

Splitting scores at VRDI 2019

taissa.gladkova + splittings group 2019

Contents

1	Why shouldn't districts split localities?	1
2	Splitting scores	2
2.1	Splits	2
2.2	Pieces	3
2.3	Parts	3
2.4	Shannon Entropy	4
2.5	Symmetric square-root entropy	4
2.6	Power entropy	4
2.7	Naked boundary	5
2.8	Pennsylvania Fouls	5
3	Relationships between scores	6
3.1	A note on ensembles	6
3.2	Human-drawn plans in Pennsylvania	6
3.3	Preliminary short burst ensembles	6
4	(Tentative) recommendations	7

1 Why shouldn't districts split localities?

The reasons why one may want districts to keep localities whole are varied. The relevance of each of these motivations may vary from state to state – for example, a state where counties play a large role in governing may consider the “nested government” argument to be stronger. We list here a non-exhaustive list of reasons which one may want to consider. We focus on the case of counties as that was the primary motivating example of a locality considered during the project’s run.

- *existing law* – nineteen out of fifty states require or advise for their congressional districts to follow political boundaries. Of the twelve that get into any specifics, all twelve specifically mention counties.
- *limiting gerrymandering* – it is possible that restricting the ability of districts to split up larger units such as counties may limit the ability of map-drawers to gerrymander. There is some evidence based on ensembles for this in the *Virginia criteria paper*. As further informal evidence for this fact, consider the following map of NC. The underlying choropleth shows the vote share from the 2016 presidential election and the county borders are in black. In yellow, we indicate the district boundaries – the districts are current as of 2019 and were part of the gerrymandering case *Ruch vs Common Cause*. One may observe that where the districts deviate from county boundaries, it is

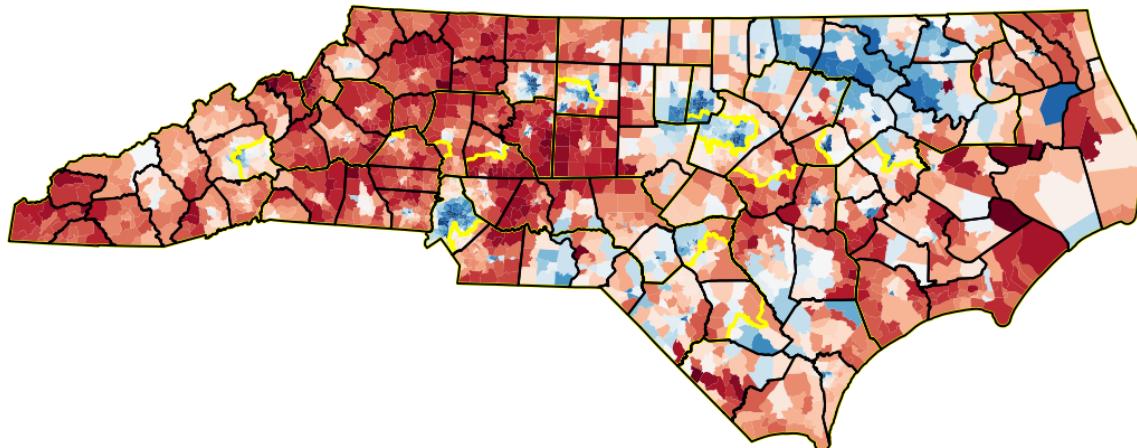


Figure 1. North Carolina enacted districts deviation from county boundaries

often seemingly to either crack or pack a blue region of precincts – one may even be tempted to say that “the gerrymandering happens in the interior of the counties”. We stress that this is just a preliminary analysis, and more study is required.

- *communities of interest* – in some cases, counties represent a community which should be kept whole for the purposes of representing the interests of that community.
- *election administration* – to the extent that counties are involved in administering elections, fewer districts means fewer ballots to manage.
- *nested government* – a general principle of governance is that the various jurisdictions at each level of government be nested.

