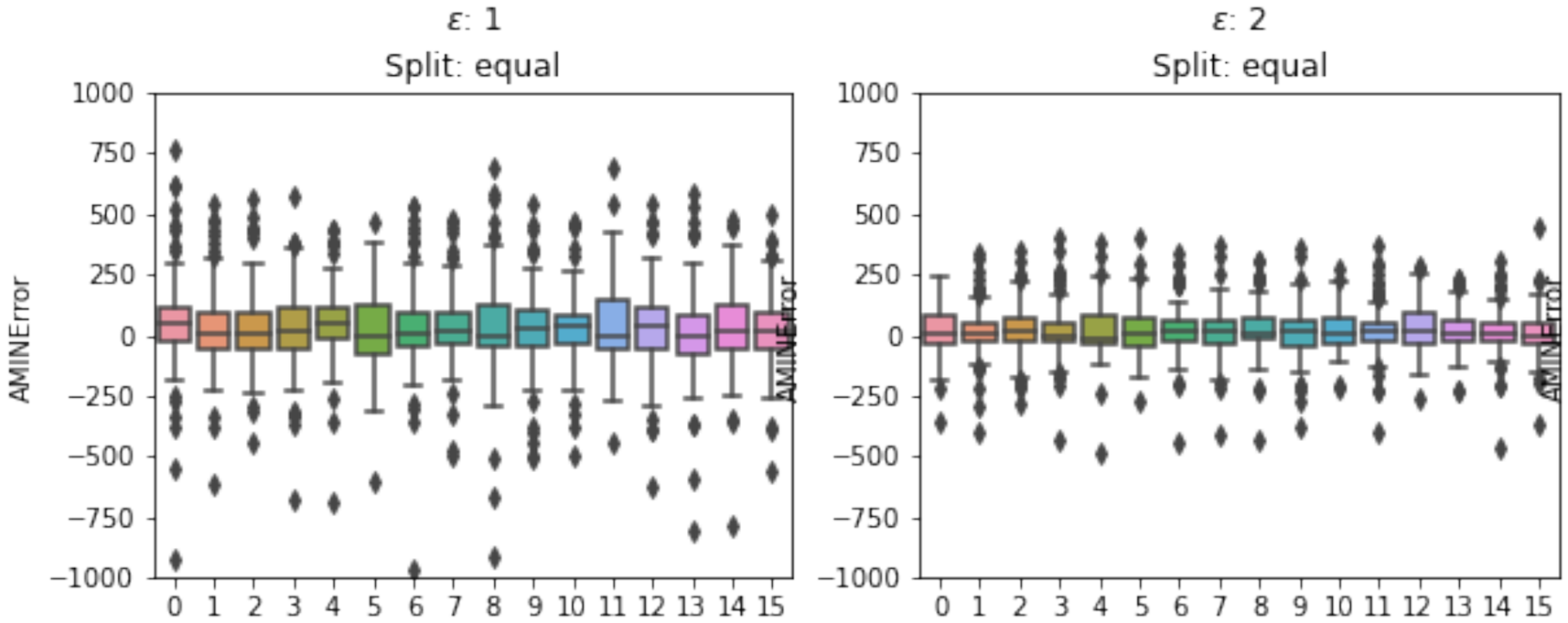
let’s see some

experiments

we’ll use a simplified model called “**ToyDown**” — see mggg.org/dp

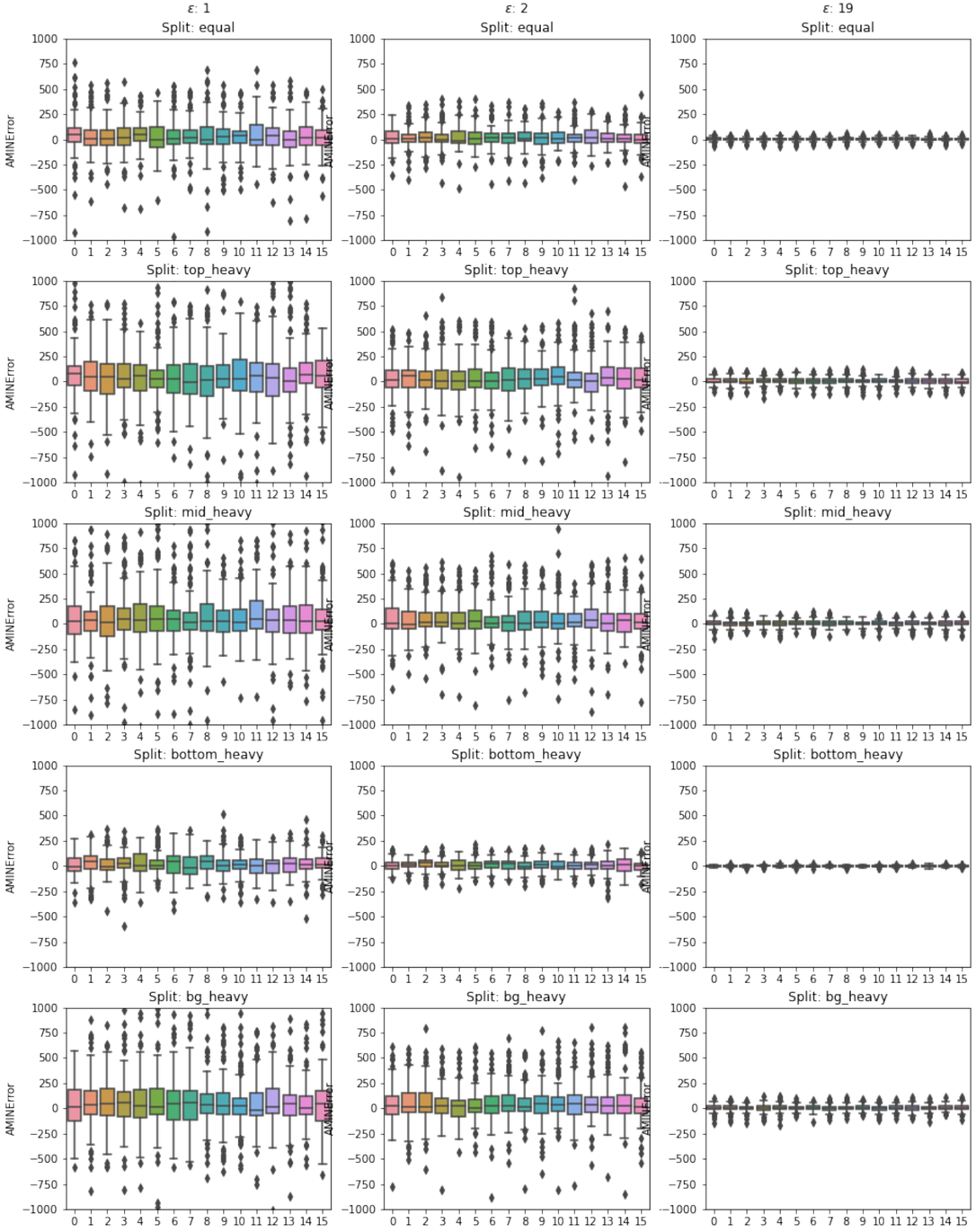
Do districts lose Native population?





