Understanding the Effect of Gerrymandering on Voter Influence through Shape-based Metrics

Jack Cackler[1] and Luke Bornn[2]

[1]Department of Biostatistics, Harvard University
[2]Department of Statistics, Harvard University

June 21, 2014

1 Introduction

Gerrymandering describes the abnormal partitioning of a political region into smaller regions of equal population. For congressional redistricting, the fifty United States are partitioned into 435 roughly equally sized congressional districts. Based on the 2010 Census (U.S. Census Bureau, 2011), the largest district was the sole Montana district with 994,416 people, while the smallest was the Rhode Island first district with 526,283 people. The mean district size was 710,767 people, and 95 percent of districts were between 640,000 and 780,000 people.

congressional districts can theoretically be partitioned in any way to roughly preserve equal populations – a non-deterministic problem with many approaches to optimization (Adams, 1978). Ideally partitions are made along consistent, intuitive guidelines such as locally compact geographical units. Districts can also be optimized to produce some desired demographic distribution – benefiting certain individuals or populations at a cost to others (Cain, MacDonald, Hui, Boyle, Lee, and Woods, 2006, Cox and Katz, 2002).

We can consider any redistricting scheme in which optimality is non-constant for individuals as not "fair" (Grofman, 1983, Sherstyuk, 1998). Specifically, in a "fair" system, an individual's political influence should be independent of the redistricting procedure. Irregular geographic boundaries are often a side effect of attempts to make districts unfair. Measures of geographic irregularity can hint at human intervention for a desired gain, and thus can be used to estimate the rate at which a district is not "fair" (Cox and Katz, 2002). As an example, the portmanteau gerrymander was coined by the *Boston Gazette* to describe the irregularity of a roughly salamander-shaped State Senate district designed by Governor Elbridge Gerry in Essex, Massachusetts. The bizarrely shaped districts in the 1812 election allowed the Democratic-Republicans to retain control of the state senate, despite a large percentage of state voters voting for Federalist candidates. Though Gerry's salamander district

looked intuitively irregular, numerical measures of irregularity allow us to explicitly quantify the presence gerrymandering, as well as measuring its impact on voter influence.

This paper will address a few major issues related to gerrymandering. First, we discuss the causes and uses of gerrymandering and provide an overview of how US congressional district boundaries are drawn. Second, we analyze two existing scores used to measure gerrymandering and introduce one new score, which has a major advantage of being able to measure regularity independent of state boundaries. Finally, we study each score, modeling their relation to a measure of voter influence.

1.1 Gerrymandering

A primary functions of gerrymandering is often to increase a political party's overall representation in Congress (Grainger, 2010, Owen and Grofman, 1988). Given infinite knowledge of how people will vote, almost any overall representation in Congress is possible. Figure 1 shows a dummy example of 49 individuals colored white and black to indicate political preference. In this example there are 25 white points and 24 black points. Suppose you wanted to split these individuals into 7 subregions of 7 points each. Two division procedures would be to split the data into rows or split the data into columns. Surprisingly, splitting the data into rows yields 6 majority black rows (each with a 57% majority), and one all white row, while splitting the data into columns yields 6 majority white columns (with a 57% majority), and one 86% black columns. By concentrating opposing voters into a small number of districts, a party is able to predictably secure a majority of the remaining districts, regardless of the actual proportions in the region.

Gerrymandering can also be used to promote incumbency (Gul and Pesendorfer, 2010). Particularly in states where legislators control the redistricting process incumbents are heavily incentivized to make their own districts secure (Ansolabehere, Snyder, and Stewart, 2000). This can manifest in an opposite incentive to the first motivation; while the Republican Party Chairman in Tennessee might want each Republican Congressman to win by a slim majority to maximize representation, those Congressmen want to win each of their seats by a large majority so that they can ensure their seat for years to come. Either case of gerrymandering will create congressional districts with supermajorities, which can be measured numerically.

There are positive reasons for optimizing over non-geographic constraints, the most typical cited being minority representation (Coate and Knight, 2007, Washington, 2006). In the current 113^{th} Congress, 42 out of 435 members are black, compared to 1 out of 100 senators, which are elected by states as a whole. The representation of black representatives in the House is much greater than in the Senate, but is still lower than the national average of 12.5 percent. Concentrating black voters into a handful of districts increases the chances for a black representative in any given state. Optimizing a redistricting scheme to promote proportional representation can be a social good. Acknowledging this, this paper will only measure and not analyze the intent of gerrymandering.

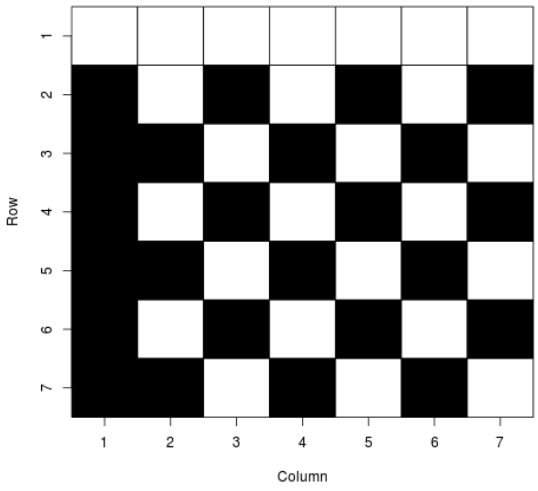

Figure 1: In this hypothetical state, two intuitive ways to split the region into seven districts would be to divide it into rows and columns. When split by rows, black has six majority districts, and white only has one, while when split by columns, the reverse is true.