2 Splitting scores

It is not always clear what should be meant by “splitting counties less”. Since it is not always feasible to find even one map that never splits a county (some counties are for example far more populous than any district), trade-offs must be made. In order to compare different plans against one another, one can come up with a score which reflects the extent to which county splitting occurs in the plan. The main body of work in the VRDI project was to propose and analyse a number of different such scores. Here we present some of the more promising scores considered. Again, we consider the case of counties, though other localities can also be considered. In each case, we list what we consider the advantages and disadvantages of the score, as well as any other properties the score has whose value is unclear.

2.1 Splits

Definition: The number of counties which intersect more than one district.

Advantages: Perhaps the simplest score, it is easy to explain and to capture in law.

Disadvantages: Is insensitive to geography and splitting counties into more than two, and does not take into account the nature of the split.

Other: All counties treated equally, regardless of population.

2.2 Pieces

Definition: For every county C and every district D , let $n(C, D)$ be the number of connected components of the intersection $C \cap D$. Then

$$\text{pieces} = \sum_{C,D} n(C, D).$$

Informally, lay the county and district borders on top of one another and count the number of resulting regions.

Advantages: A fairly simple score to explain, and is sensitive to additional splittings beyond the first. The intersection of a county and district containing more than one connected component often fails the “eyeball” test.

Disadvantages: Is insensitive to geography, and does not take into account the nature of the split.

Other: All counties treated equally, regardless of population. Is inherently symmetric in counties and districts.

2.3 Parts

Definition: For every county C , let $s(C)$ be the number of districts intersecting C . Then

$$\text{parts} = \sum_C s(C).$$

Advantages: A fairly simple score to explain, and is sensitive to additional splittings beyond the first.

Disadvantages: Is insensitive to geography, and does not take into account the nature of the split.

Other: All counties treated equally, regardless of population. Symmetric, as the following lemma shows.

Lemma 1. *For every district, let $s(D)$ be the number of counties intersecting D . Then*

$$\sum_{C \in \text{counties}} s(C) = \sum_{D \in \text{districts}} s(D)$$

Proof. Both sides are merely the number of non-empty sets in the collection

$$\{C \cap D \mid C \in \text{counties}, D \in \text{districts}\}$$

□

2.4 Shannon Entropy

Definition: An information-theoretical concept which asks: on average, how many additional bits of information are required to specify a person's district once the county is known? It is defined by the formula:

$$\text{ShEnt} = \sum_{C \in \text{counties}} \frac{\text{pop}(C)}{T} \sum_{D \in \text{districts}} -\frac{\text{pop}(C \cap D)}{C} \log \frac{\text{pop}(C \cap D)}{C}$$

where T is the total population of the state. Note: another variant of this formula considers *number of VTDs* instead of population.

Advantages: Takes into account the nature of the split. In particular, a 50 – 50 split of a county is worse than a 99 – 1 split. This may or may not be considered an advantage depending on user preference. Also takes into account county populations.

Disadvantages: Does not consider geography. The score is unitless, and does not make sense to compare across states. It is also not easy to explain to non-experts.

Other : Counties are not treated equally, and are weighted by population.

2.5 Symmetric square-root entropy

Definition: A symmetrization of Shannon entropy, replacing log by square root.

$$\text{SqEnt} = \sum_{C \in \text{counties}} \text{pop}(C) \sum_{D \in \text{districts}} \sqrt{\frac{\text{pop}(C \cap D)}{C}} + \sum_{D \in \text{counties}} \text{pop}(D) \sum_{C \in \text{counties}} \sqrt{\frac{\text{pop}(C \cap D)}{D}}$$

This score was recommended in a footnote of the *Virginia criteria paper*.

Advantages: Takes into account the nature of the split. In particular, a 50 – 50 split of a county is worse than a 99 – 1 split. This may or may not be considered an advantage depending on user preference. Also takes into account county populations. The symmetry may be appealing in situations where counties differ widely in population relative to an ideal district size.