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

…truly tiny at $\varepsilon = 19$



$\varepsilon = 1, 2, 19$

Navajo County

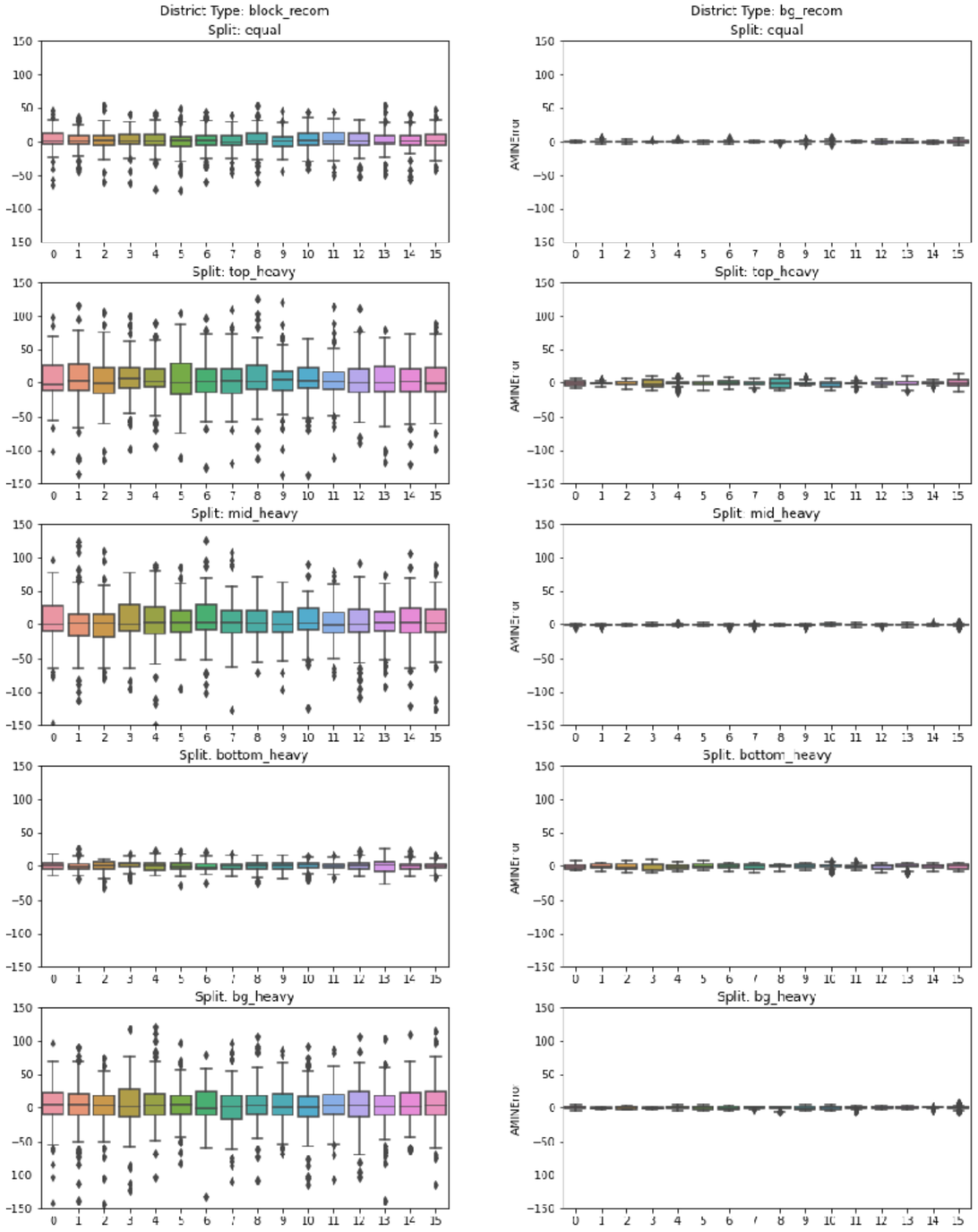
k=5 districts, population 20K

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**



built from blocks vs. block groups

Navajo County

k=5 districts, population 20K

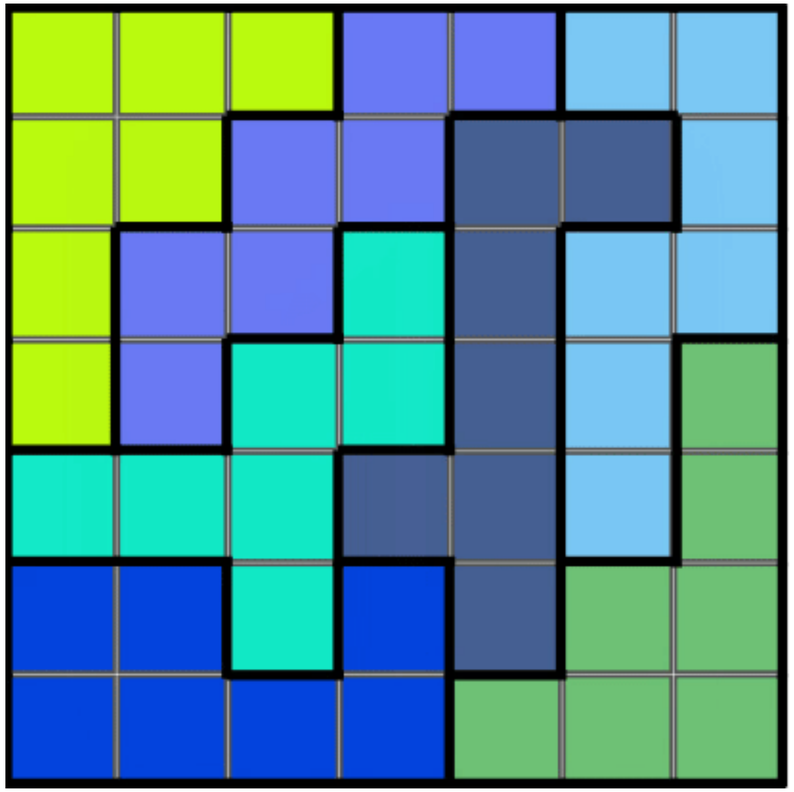
these plots show the discrepancy introduced by top-down style differential privacy

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

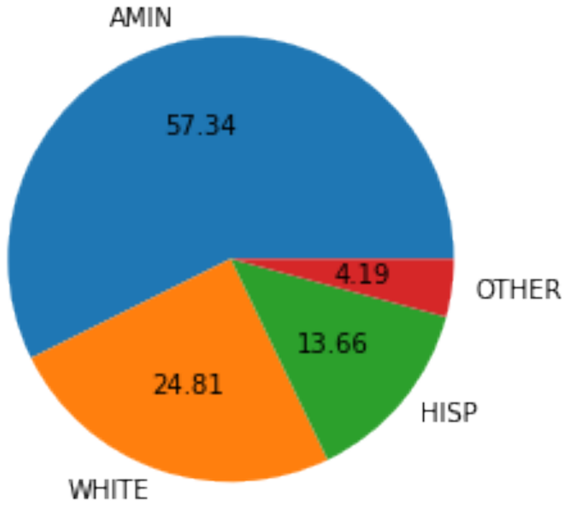
construction matters!

far better accuracy on districts built from larger pieces

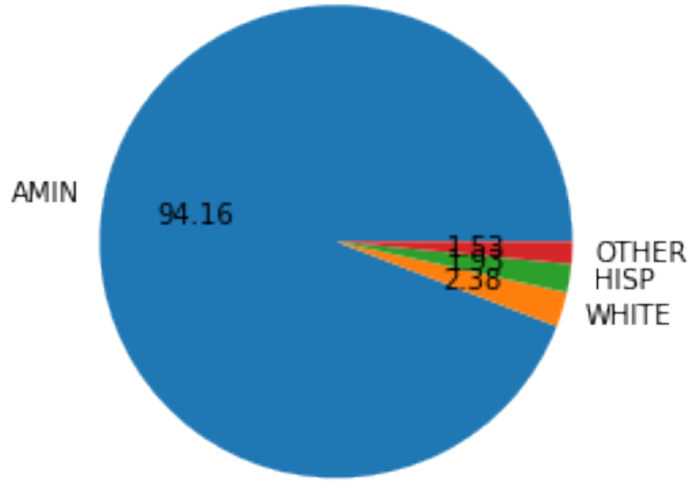
Do districts change their overall racial composition?



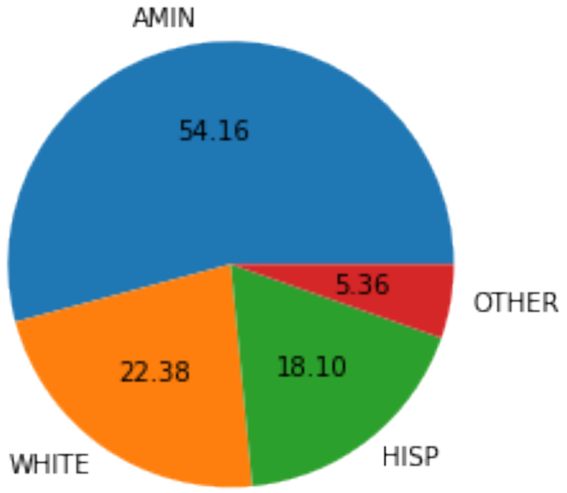
we will noise these 16 times with $\varepsilon = 2$ and equal allocation over the geographical levels