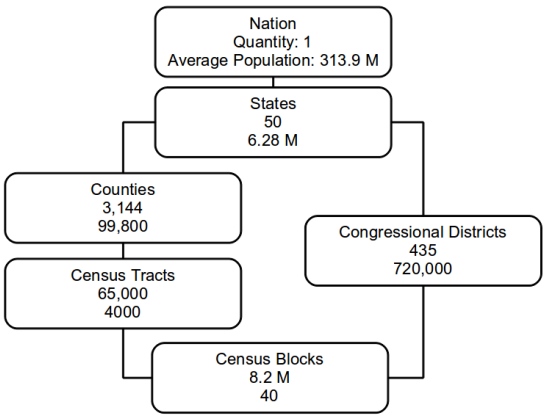

Figure 2: US Geographic Organization, showing the different levels of governance the country is subdivided into. Also shown are the total number of each category, and the average population in each category.

1.2 Geographical Organization

The US is subdivided into regions at many levels as shown in Figure 2. The main subdivisions considered in this paper are at the state level, of which there are 50, the congressional district level, of which there are 435, and the Census tract level, each of which have approximately 4000 people, so there are roughly 175 Census tracts in any congressional district. All data was obtained from the US Census 2010 (U.S. Census Bureau).

2 Generating District Shapes

2.1 Data Sources

The three levels of geographic organization considered in this paper are the state level, the congressional district level, and census tract level. All three boundary levels were obtained from the US Census TIGER/Line Database (U.S. Census Bureau), using information current to the 2010 Census and the 113^{th} congressional district. A few modifications to the data were necessary. Districts 200 and 338 were not actually congressional districts, but rather were portions of states entirely in water, in Connecticut and Michigan. Several states and congressional districts had portions of their boundaries over water, but as all state boundaries

are removed, this does not affect analysis of district lines. For one district in Connecticut and three districts in Tennessee, tiny discontinuous islands are coded into the data, each with an area of less than one square kilometer. This made perimeter calculations nontrivial, and so these regions were removed for simplicity. The 48 continental states and their congressional districts are displayed in Figure 3. All of the Census Tracts in Alabama are displayed in Figure 4.

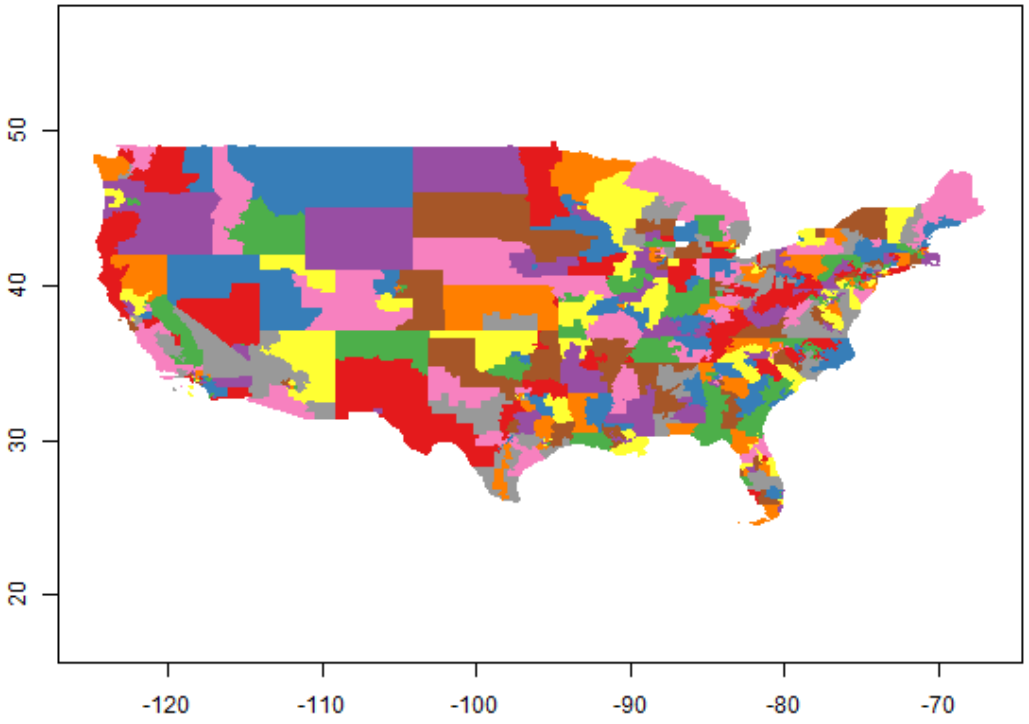

Figure 3: A color-divided map of every congressional district in the continental United States.

2.2 Removing State Borders

One area for major improvement in existing numerical analyses of gerrymandering is that these studies have not considered removing the boundaries. If our goal is to make Congressional reapportionment "fair", one boundary condition that seems reasonable is that state boundaries are immutable. Bearing this in mind, it makes little sense to penalize

congressional districts containing irregular state boundaries. Mathematically, we can write

$$T_n = \sum_j S(\delta_j)$$

where T_n is the score for a state n, δ_j is a district j within state n, and S is a score function of a district. We want to optimize T_n over S, and proposed score functions are a function of the shape of the district. We can further partition S into S_b and $S_{b'}$, where S_b represents the portion of the score contributed by a district border, and $S_{b'}$ represents the portion of the score contributed by other factors, such as area. We can now rewrite

$$T_n = \sum_i S_b(\beta_i) + \sum_i S_b(\iota_i) + \sum_j S_{b'}(\delta_j)$$

where β_i represents components on the state border, ι_i represents district borders on the state interior. Because $\sum_i S_b(\beta_i)$ is a constant for any S, independent of the conformation of congressional districts within a state, holding a district accountable for edges on the state border will not improve statewide estimates of irregularity. More intuitively the entire state border will be included in any configuration of districts, and so we should remove state borders from consideration to get a more accurate inference of actual gerrymandering of a district.

2.3 Replacing State Borders

A simple approach to dealing with state borders would be to remove portions of a congressional border on a state border and do nothing else. A problem with this is that while we can optimize T_n over $S_b(\iota_i)$, several useful metrics contained in $S_{b'}$ become degenerate. Because a congressional district with a removed edge is no longer a closed polygon, metrics that incorporate district area are indeterminate. Both existing score functions assume congressional districts are closed polygons. For a useful comparison between the three scores, we must derive a transformation to turn a congressional district with missing edges back into a polygon.