Disadvantages: Does not consider geography. The score is unitless, and does not make sense to compare across states. It is also not easy to explain to non-experts.

Other : Counties are not treated equally, and are weighted by population.

2.6 Power entropy

Definition: This score is based on Shannon entropy, but adapted to penalize small splits more, and to weight low population counties more than high population ones (this apparently inspired by Iowa's recommendation that “[w]hen there is a choice between dividing local political subdivisions, the more populous subdivisions shall be divided before the less populous”). It is given by the formula

$$\text{PEnt} = \sum_{C \in \text{counties}} \frac{T}{\text{pop}(C)} \left[\left(\sum_{D \in \text{districts}} \left(\frac{\text{pop}(C \cap D)}{\text{pop}(C)} \right)^{1-\alpha} \right) - 1 \right]$$

where α is some parameter between 0 and 1. We refer the reader to the VRDI 2018 report *Entropy Functions for Analysis of Splittings in Redistricting* for full motivation of this score, where $\alpha = 0.8$ is recommended. Note: as with Shannon entropy, there is a variant using number of VTDs instead of population.

Advantages: Takes into account the nature of the split. In particular, a 50 – 50 split of a county is worse than a 99 – 1 split. This may or may not be considered an advantage depending on user preference. It penalizes a 99 – 1 split and additional small splits more heavily than Shannon entropy does. Also takes into account county populations, though in the opposite sense of Shannon entropy.

Disadvantages: Does not consider geography. The score is unitless, and does not make sense to compare across states. It is also not easy to explain to non-experts. Compared with the Shannon version, it also lacks a historical underpinning from information theory.

Other : Counties are not treated equally, and are *inverse*-weighted by population.

2.7 Naked boundary

Definition: The length of district boundary which is not coincident with a county boundary. This “length” may be computed either

- exactly (in miles), or
- approximated by the number of cut edges between districts whose ends lie in the same county.

Advantages: Considers geography, and penalizes “weird” intra-county district boundaries which may fail the eyeball test. Simple to explain and in natural units when computed in miles. Appears to line up well with legal language surrounding “political boundaries”.

Disadvantages: Conflated with compactness in that more compact plans may naturally score lower. Both the miles and cut edges versions have measurement issues (miles may suffer from the coastline effect, while cut edges is sensitive to distribution of units e.g. VTDs). Does not take county population into account. Does not penalize small “nibbles” very much.

Other: Not symmetric.

Note: An early version of this score was called “coincident boundary”. While the terminology is obsolete, this name appears in some plots below where it should be understood to mean this score.

2.8 Pennsylvania Fouls

Definition: Based on legal language from Pennsylvania’s redistricting legislation, a county C is considered to a “Pennsylvania foul” if it intersects more than

$$P(C) = \lceil \text{pop}(C)/I \rceil + 1$$

districts, where I is the ideal population of a district. The score is then the total number of Pennsylvania fouls.

Advantages: Depends on the state and county layout, and so attempts to be comparable across states. Is not as complicated or opaque as entropy-based scores, but can be ambiguous unless states precisely.

Disadvantages: Is insensitive to geography and splitting counties into more than a certain threshold, and does not take into account the nature of the split. No penalty for splitting every county into two pieces.

3 Relationships between scores

3.1 A note on ensembles

While neutral ensembles of plans generated from a Markov chain can be effective tools in determining a baseline for various metrics, these ensembles typically split up counties far more than human-made plans, since they are usually blind to these boundaries. Moreover, it is challenging even to bias a MCMC run towards less county splittings, since neither ReCom nor flip are particularly well-suited to proposing plans which split counties less. As a result, care must be taken when using ensemble methods to compare or evaluate these scores.

3.2 Human-drawn plans in Pennsylvania

In order to get some feeling for the scores and how they compare on human-drawn plans, we consider eight plans available in the shapefile for Pennsylvania on github.com/mggg-states. The figures below show the various plans, with districts indicated by colored regions and county boundaries displaying in black.