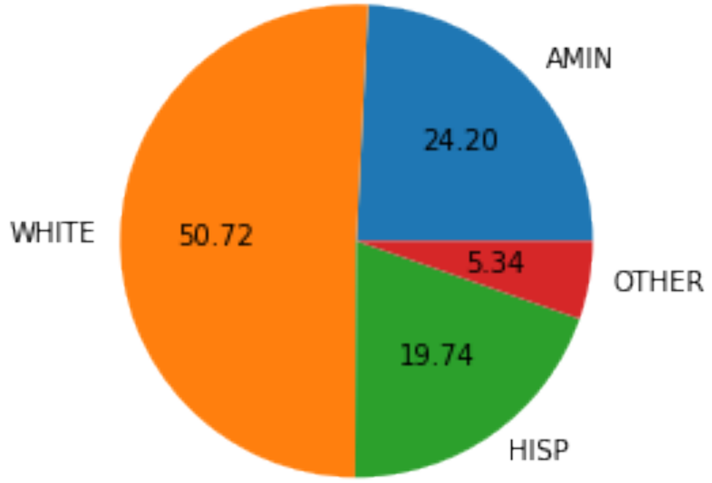
random district #2



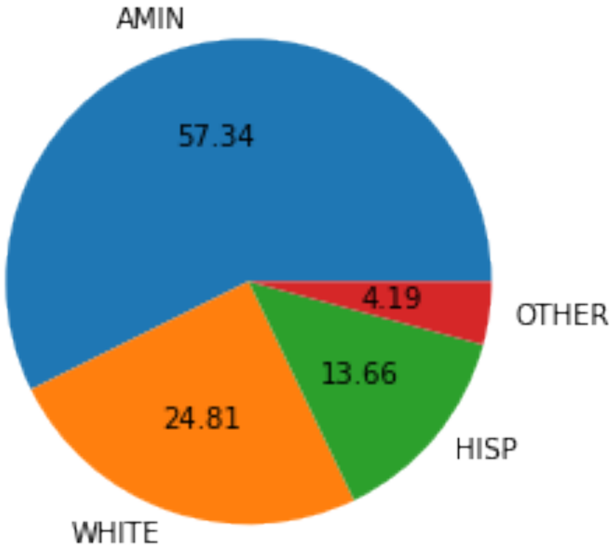
random district #9



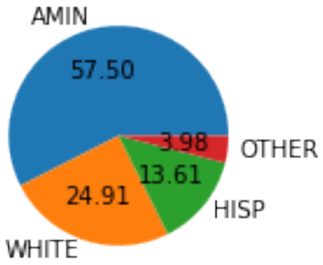
random district #13

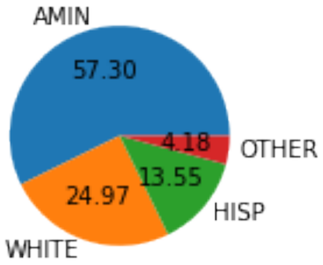


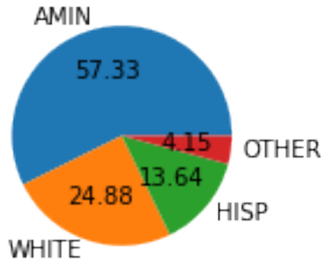
random district #46

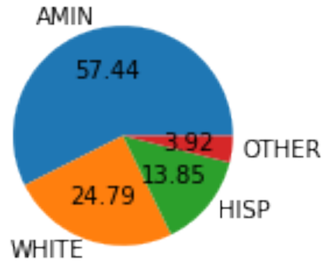


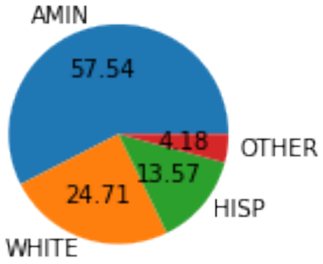
random district #2

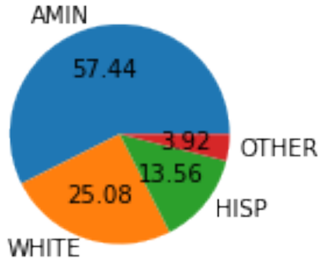