The procedure used is determined as follows. First, portions of a congressional district boundary on a state boundary are removed. For each removed curve, we calculate the total path length, p_n, bounded by points a and b, the endpoints of the removed curve. There are only two unique planar arcs of a circle passing through both a and b with length p_n, which are reflections of each other through the line between a and b. We arbitrarily pick the arc on the far side of the line between a and b the district centroid. Figure 6 demonstrates this procedure with eight congressional districts. The effect is most pronounced in the Texas 16^{th} district and the West Virginia 2^{nd} district, both of which are shown.

This construction accomplishes a number of goals. It preserves the overall perimeter, and does not change the weight of the contribution to T_n from $S_b(\iota_i)$, but regularizes $S_{b'}(\beta_i')$, where β_i' is the new arc. As the perimeter is kept constant but the arc is directed outwards from the center of the district, the area of the new district is generally slightly larger than the original.

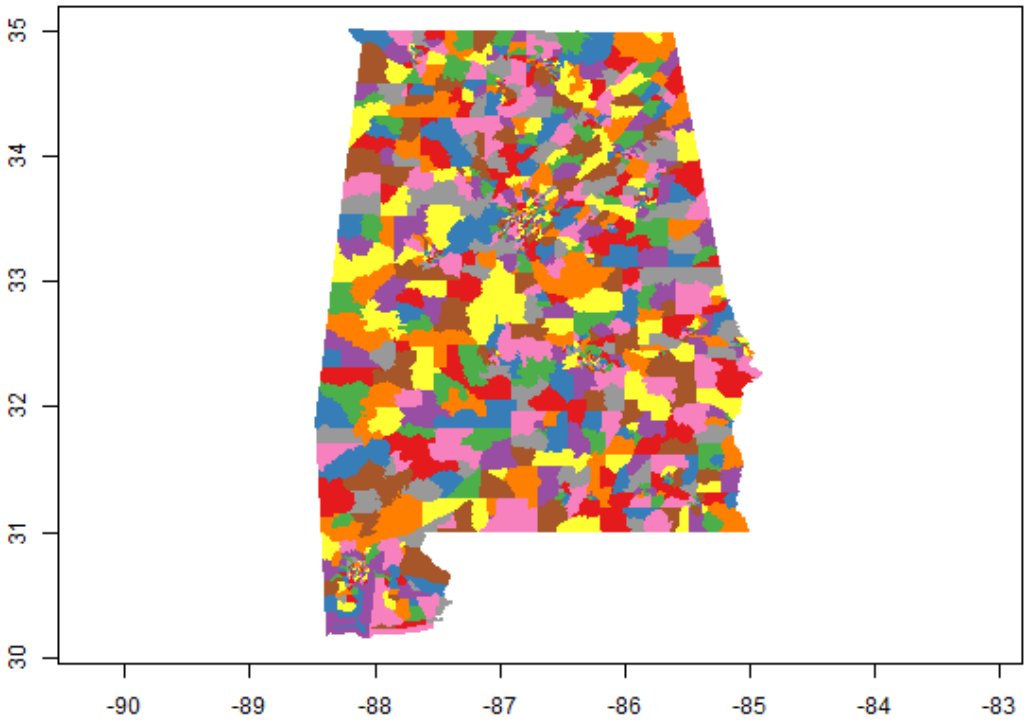

Figure 4: A color-divided map of every census tract in the State of Alabama. Each tract has roughly 4,000 people.

2.4 Regular Spacing

Census shape files generally list vertices of a district, and thus districts with long straight sections can have wide spacing between points. For reasons described in the next section, a maximum spacing interval of 200 meters was chosen between points.

3 Measures of Regularity

Two existing score functions that have been used to numerically describe gerrymandering are a score based on area and a score based on convexity. Additionally, we propose one new measure, the Hausdorff dimension.

3.1 Area Score

An area-based measure of regularity is the Isoperimetric Quotient (IQ), $\frac{P^2}{4\pi A}$, where P is the perimeter of a district and A is the area (Case, 2007, Young, 1988, Gilligan and Matsusaka, 1999). For better comparison with other scores, we define the Area Score as $A_j = \frac{1}{IQ}$, so that the Area Score ranges on (0,1) and is 1 for a circle. This score has the advantage of intuitive understanding and fast computation time (Friedman and Holden, 2008). It should be noted that all metrics are approximated on projections of districts onto a plane. Area will technically be different than when computed on a three-dimensional surface, but approximating a small local region of a spheroid as a plane generally preserves area and distance.

3.2 Convexity Score

Convexity Score (Hodge, Marshall, and Patterson, 2010) can also be used to describe polygon regularity, and has been used to interpret gerrymandering. We can define the Convexity Score as $C_j = P(l(j_1, j_2) \in d_j)$, where j_1, j_2 are points in a district j, l is the line connecting them, and C_j is the probability that line is entirely within d_j, assuming j_1, j_2 are drawn uniformly from d_j. Convexity Score is typically approximated numerically rather than calculated analytically. In our case, for each district 2500 sets of points were randomly selected from the interior, and C_j is the proportion of lines inside the district. The choice of 2500 lines was made in order to bound the approximation error. Specifically, as we model the Convexity Score as a binomial distribution with p as the probability of a line segment between two points staying in the interior. The standard deviation of the estimate of convexity of a district is $\sqrt{\frac{(p)(1-p)}{n}}$, which is maximal at $p = .5$ with a value of $\frac{.5}{\sqrt{n}}$, which at $n = 2500$ is .01.