The bar charts below indicate the performance of these eight plans on some of the scores considered above. Note that the VTD variants (not the population variants) of Shannon entropy and Symmetric square-root entropy were used.

Finally, one can compare these scores on the eight plans with scatter plots. While an exhaustive pairwise comparison is possible, for now we restrict to comparing them all against one score (Pieces), which is very simple but also more sensitive than Splits.

3.3 Preliminary short burst ensembles

We consider using short bursts to optimize for one of these scores and compare how the other scores change. We start these runs from the the remedial PA districting plan and use 500 bursts of 10 steps. We minimize the naked boundary score and compare the other scores on these new districting plans.

The remedial PA districting plan has 596 cut edges that are not coincident with a county line and this short burst minimization algorithm decreases the number of cut edges to about 460. We find

that the naked boundary score is anti-correlated with the number of pieces. In the scatterplot, we have only included the minimum naked boundary districting plan for each short burst.

We also find that the naked boundary score is anti-correlated with power entropy scores (such as square root entropy and symmetric entropy). However, we find very little correlation with log entropy scores such as Shannon entropy. We can also compare the districting plans of the remedial PA plan and the plan found after 500 short bursts. We find that these plans are quite similar. The minimum naked coincident boundary plan has more splits and pieces than the remedial plan because it takes small nibbles out of counties that add only a few cut edges, but increase the splits and pieces metrics significantly.

A lot more work is needed to design MCMC runs which fully explore plans with similar county splitting properties to human-made plans, to determine if (in this range) these scores are correlated and to what degree. The question of optimizing these scores effectively with MCMC is also the subject of on-going research.

4 (Tentative) recommendations

The scores seem to naturally group into three types:

- “Simple counts” – scores which are easy to understand and compute such as Splits, Parts and Pieces.
- “Population-based methods” – the main characteristic of these scores is they care how the population of a county is divided up (e.g. entropy methods).
- “Geography-based methods” – only Naked Boundary represents this class in this report, the main characteristic being that this score cares how the *geography* of a county is divided up.

One can find good reasons to be interested in scores from each of these classes. One should take into account, however, that the simpler scores are easier to codify in law and explain to non-experts. We may tentatively want to recommend that simple counts form the bedrock of questions about county splittings, and thereafter the stakeholders determine for themselves which additional, finer aspects of the problem they care about and whether they are willing to deploy more nuanced and complicated scores.