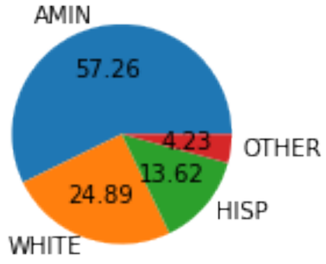


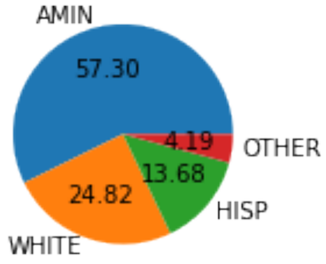


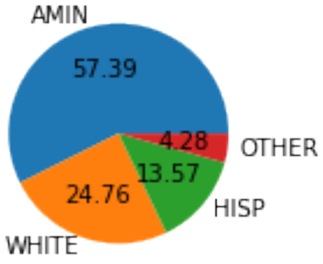


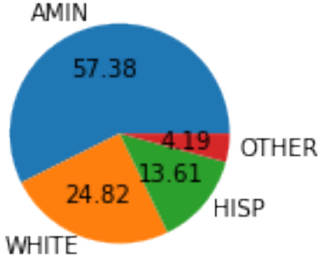


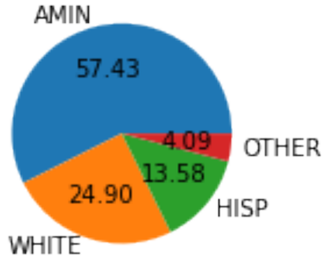


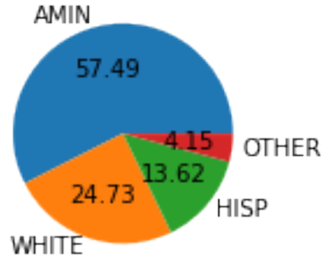


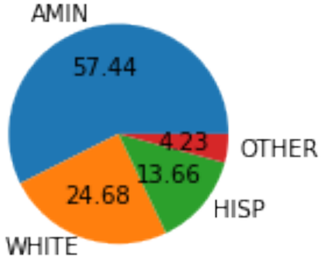


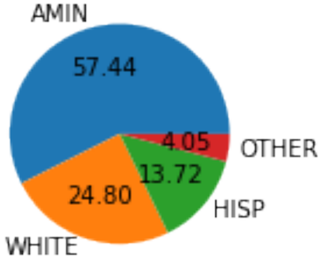


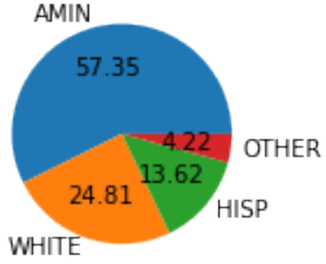


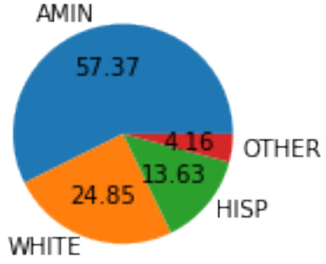