3.3 Hausdorff Score

The Hausdorff Dimension is a well characterized concept that has been applied to a number of applications (Mandelbrot, 1967). The Hausdorff dimension is an estimate of fractal dimension, so just as a plane has dimension 2 and a line has dimension 1, a curve that widely covers a 2 dimensional plane can be said to have a dimension somewhere between 1 and 2. We compute the Hausdorff dimension of the boundary of any given district as follows. If we subdivide a rectangle bounding a district into ever smaller regions, we interpret the Hausdorff dimension as

$$H_D = lim_{n\to\infty} \frac{\log k}{n \log 2}.$$

where we subdivide the bounding rectangle into 2^{2n} similar rectangles (so each axis is subdivided into 2^n segments), and k is the number of those rectangles containing a section of the boundary. The maximum potential dimension for a district is if the boundary manages to pass through every single point, yielding $k = 2^{2n}$, so $H_D = 2$. The minimum potential

dimension is if the district just stays along the bounding rectangle, yielding $H_D = 1$ in the limit. To choose the discretization of boundary lines for the calculation, 50 sample districts were divided into intervals ranging from 2 kilometers to 25 meters, and a distance of 200 meters was the largest distance for which Hausdorff dimensions were within .01 of the 25 m score, so a distance of 200 meters was chosen for speed. Finally, we normalize the Hausdorff Score as $H_j = 2 - H_D j$, so that all three scores lie on (0,1), and are more regular at 1 and less regular at 0. Of the observed districts we ran, the actual range for H_j was between .489 and .821.

3.4 Score Comparison

Figure 5 shows each of the three types of scores on the same plot, along with the Area Score and the convex score for the full districts with state borders included. The districts are ordered from left to right by ascending Area Score. The correlation between Area Score and the full Area Score is .768, and the equivalent for Convexity Score is .839. The correlation between Area Score and Convexity Score is .775, between Area Score and Hausdorff Score is .830, and between Convexity Score and Hausdorff Score is .727.

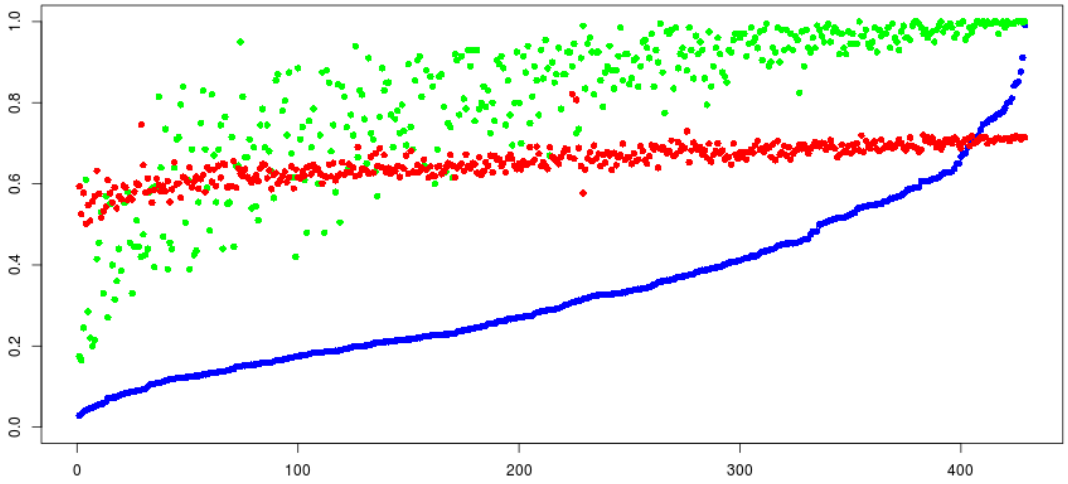

Figure 5: Comparison of three type of scores, ordered by Area Score. The Area Score is in blue, convex score in green, and Hausdorff Score is in red. This demonstrates that while each score is somewhat positively correlated, they are providing unique information about a district, and each may be helpful to assess a different aspect of irregularity.

Figure 6 shows eight districts (with and without replaced state borders), demonstrating that area, convexity, and Hausdorff Scores provide different types of information about the irregularity of a district. TX-16 and CA-13 have similar Hausdorff and Convexity Scores, but

very different Area Scores due to the lengthy eastern perimeter of CA-13. In contrast, CA-13 and CA-37 have similar area and Convexity Scores, though the smooth western border of CA-13 leads it to have a higher Hausdorff Score. Studying the figure, it is seen that Area Score penalizes "knobby" districts, with large perimeter relative to area. In contrast, the Haussdorff score penalizes small-scale perturbations in the borders, hence "knobby" districts where the borders consist of long straight segments will have low Area Score but high Hausdorff Score. So the Hausdorff Score penalizes districts with borders defined by many short line segments, which is more reflective of the fine-scale border perturbations induced by gerrymandering. To evaluate the Area Score of all congressional districts took 3.267 seconds on a personal computer. With the parameters described above, the Convexity Score took 923.143 seconds, and the Hausdorff Score took 214.135 seconds. Neither score takes an insurmountably long time to calculate, but the Area Score is considerably faster than the Hausdorff Score, which is considerably faster than the Convexity Score. Perhaps the biggest benefit the Hausdorff Score offers is that it does not require a closed polygon to evaluate a score, while the other two measures do. Particularly for a problem like this in which we are really only interested in sections of the district boundaries not along a state border, this can be used to offer a more meaningful measure of the true irregularity.