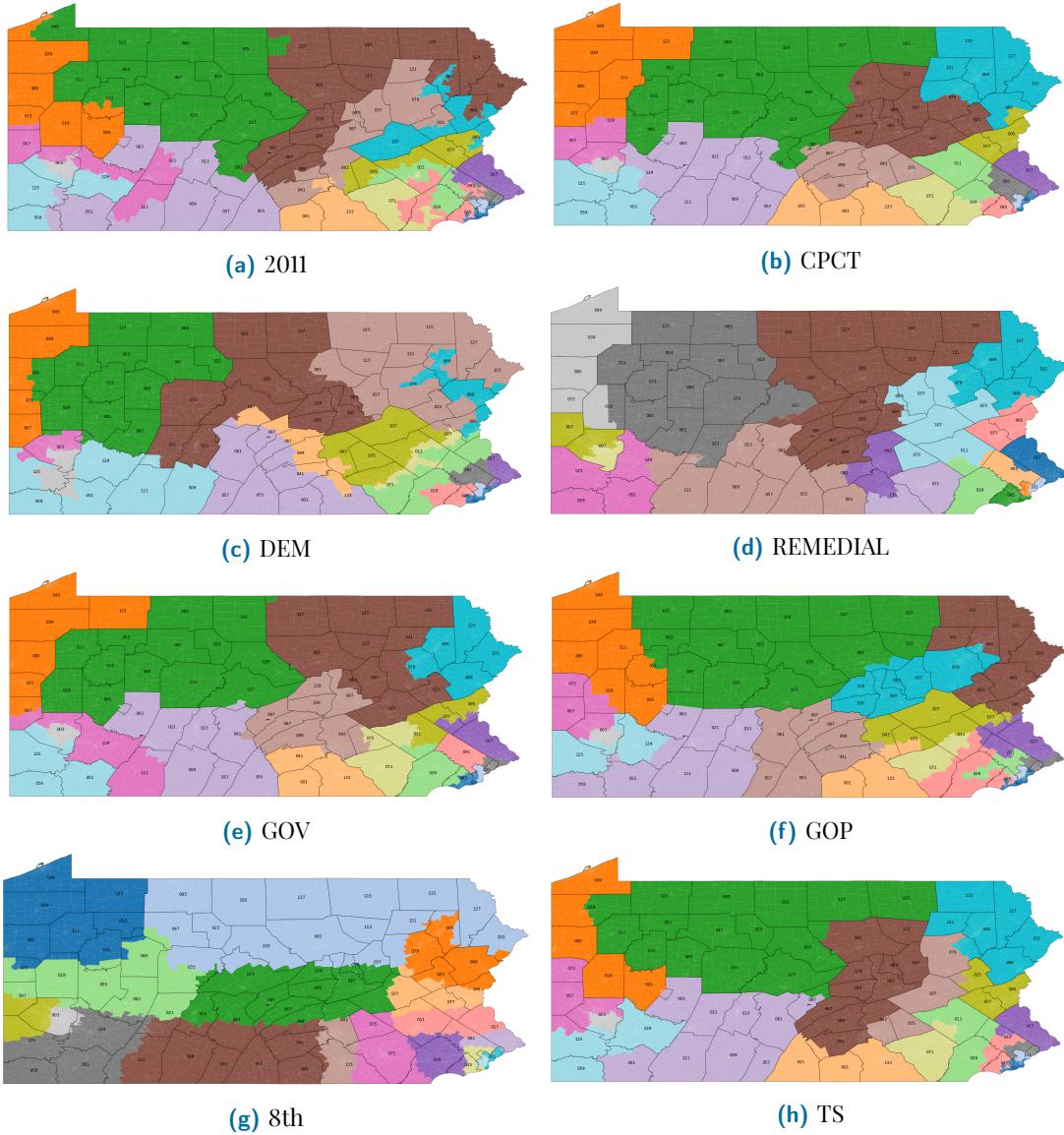