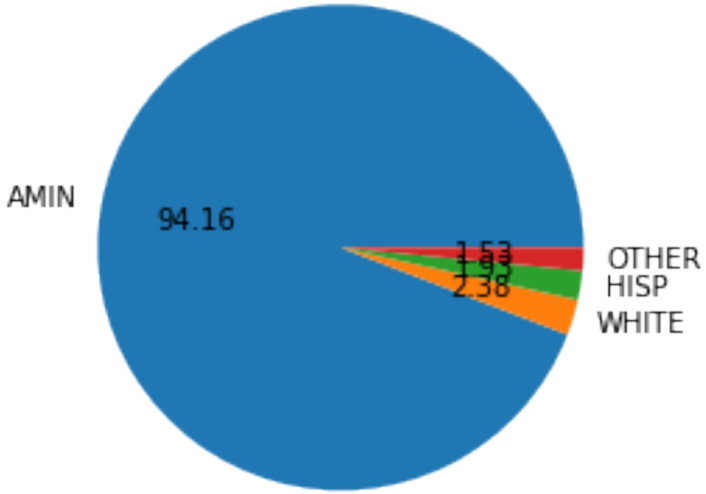




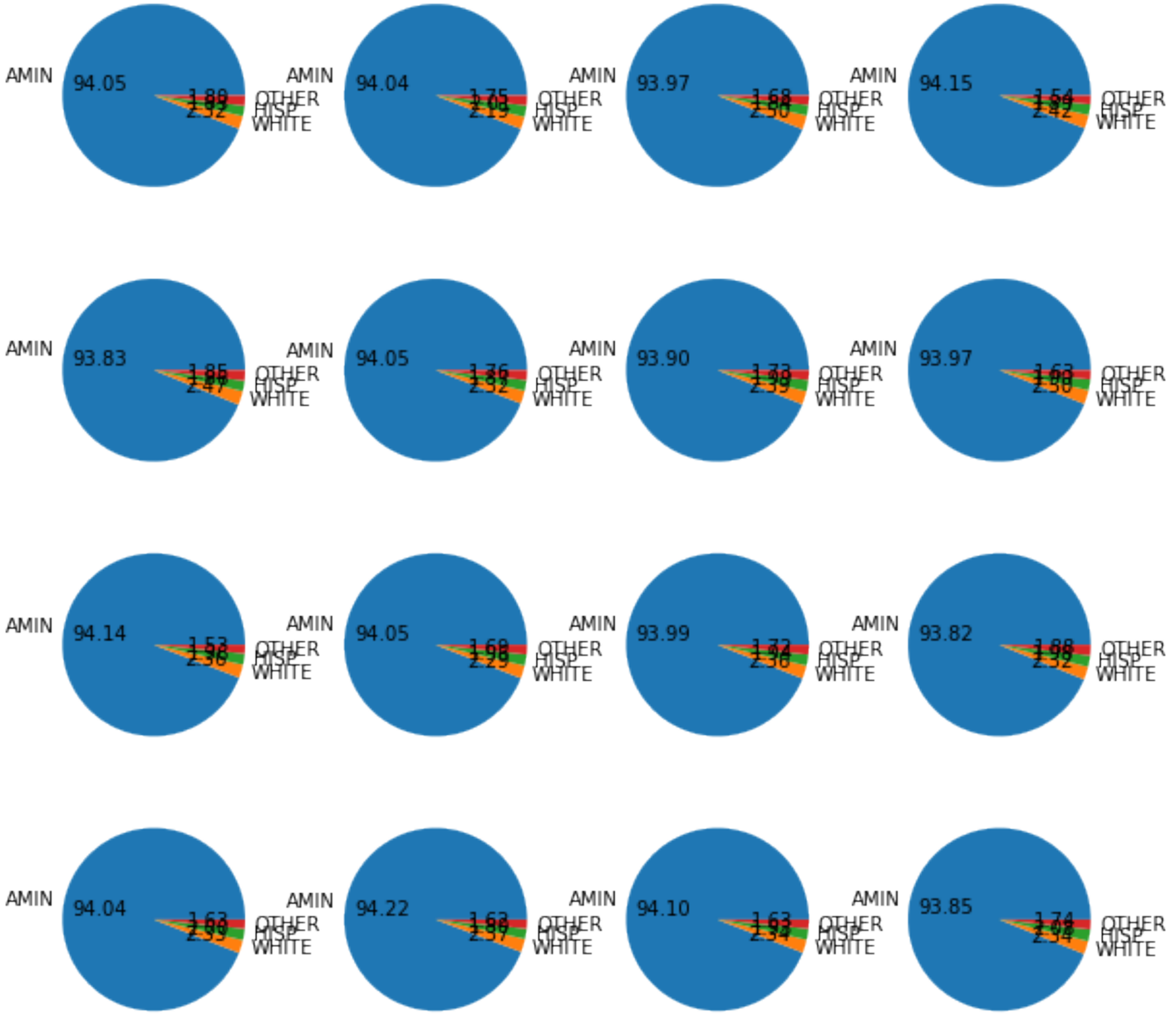


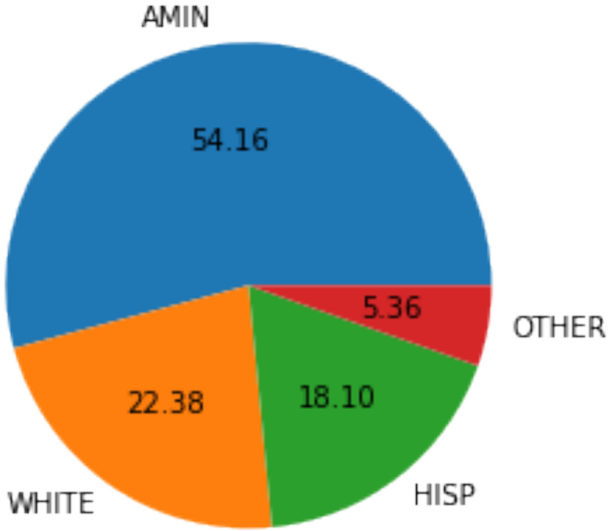




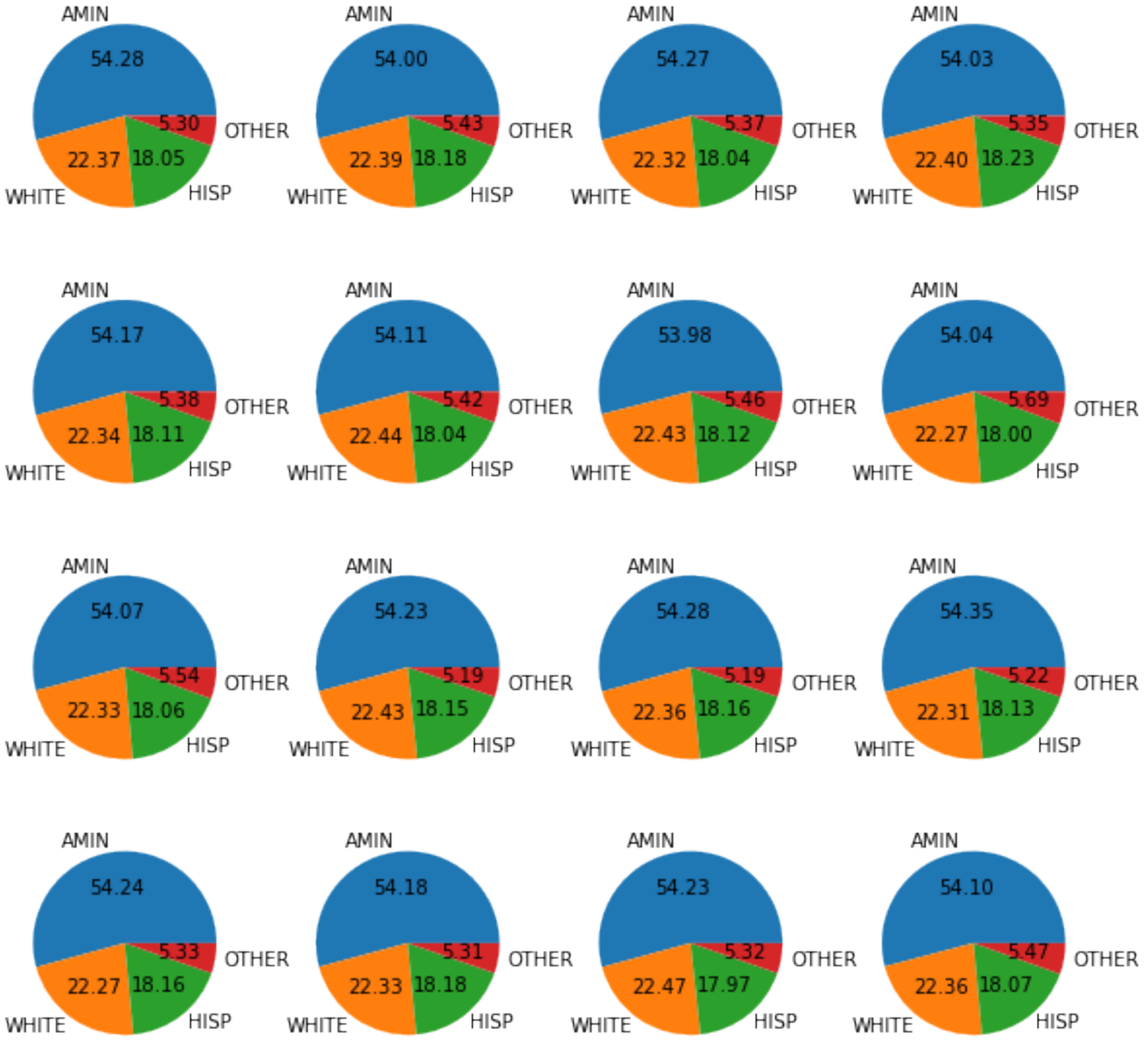


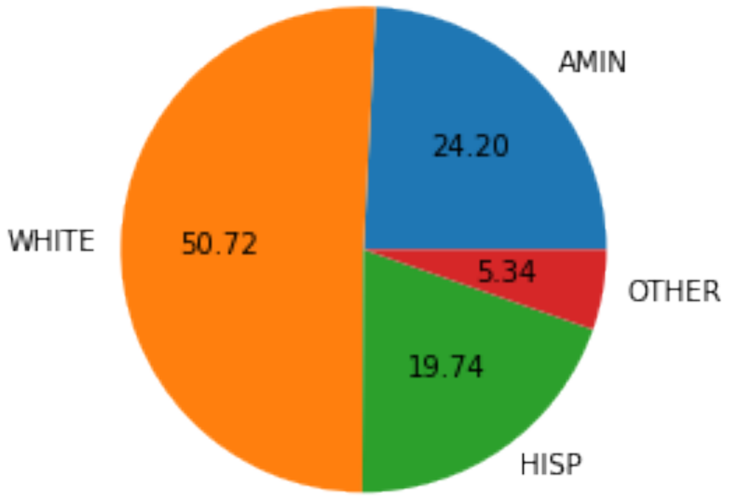
random district #9



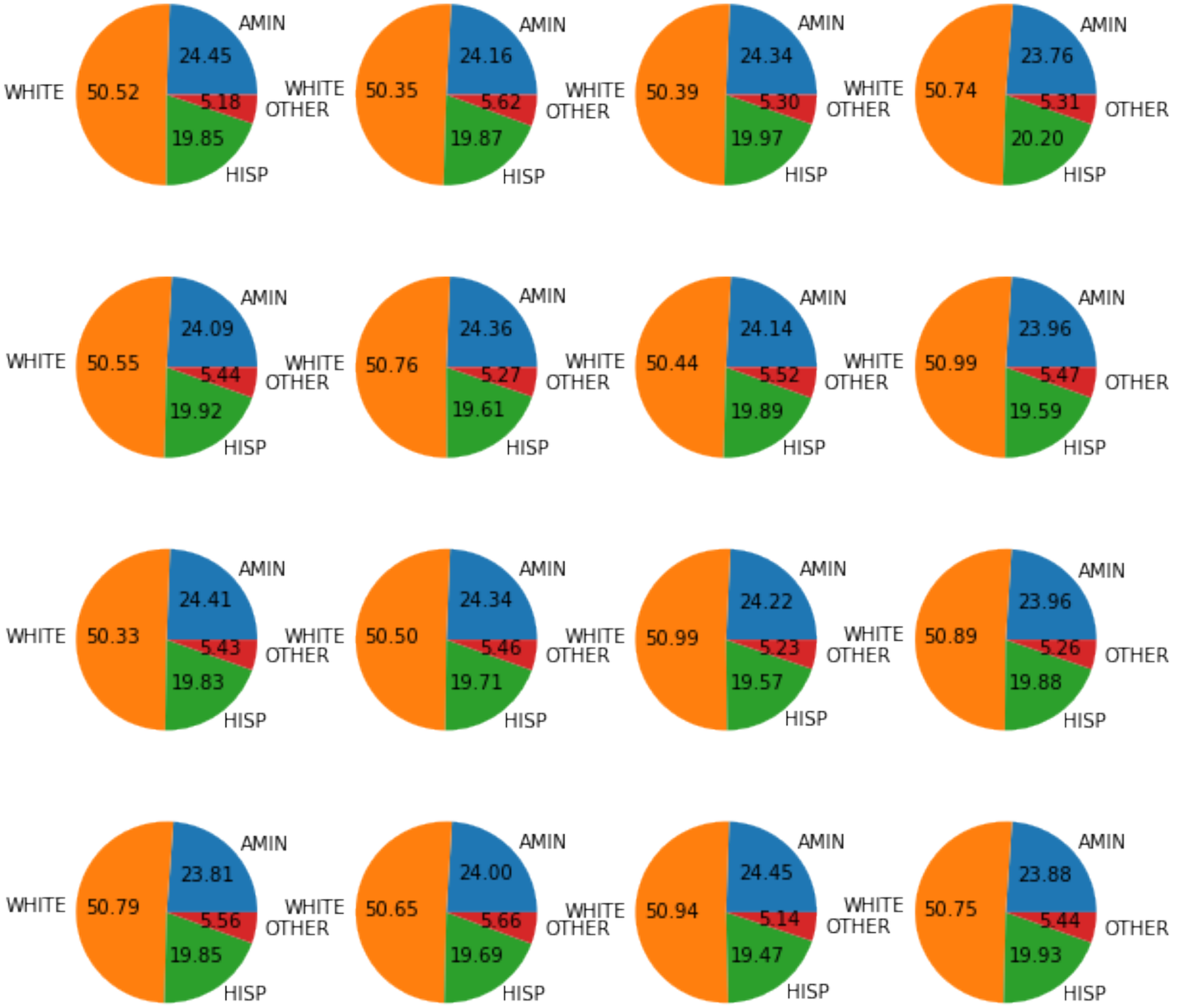


random district #13

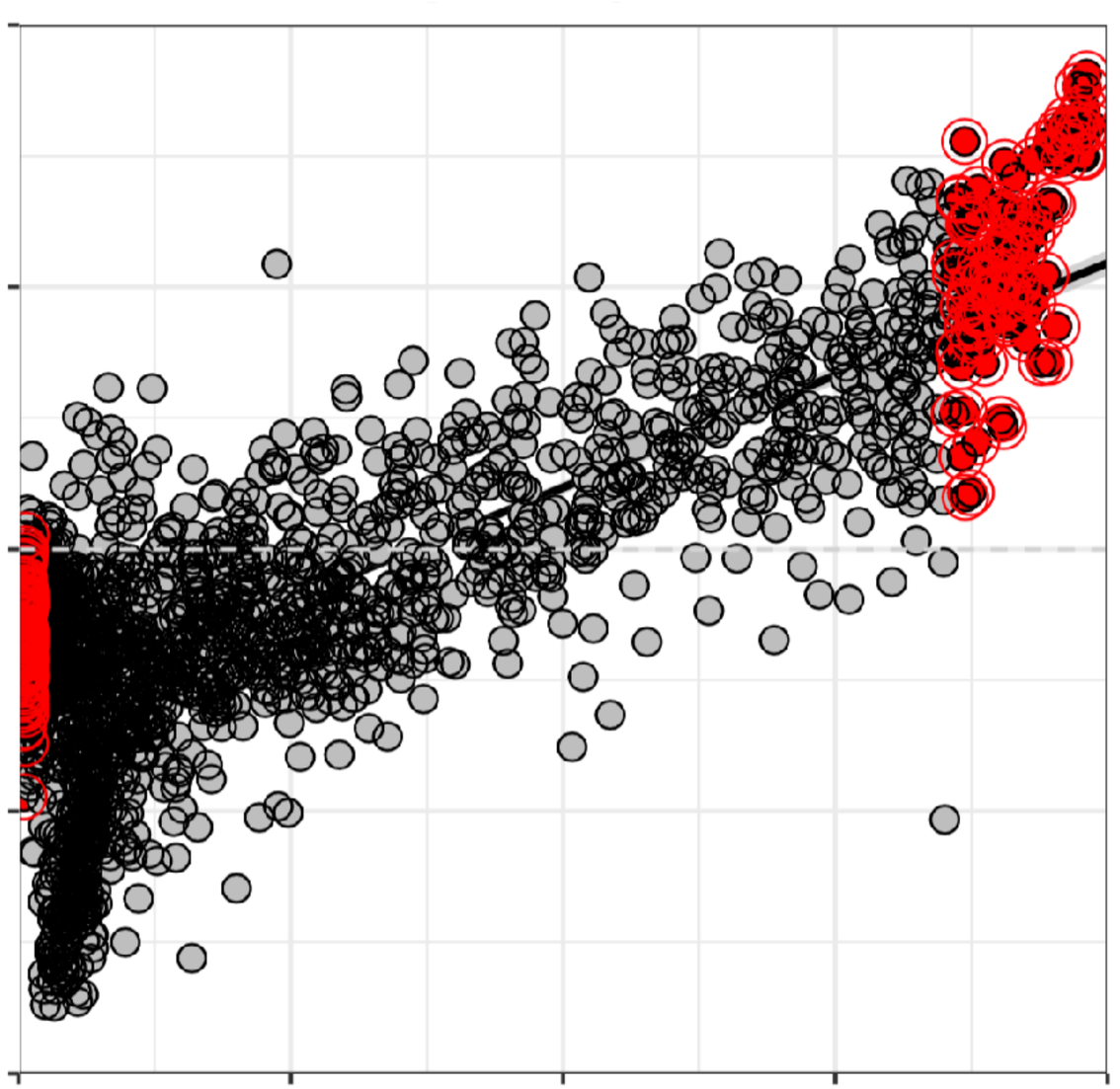




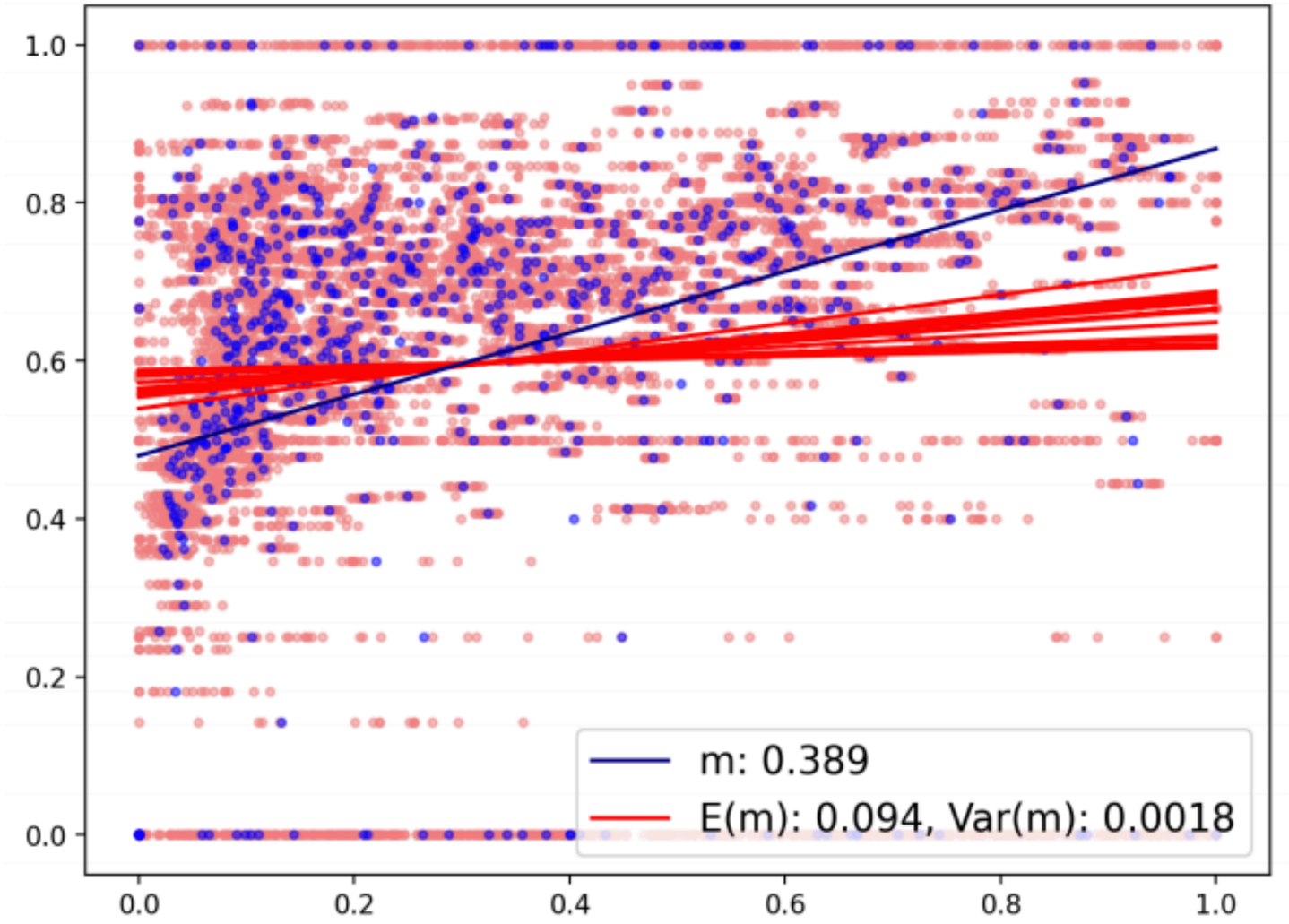
random district #46



Can we identify racially polarized voting?