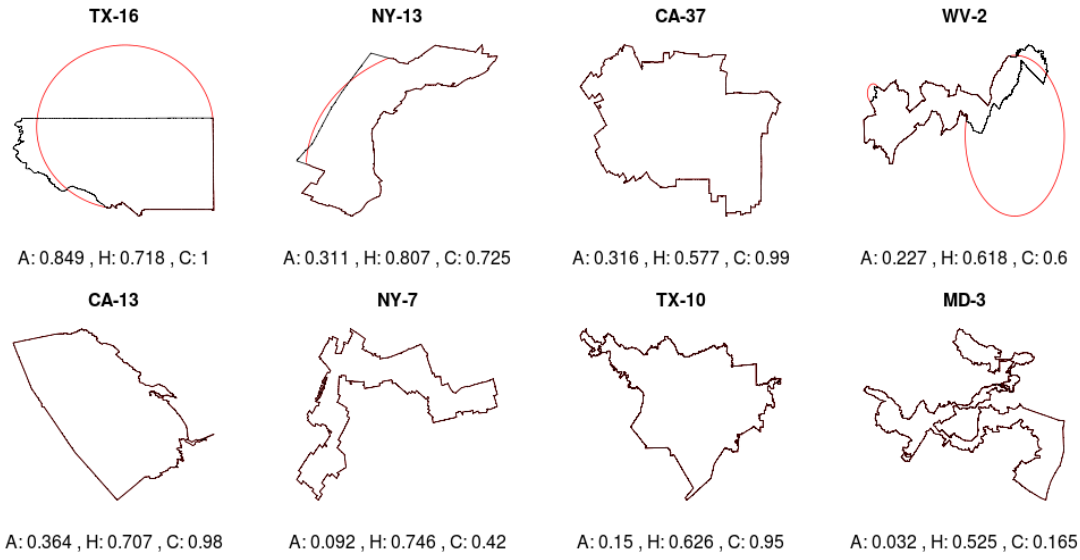

Figure 6: A comparison of congressional districts, divided by their relative scores in each metric. The full districts are shown in black, and the districts without state borders whose scores are used are shown in red. Districts on the top row have relatively higher Area Scores as compared to the bottom row, districts on the left half have relatively higher Hausdorff Scores compared to the right half, and districts in columns 1 and 3 have relatively higher Convexity Scores compared to columns 2 and 4.

4 Model Generation

4.1 Data Collection

Census data at the census tract level was collected from the Investigative Reporters and Editors Reporters and Editors (2013), using US Census 2010 Tables P3 and P12. Election data was collected for the 2012 US Presidential Election for each US County, and the position of each US County was acquired from Gothos Gothos (2013). At the time of publication, voter data was not available all the way down to a census tract, and so an expected vote total for each census tract was generated using a spatial averaging of the counties in the state. Each census tract was estimated to have a percentage of democratic voters given as a weighted average of all the counties in the state, where the weight was proportional to the log of the population divided by the distance in kilometers. States were grouped by redistricting process into five groups: states with a redistricting process governed by a Democrat-controlled legislature, states with a redistricting process governed by a Republican-controlled legislature, states with an independent commission, states with a bi-partisan commission, and states with only one district. For the 2010 redistricting, there were 17 Republican-led redistrictings, 7 Democrat-led redistrictings, 5 with an independent commission, and 14 with a bipartisan task force.

4.2 Voter Influence

To measure voter influence, elections were modeled as a binomial distribution based on the population n and the percentage of voters in the district voting for a Democratic candidate in the 2012 election, ρ. Define

$$\Phi = \log(P(\text{Bin}(\frac{n}{s}, \rho) = \lfloor \frac{n}{2s} \rfloor)$$

If $s = 1$,then Φ is the log probability that an individual voter determines the outcome of an election. Given that n averages 720,000, for $s = 1, \Phi$ is incredibly small. To make the numbers more interpretable, the scale factor $s = 1000$ was used. Φ is maximized at $\rho = .5$, at which point Φ = -3.502. For $\rho = .4$, Φ = -17.789, so Φ drops off quite quickly as a district gets less competitive.

4.3 Covariates

At the Census Tract level, we analyzed data on total population, race, sex, age, percent voting Democrat, Voter Influence, and state redistricting type. For race, we considered the proportion of voters that were White, Black, American Indian, Asian, Pacific Islander, Hispanic, and Multiracial. For age, we looked at the proportion in each Census Tract in the age groups (0-17,18-24,25-34,35-44,45-54,55-64,65+).

4.4 Fitting a Model

For any given congressional district, covariates for that congressional district were generated by summing or averaging the covariates of all Census Tracts whose centroid lie within the bounds of that District. While there are Census Tracts that straddle the border between two congressional districts, this is a reasonably close approximation to the true model. The full model used was

$$\log(-\Phi) \sim \sum \rho_i p_{\text{race}i} + \phi p_{\text{female}} + \sum \alpha_i p_{\text{age}i} + \delta p_{\text{democrat}} + \sigma_i \text{factor(statetype)}$$

p_x is defined as the proportion of the population in a region with quality x. This model was evaluated using a simple linear regression over each congressional district in a state with more than one district, where negative voter influence is measured on the log scale.

5 Results

5.1 Linear Model

Using backwards selection from the full model, we ran an F-test of the full model against each of five reduced models without either race, sex, age, proportion of democrats, or state-type. The respective p-values of these F-tests were of .0121, .0327, .00412, .00136, and .00539 all of which were significant below the .05 level. Each term thus contributes meaningful information to Voter Influence. In particular the state type term is highly significant. We can define $I_j = -\log(-\Phi_j)$, which is defined on $I_j < 0$, with a higher I_j corresponding to a higher Φ. As compared to states whose redistricting is controlled by Republican legislatures, states whose redistricting is controlled by Democratic legislatures have I_j values estimated to be 7.24 ± 3.21 percent lower, controlling for race, sex, age, and proportion of democrats. States whose redistricting is done by non-partisan commissions have I_j values estimated to be 16.72 ± 4.71 percent higher, and states with bi-partisan redistricting committes have I_j values estimated to be 19.82 ± 3.58 percent higher than states with Republican legislatures. Without even looking at district scores, simply knowing what type of redistricting a state uses offers tremendous insight into voter influence, and non-partisan and bi-partisan states have significantly higher voter influence than partisan states.

5.2 Tests of Regularity

Additional models were generated in which each score was added to the full model. Another F-test was performed on each model compared to the full model, which generated p-values of .0342, .0498, and .0101, for the models adding information from the Area Score, Convexity Score, and Hausdorff Score respectively. Thus, each of the scores provides a significant source of information for Voter Influence. Generating a final model combining all three scores generates a p-value of .00981. It can thus be seen mathematically that the Hausdorff Score provides relatively more information than Area Score or Convexity Score, and all

three together provide useful information. Analyzing the coefficients in the extended linear model suggests that an optimal strategy for maximizing voter influence lies in maximizing the quantity

$$.163A_j + .183C_j + .654H_j,$$

a weighted average of the three scores analyzed, over all congressional districts in a state. All possible combinations of coefficients for each of the three scores such that the coefficients summed to 1 and were multiples of .001 were tested. Among these, the regression using the above coefficients was found to have a lower p-value than any other combination.