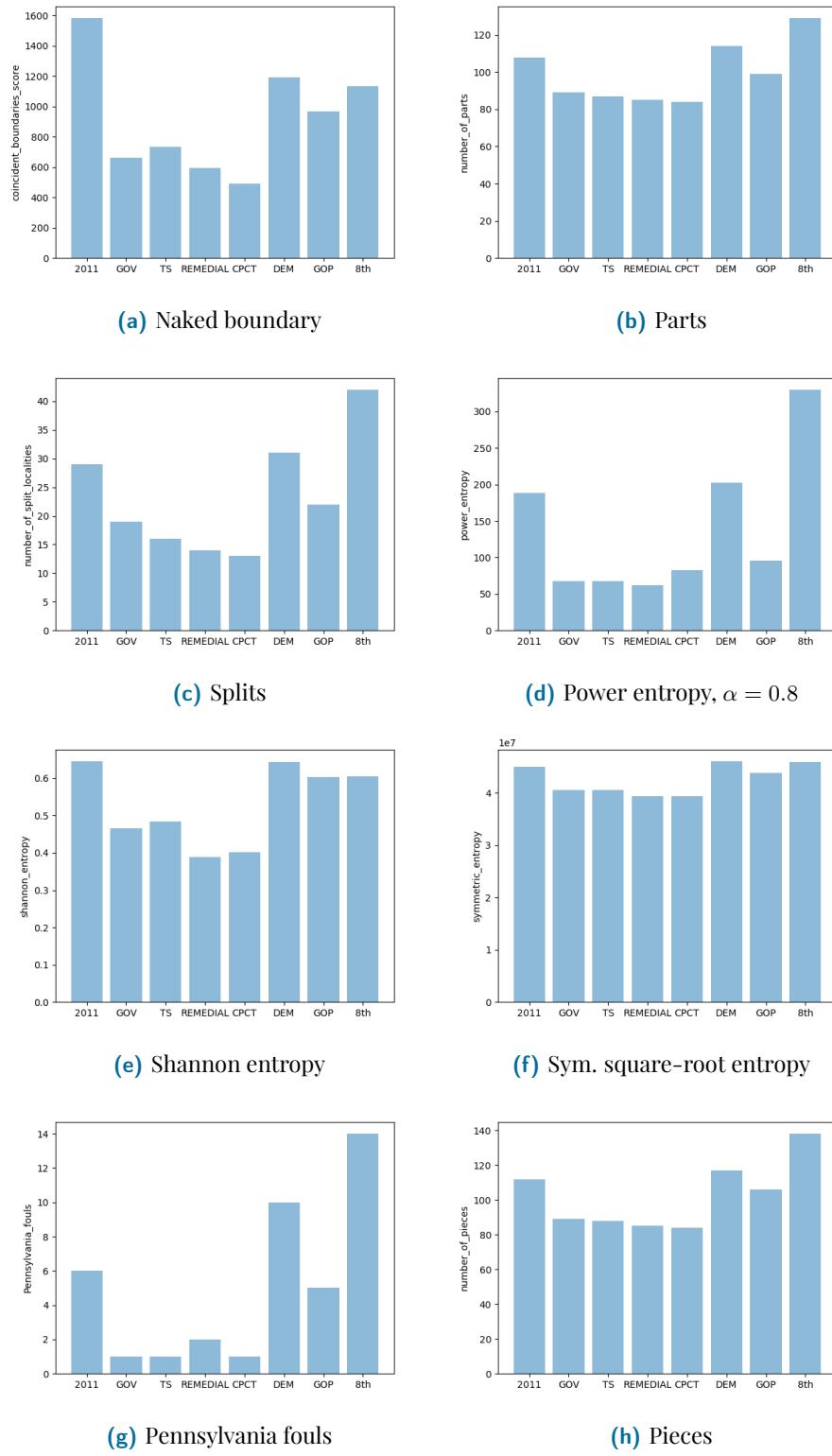