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



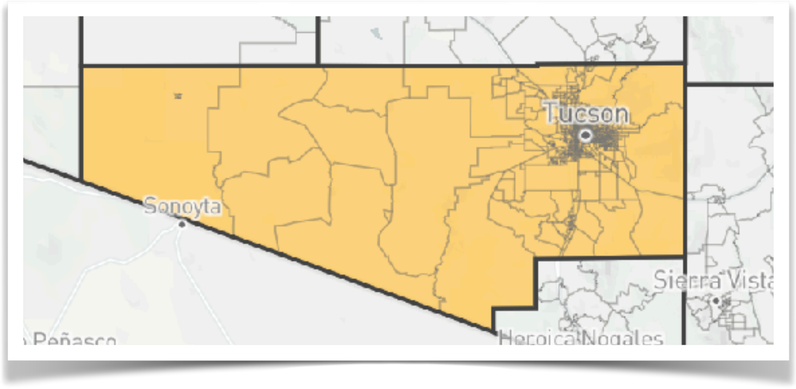
the nightmare scenario

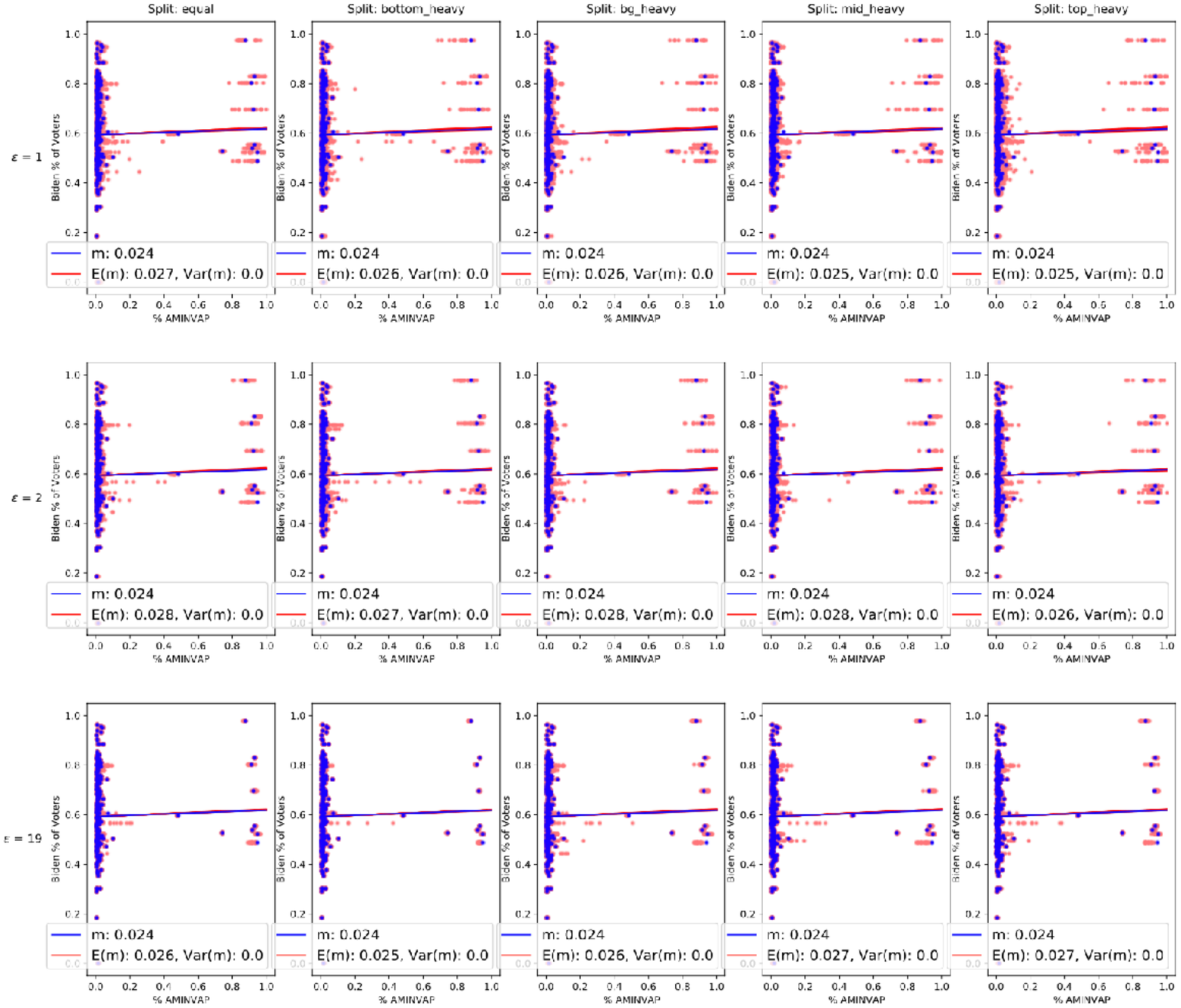
adding noise loses the signal of racially polarized voting