5.3 Resampling

We wanted some measure of how suboptimal the Voter Influence was in a given state's arrangement of congressional districts. To accomplish this, we generated many random samples of congressional districts to generate a null distribution for Voter Influence. The algorithm to generate a random sample in a given state proceeded as follows. First, we generated 1 point for every thousand people (rounded up) in a Census Tract located at the centroid of each Census Tract. Selecting a thousand was an arbitrary scale parameter to speed up runtime. We then ran a modified k-means clustering algorithm, which split each state into the correct number of districts of equal size. To do this, we began by splitting each set of points into k clusters, where k is the number of congressional districts in that state using a standard k-means algorithm. We then reshuffled points until the clusters were of equal size, moving points from larger districts to smaller districts. This is non-deterministic, but generates a set of equally sized clusters that are somewhat locally compact; an intuitive way to partition a state into equally sized regions. We then have a new set of districts, and for each district we can average and sum the covariates of the points assigned to the district to get a new set of covariates. Based on this we can calculate Voter Influence for each of the new districts. We run this algorithm 1000 times for each state to obtain a null distribution for Voter Influence, Φ.

Though we are guaranteed to have k districts each iteration, the ordering of the resampled districts has little meaning. Therefore, for comparison, we look at the order statistics of Voter Influence of the resampled districts compared to the order statistics of Φ of the actual districts. We can define

$$M_j = P(\Phi_{(j)} > \hat{\Phi}_{(j)}),$$

where $\hat{\Phi}$ is the estimated Voter Influence for the resampled districts and j is the order statistic. Analyzing M_j in each state will tell us how much better or worse the current distribution is than a random resampling. A low M_j indicates that a state's current congressional district distribution gives its voters less influence than assigning districts through the equal size k-means clustering method described above.

5.4 Shortest Splitline

A method for redistricting that can be shown to optimal under all three score functions is the Shortest Splitline method (for Range Voting, 2013, Benn and German, 2008, Altman, 1997). The shortest splitline algorithm is a recursive process that works as follows. To split any state into k districts, if k is even take the shortest possible line such that half of the individuals in the state are on one side of the line and half are on the other. If k is odd take the shortest line such that there the proportion of individuals on one side of the line is $\frac{k-1}{2k}$. Iterate the algorithm with both subsections, and continue until there are k subsections.

Shortest splitline guarantees convex districts (when state boundaries are removed) because each line is bounded by either a state boundary or another split line. Thus it is impossible to create a region where one district is concave. It also gives optimal Hausdorff Score, as each new district will be composed of a small number of straight lines and arcs, each of which have dimension approximately 1. For Area Score, the shortest splitline method creates districts that are largely circular with a few straight lines in place of arcs, and so keeping the lines as short as possible also optimizes the Area Score.

Another method that uses additional data to optimize fairness incorporates both population density and compactness (Belin, Fischer, and Zigler, 2011). The rationale behind incorporating population density is that there is a high correlation between population density and political preference, and so making divisions that are both compact and with low variation in population density could further increase statewide competitiveness.

We ran the shortest splitline method and again, determined covariate values for each new district using the same spatial averaging method defined in 4.1. This allowed us to calculate Φ for each new district, and we can once again calculate how much better the optimal shortest splitline solution does compared to both existing districting as well as the null distribution defined by the modified k-means algorithm in Section 5.3 .

5.5 Comparing Clusters

For each state, we generated a null distribution through 1000 iterations of the modified k-means algorithm, and also calculated the optimal splitline method. Figure 7 shows for the state of Alabama, California, Colorado, and Connecticut the density of the Φ scores for each of the districts in each set of clusters, along with black vertical lines showing Φ for current districts, and red vertical lines showing Φ for districts optimized by splitline. For the actual, each of the random districts, and the optimal districts in each state we can order the districts by Voter Influence.

It should be noted that the order has little geographic significance; the actual Colorado district with the most Voter Influence may be in a different region of the state than the optimal region with the highest influence. Comparing districts of equal ranks gives a numerical assessment of how different redistricting schemes affect Voter Influence in a state as a whole. So for any two redistricting schemes, we can compute the probability that the n^{th} highest score in one set of districts is higher than the n^{th} highest score in another set. We do this for the optimal and actual districts, and then for both scores with each random set of districts,