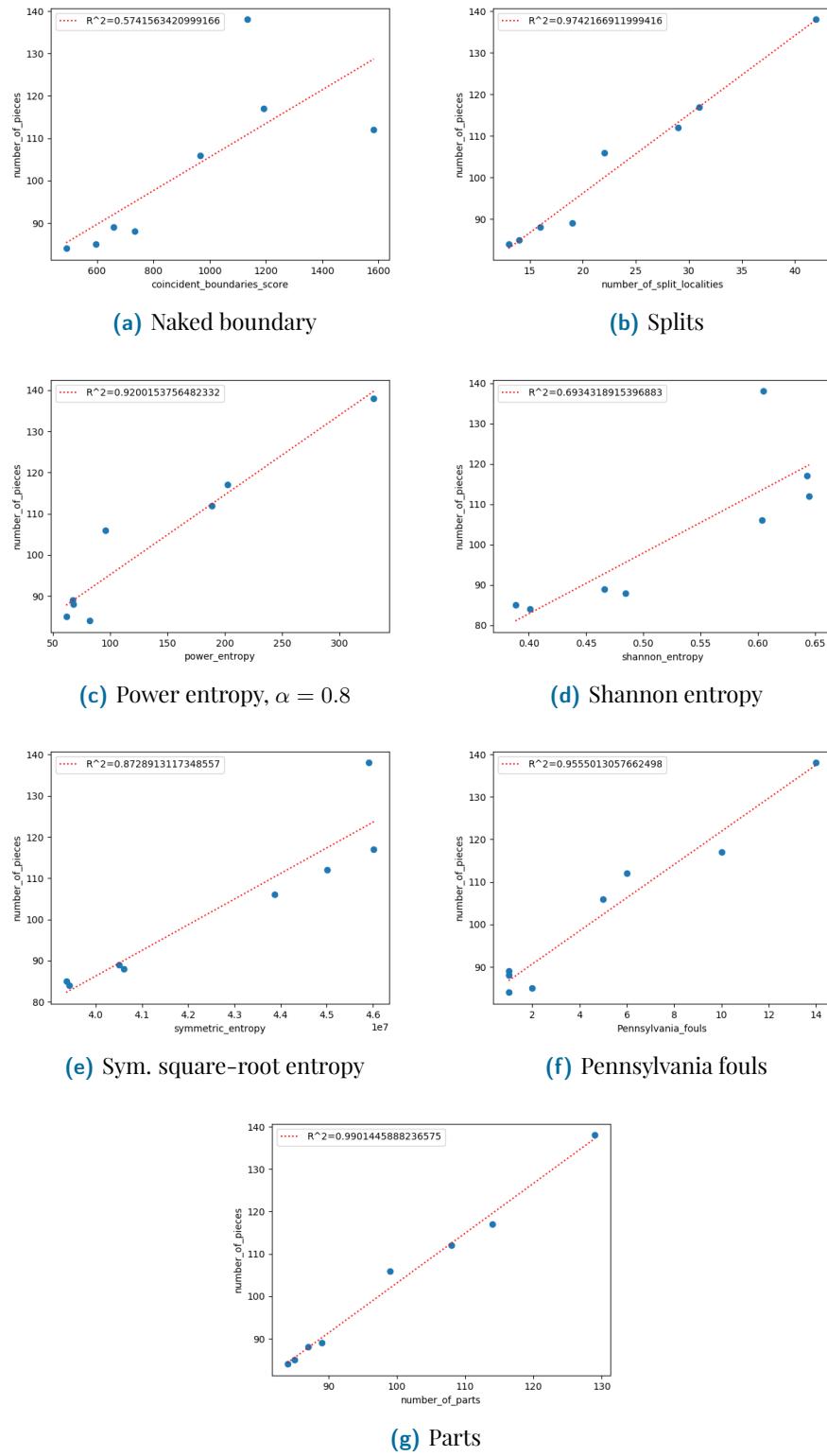