might be unable to test merit of VRA claims

Pima County

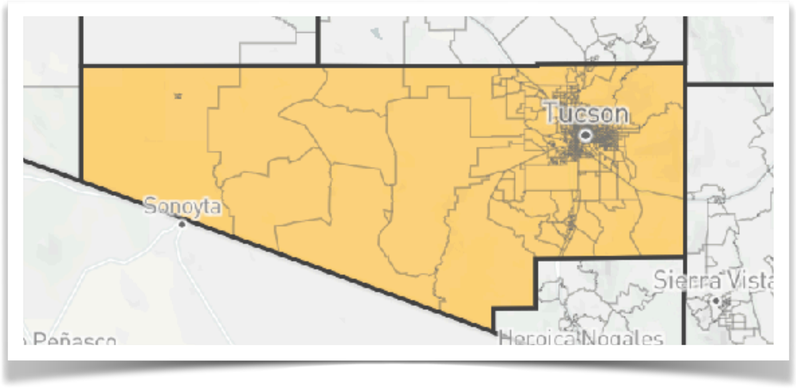
AMIN support for Biden

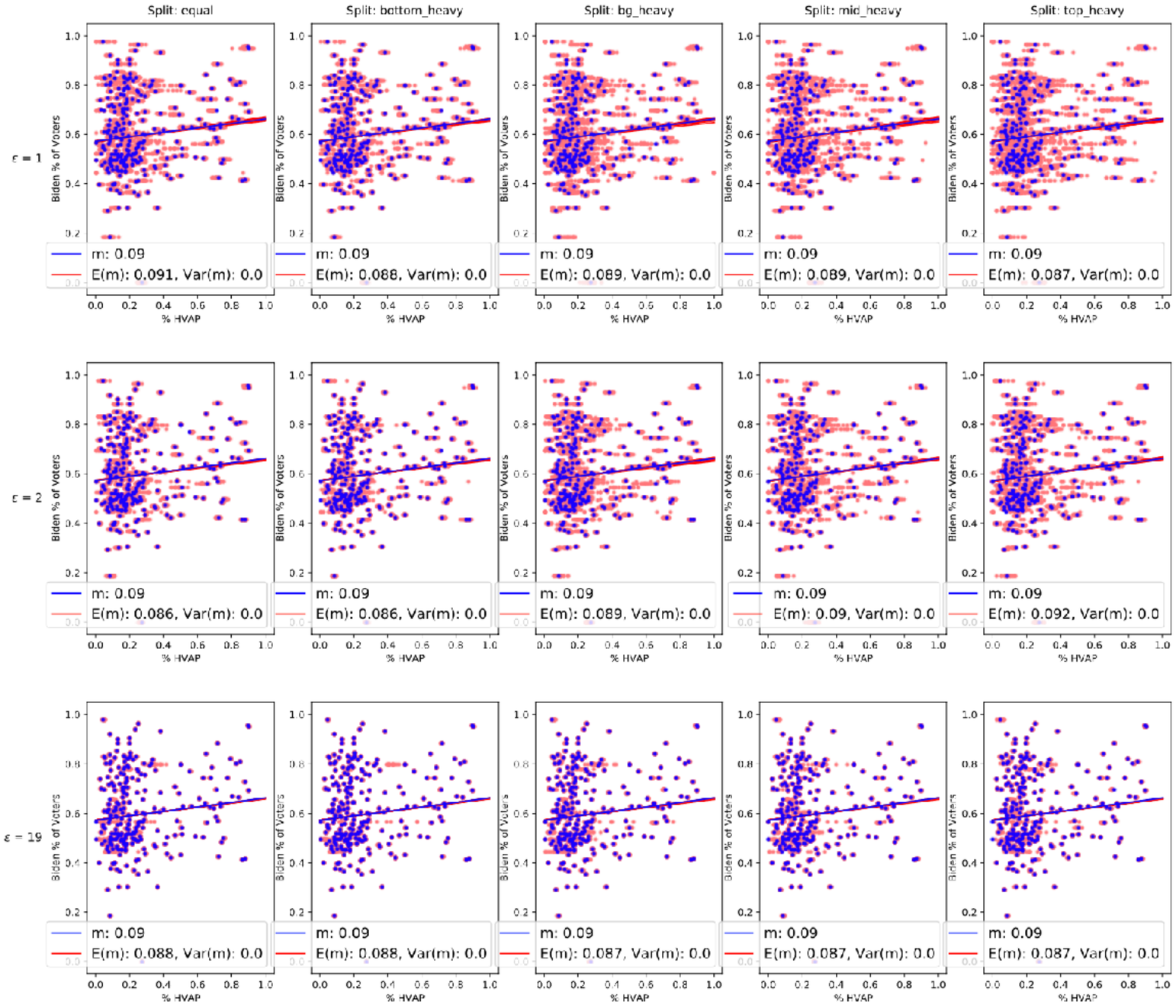




Pima County

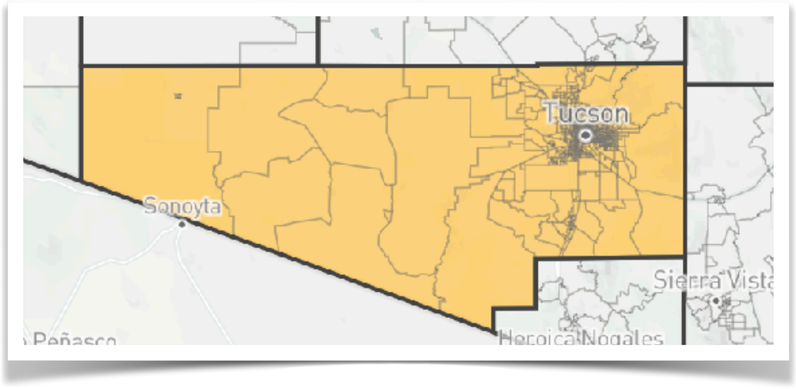
HISP support for Biden

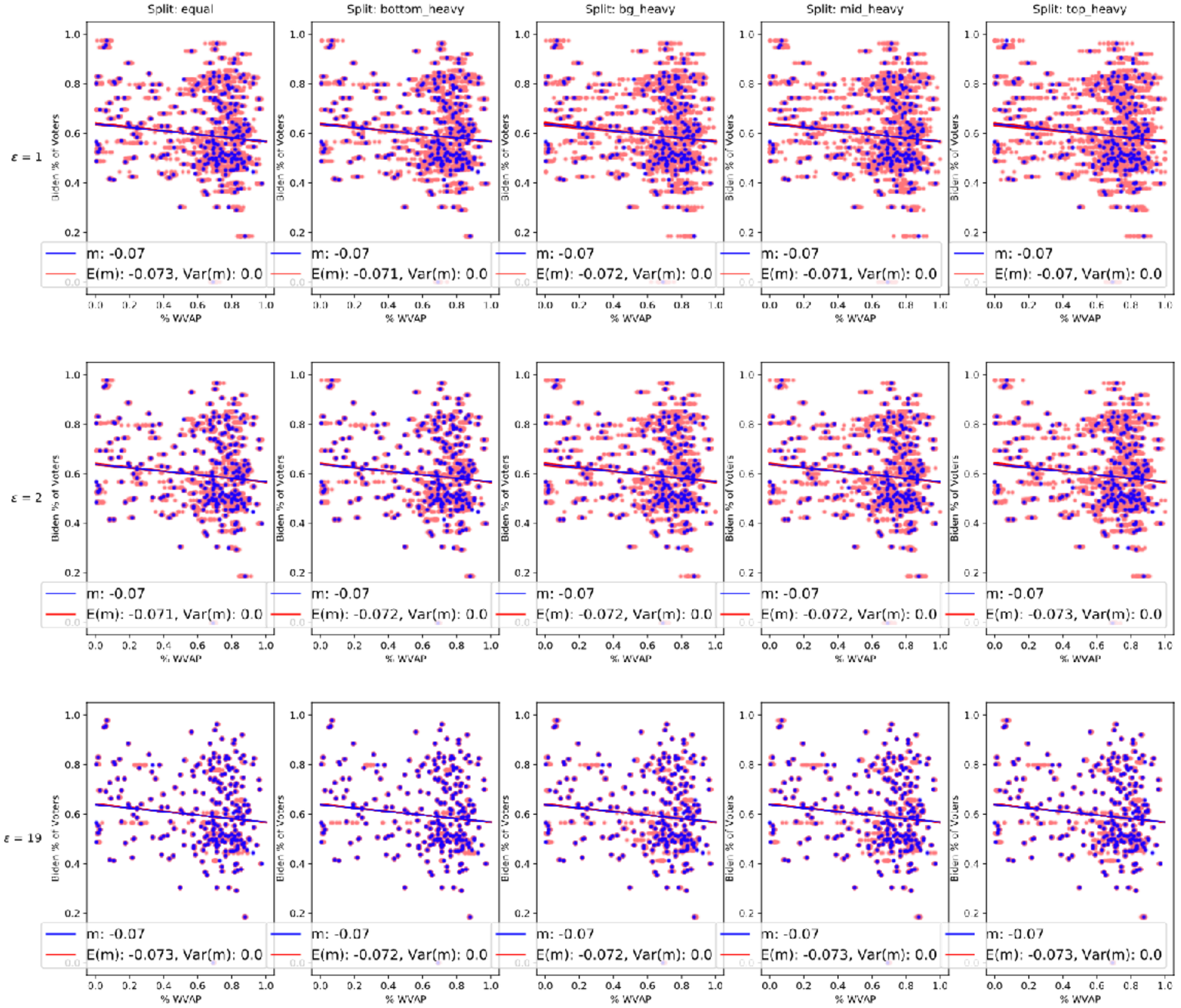




Pima County

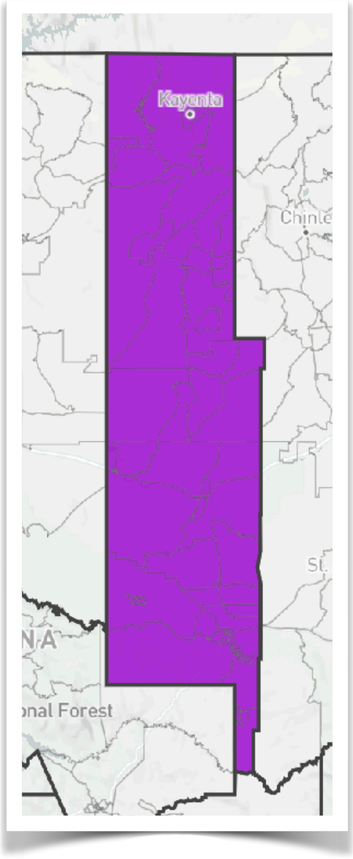
W support for Biden

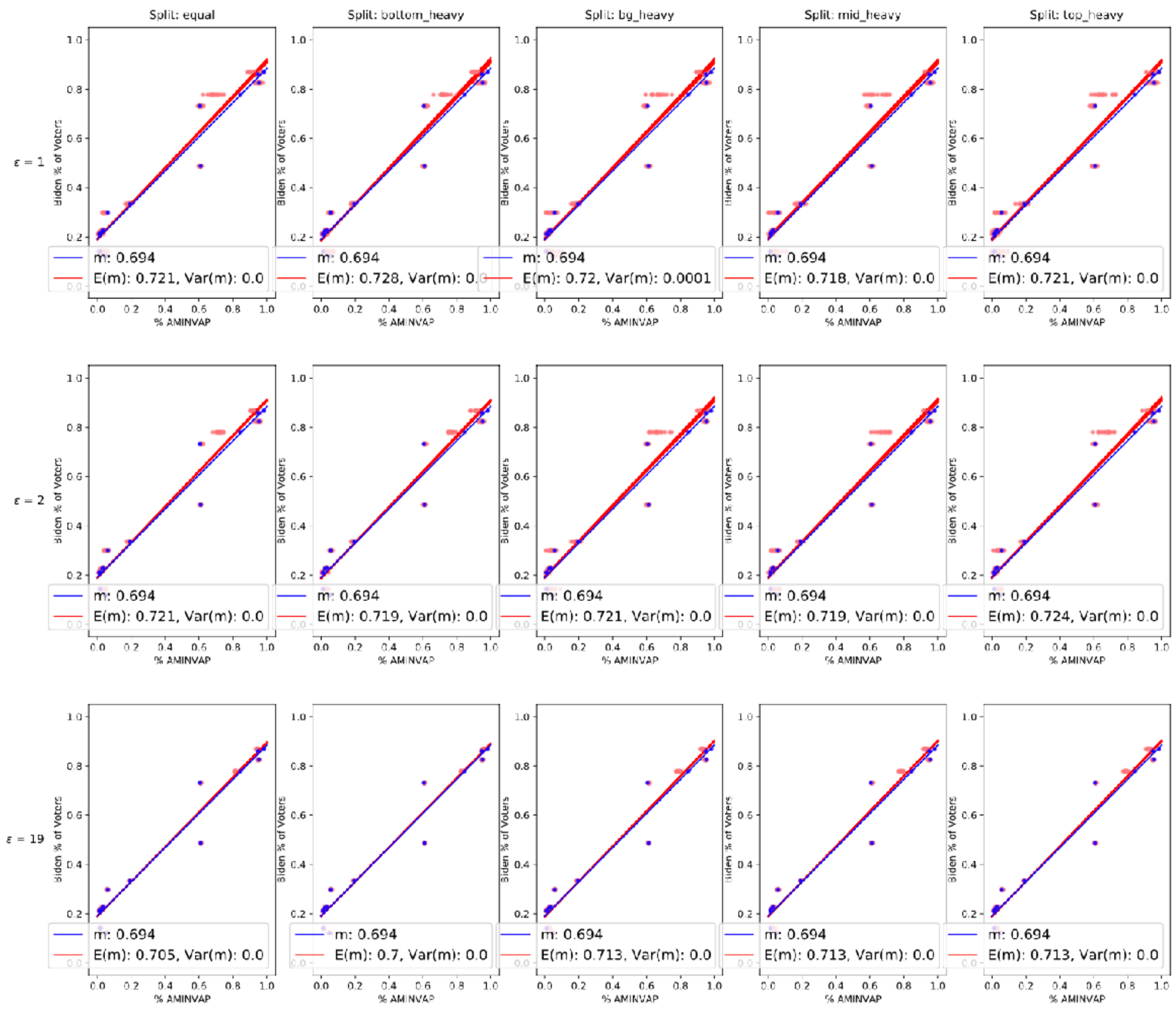




Navajo County

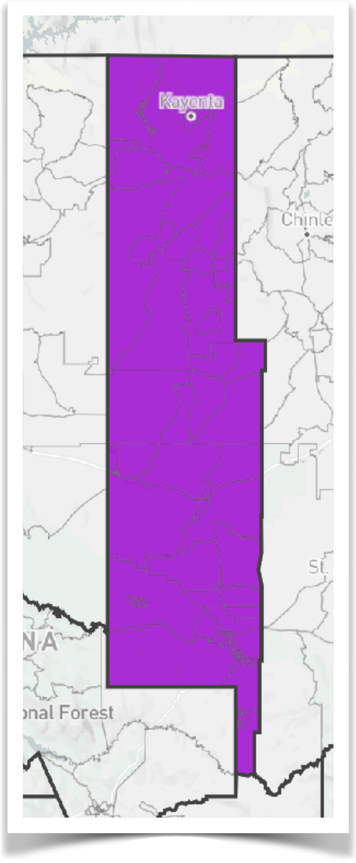
AMIN support for Biden

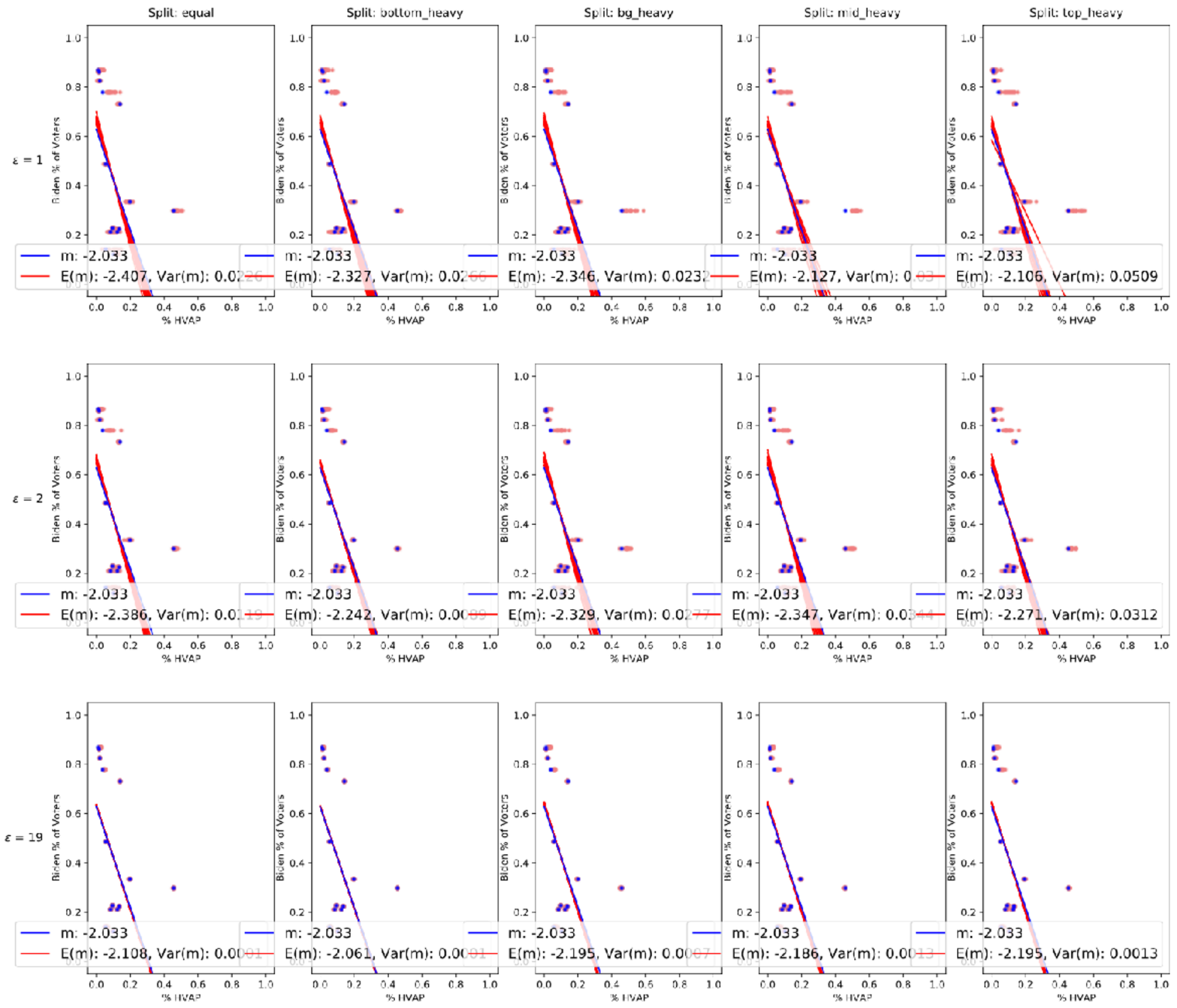




Navajo County

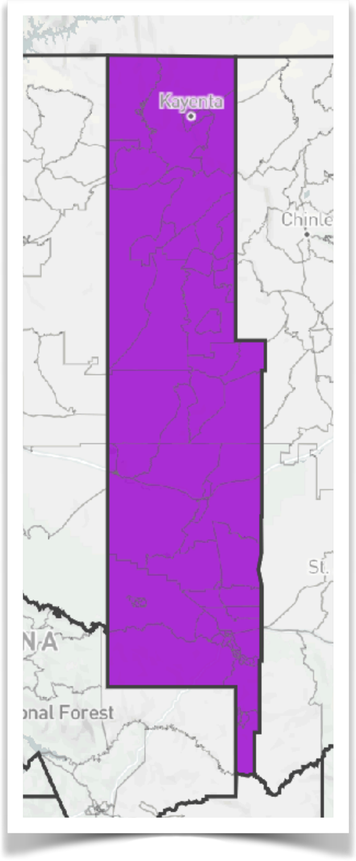
HISP support for Biden

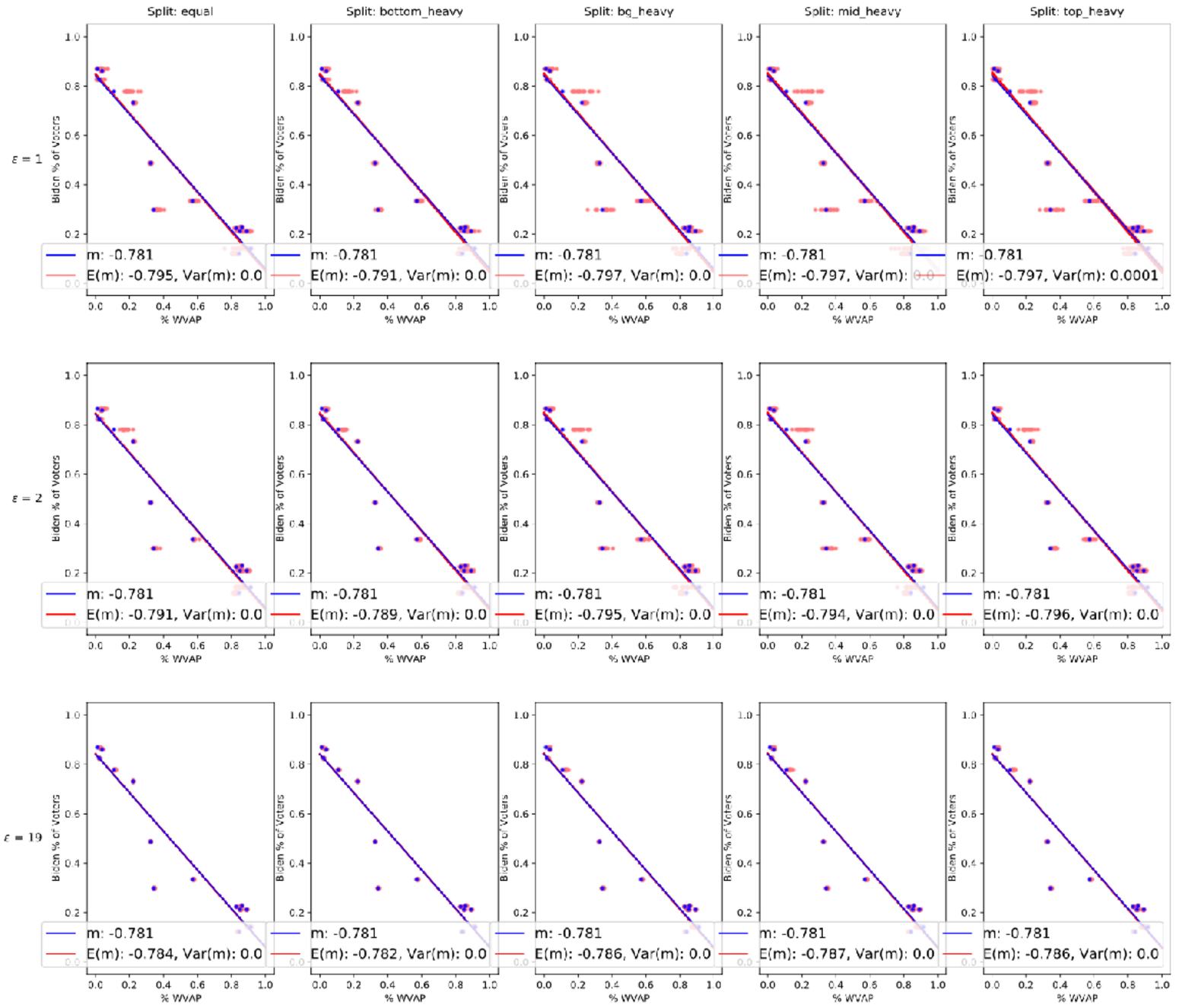




Navajo County

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!

We ran the actual Census Bureau code

Our team ran Census TopDown on our own server (AWS gets costly): Repeated executions, varying epsilon, varying hierarchical allocation

First, run on 1940s data

1013	0.85,0.05,0.05,0.05	53624.29	0.24	0.24	0.2499	2020-02-04T15:1	Rerun of 1012
1014	0.85,0.05,0.05,0.05	53,237.91	0.24	0.24	0.2499	2020-02-05T13:(	Rerun of 1013 to see if there's any
1015	0.97,0.01,0.01,0.01	52,622.03	0.55	0.55	0.2499	2020-02-06T13:3	Really move the privacy budget ar