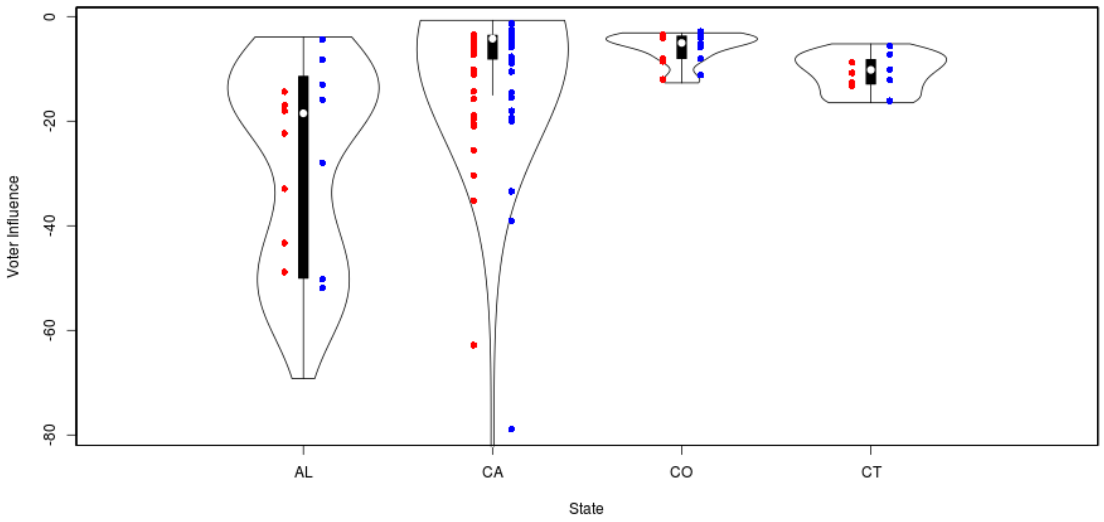

Figure 7: A violin plot showing the relative Voter Influence in Alabama, California, Colorado, and Connecticut. In red, the Voter Influence in the actual congressional districts are shown. In blue, the Voter Influence in the optimal configuration from the shortest splitline method is shown. The violin plot shows the distribution of Voter Influence scores from 1000 simulations in each state. In each case, the optimal distribution of districts is better than the "random" simulations, which are in turn better than the current configuration. Alabama has a republican controlled process, California has a non-partisan process, Connecticut has a democratic process, and Colorado has a bipartisan process.

and average those results, shown in Table 1. As can be seen, in no current redistricting process is the voter influence of a district better on average than in the random case, although the bipartisan process is closest. The splitline method outperforms random distributions in terms of voter influence roughly 58 percent of the time, and widely outperforms the actual districts in each of the four categories. Figure 8 shows the proportion of actual districts within each state that have higher voter influence than the randomly generated districts.

6 Conclusions

This paper developed a number of interesting methods and results for understanding the impact of gerrymandering on voter influence, Φ. The main goal of the paper was to develop a numerical estimator of a set of congressional districts to predict voter influence. We looked at two scores that have been used in this context, Area Score and Convexity Score, and compared them to a novel score based on the Hausdorff dimension. Additionally, we transformed each district to replace state boundaries with arcs of circles of equal perimeter to provide a more meaningful analysis of the actual redistricting. We analyzed at the Census

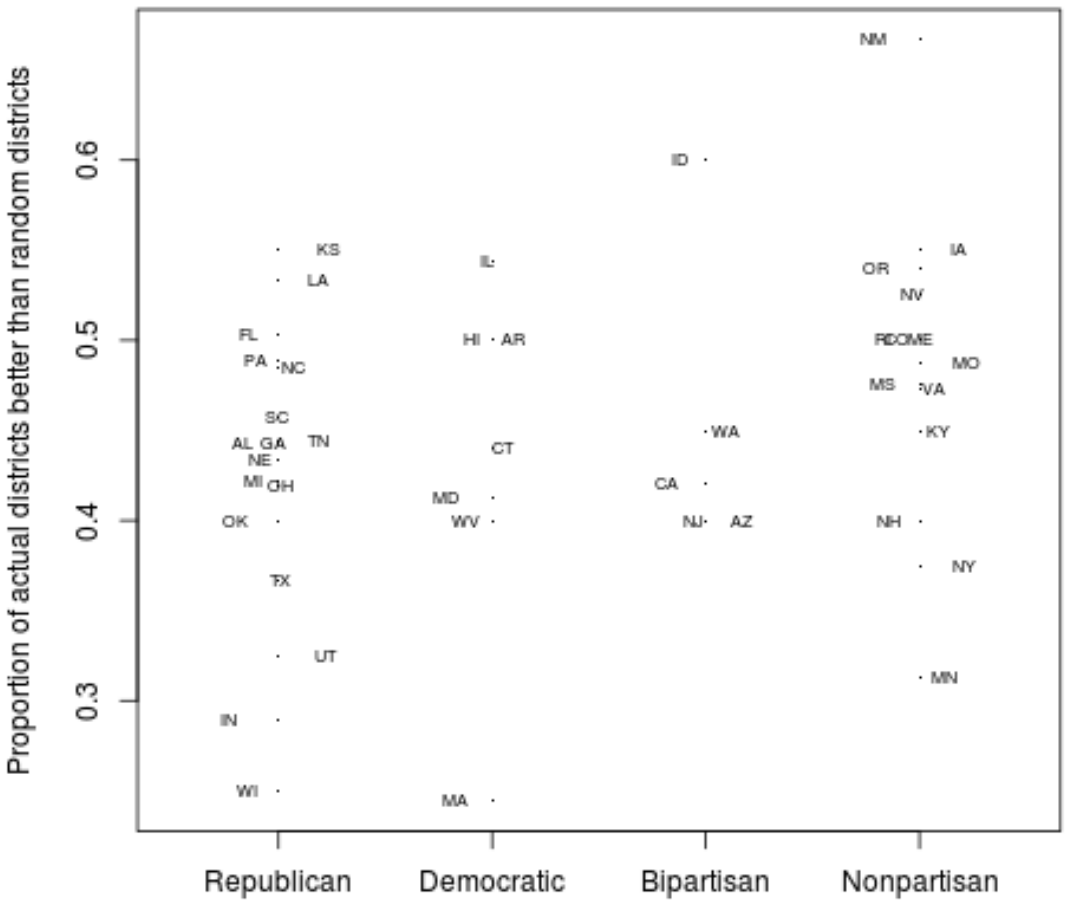

Figure 8: A plot of all 43 states with more than one district showing the proportion of actual districts with higher voter influence scores than the randomly generated districts, delineated by the type of redistricting process. As can be seen, the states with bipartisan and nonpartisan redistricting processes tend to have better actual processes as compared to the random realignments, but even they fall short, with only Florida, Idaho, Illinois, Iowa, Kansas, Louisiana, Nevada, New Mexico, and Oregon performing better than at random.

Table 1: For each state, three sets of districts were analyzed: the actual districts (A), the optimal districts as defined by the shortest splitline method (O), and 1000 different sets of "random" districts derived from the equal sized k-means method (R). For each type of redistricting process, the probability that a voter has higher influence in each set of districts is shown. For example, among states with Republican controlled redistricting processes, districts generated using the shortest splitline method will have voters with higher influence than the actual districts today in 78.72 percent of cases.