Figure 2. Human-drawn plans on PA over county boundaries

**Figure 3.** Scores for human-drawn plans on PA

**Figure 4.** Scatter plots vs Pieces for human-drawn PA plans

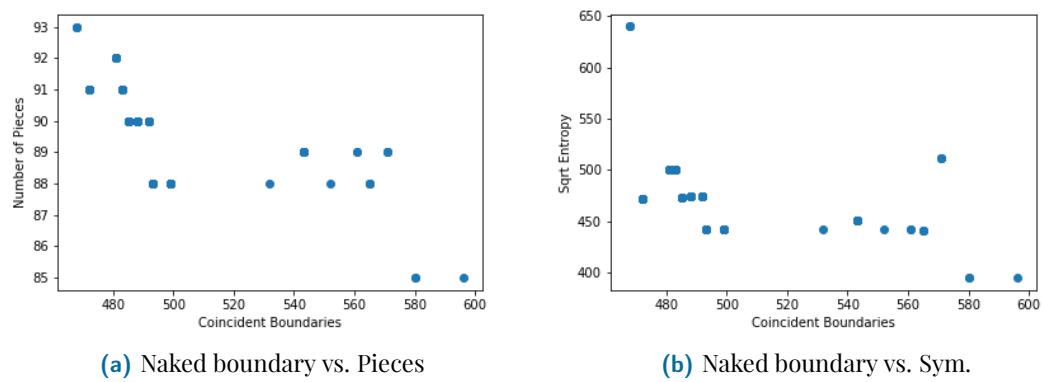
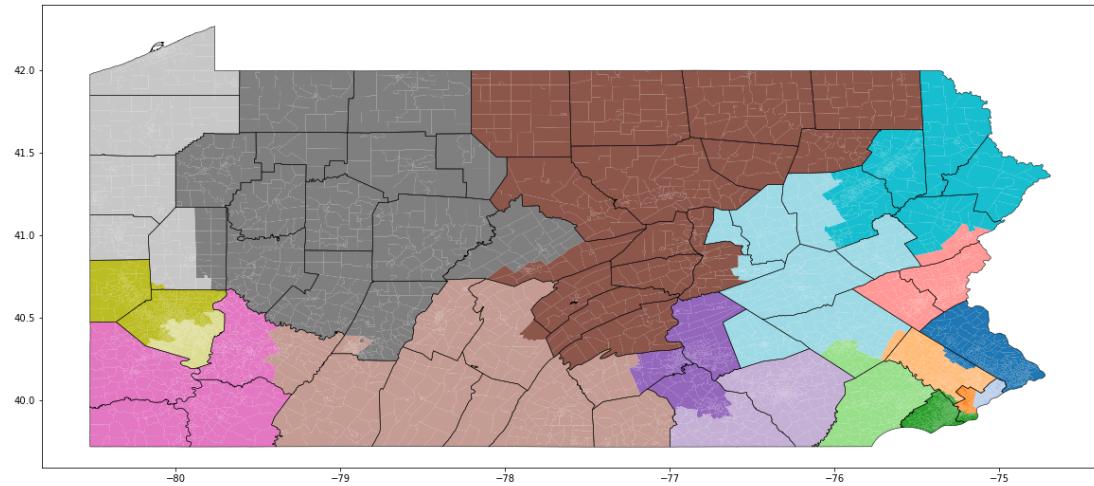
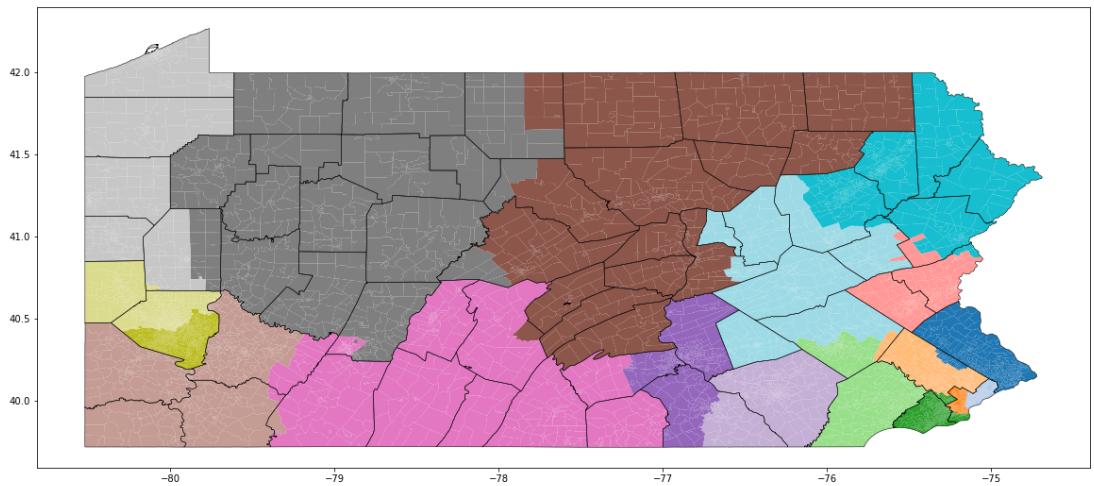


Figure 5. Short burst runs on PA



(a) REMEDIAL plan



(b) Min. naked boundary plan