1016	0.01,0.01,0.01,0.97	51,889.44	0.06	0.06	0.2499	2020-02-07T14:(	Now push it all to the smallest leve
1017	0.01,0.01,0.01,0.97	51,171.57	0.06	0.06	0.2499	2020-02-09T18:2	Repeat to test values.
1018	0.1,0.4,0.4,0.1		0.15	0.15	0.2499	.log	Data recovered from certificate
1019	0.01,0.49,0.49,0.01	59,833.49	0.55	0.55	0.2499	2020-02-10T21:1	Started up new t2.large server. A l
1020	0.1,0.7,0.1,0.1	16,965.16	0.15	0.15	0.2499	2020-02-10T21:3	Now let's try a t2.xlarge. Is the RA
1021	0.1,0.1,0.7,0.1	17,037.97	0.15	0.15	0.2499	2020-02-11T11:3	Let's reuse the t2.xlarge. The RAM
1022	0.01,0.01,0.49,0.49	61,339.34	0.04	0.04	0.2499	2020-02-11T14:28:12-1440	
1023	0.1,0.1,0.4,0.4	52,593.95	0.05	0.05	0.2499	2.log	
1024	0.4,0.4,0.1,0.1	16,909.09	0.15	0.15	0.2499	2020-02-11T18:5	The t2.xlarge really cranks. I wond
1025	0.25,0.25,0.25,0.25	42,190.10	0.07	0.07	0.2499	2020-02-13T19:5	This one doesn't seem to be as fa
1026	0.15,0.15,0.35,0.35	37,115.80	0.05	0.05	0.2499	2020-02-14T12:37:32-2598	

Then, run on reconstructed 2010 TX data

Note: we were the only team outside the Bureau to figure out how to do this! Later, the government of Singapore asked for our help.

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