	P(A>R)	P(O>R)	P(O>A)
Republican Controlled	.4288	.5841	.7872
Democratic Controlled	.4462	.5973	.7355
Non-partisan	.4285	.5771	.9000
Bipartisan	.4721	.5893	.7312

Tract level a number of covariates, including percent voting for Democrats, race, sex, age, and type of redistricting in that state. A voter influence score was generated for any district, real or hypothetical, given as the log probability that a single voter could change the outcome of an election.

There are a few areas of data collection that could be a source for future research. First, running the same tests looking at multiple different elections would reduce variance of our estimates. Second, by looking at how districts are realigned each decade, we could infer a causal effect of state redistricting type on voter influence over time. Finally, while this study analyzed down to the Census Tract level, analysis at the Census block level might provide additional accuracy and insight.

We demonstrated that states with bipartisan and nonpartisan redistricting procedures had higher estimated Φ scores, even after adjusting for multiple baseline covariates. Extended models were then fit incorporating each of our three types of scores, both separately and jointly, and found the Hausdorff Score to provide the most information, but all three were useful together. In our final model, Φ was optimized by maximizing the quantity $.163A_j + .183C_j + .654H_j$. We then generated a null distribution of congressional districts by dividing each state into equally sized clusters using a modified k-means algorithm. An optimal distribution was also generated using the Shortest Splitline method, which is optimal over all three scores. In each of the four types of states, the actual districts had lower voter influence on average than the randomly generated districts, which in turn had lower voter influence than the optimal districts. Relatively speaking, states with bi-partisan committees underperformed to a lesser degree than other types.

There are many considerations to take into account when addressing the issue of redistricting. This paper has developed a numerical approach to measuring the amount of gerrymandering, and has shown that in states with a higher degree of gerrymandering voters have relatively less influence. While voters in states with bi-partisan committees have more influence than in other states, even they do not even perform as well as a randomly aggregated set of geographically compact districts. The Shortest Splitline method optimizes

congressional districts both over measures we looked at for regularity and voter influence, demonstrating that using this algorithm is an improvement for voters.

References

Adams, B. (1978): *The Unfinished Revolution: Beyond One Person, One Vote*, National Civic Review.

Altman, M. (1997): *Is automation the answer: The computational complexity of automated redistricting*, Rutgers Computer and Law Technology Journal.

Ansolabehere, S., J. Snyder, and C. Stewart (2000): *Old Voters, New Voters, and the Personal Vote: Using Redistricting to Measure the Incumbency Advantage*, American Journal of Political Science, volume 44, number 1 edition.

Belin, T., H. Fischer, and C. Zigler (2011): *Using a Density-Variation/Compactness Measure to Evaluate Redistricting Plans for Partisan Bias and Electoral Responsiveness*, Statistics, Politics, and Policy, volume 2, issue 1 edition.

Benn, A. and D. German (2008): *Unbiased Congressional Districts*, Conference on Computational Geometry.

Cain, B., K. MacDonald, I. Hui, N. Boyle, A. Lee, and A. Woods (2006): *Competition and Redistricting in California: Lessons for Reform*, Berkeley Institue of Government Studies.

Case, J. (2007): *Flagrant Gerrymandering: Help from the Isoperimetric Theorem?*, SIAM News, volume 40, number 9 edition.

Coate, S. and B. Knight (2007): *Socially Optimal Districting: A Theoretical and Empirical Exploration*, The Quarterly Journal of Economics.

Cox, G. and J. Katz (2002): *Elbridge Gerrys Salamander: The Electoral Consequences of the Apportionment Revolution*, Cambridge University Press.

for Range Voting, T. C. (2013): URL `http://rangevoting.org/`.

Friedman, J. and R. Holden (2008): *Optimal Gerrymandering: Sometimes Pack but Never Crack*, American Economic Review.

Gilligan, T. and J. Matsusaka (1999): *Structural Constraints on Partisan Bias under the Efficient Gerrymander*, Public Choice, volume 199, number 1-2 edition.

Gothos (2013): URL `http://gothos.info/`.

Grainger, C. (2010): *Redistricting and Polarization: Who Draws the Lines in California?*, Journal of Law and Economics.

Grofman, B. (1983): *Measures of Bias and Proportionality in Seats-Votes Relationships*, Political Methodology.

Gul, F. and W. Pesendorfer (2010): *Strategic Redistricting*, The American Economic Review.

Hodge, J., E. Marshall, and G. Patterson (2010): *Gerrymandering and Convexity*, The College Mathematics Journal, volume 41, number 4 edition.

Maceachren, A. (1985): *Compactness of geographic shape: Comparison and evaluation of methods*, Geograska Annaler, series b, human geography 67 edition.

Mandelbrot, B. (1967): *How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension*, Science, volume 156, number 3775 edition.

Owen, G. and B. Grofman (1988): *Optimal Partisan Gerrymandering*, Political Geography Quarterly, volume 7, number 1 edition.

Reporters, I. and Editors (2013): URL `http://census.ire.org/`.

Sherstyuk, K. (1998): *How to Gerrymander: A Formal Analysis*, Public Choice, volume 95, number 1-2 edition.

Stephanopoulos, N. (2012): *Spatial Diversity*, Harvard Law Review.

Taylor, P. (1973): *A new measure for evaluating electoral district patterns*, The American Political Science Review, volume 67 edition.

U.S. Census Bureau (2011): *2010 Census Summary File 1*.

Vickrey, W. (1961): *On the prevention of gerrymandering.*, Political Science Quarterly, 76 edition.

Washington, E. (2006): *How Black Candidates Affect Turnout*, Quarterly Journal of Economics, volume 121, number 3 edition.

Young, H. (1988): *Measuring the Compactness of Legislative Districts*, Legislative Studies Quarterly.