Mathematical Aspects of Gerrymandering

Mary Solbrig

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2013

Committee:

Sara Billey

John Sylvester

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

Mathematical Aspects of Gerrymandering

Mary Solbrig

Chair of the Supervisory Committee:
Professor Sara Billey
Mathematics

Every 10 years the United States performs a census, and this census determines how many members of congress will represent each state. Then begins an unfortunate battle, as cartographers manipulate the boundaries for political power in a process called *gerrymandering*. A city can become one strong Democratic hold out, or divided up creating multiple more moderate districts. A previously strong Republican district can be changed to include some more Democratically leaning districts, resulting in a competitive race where previously none existed. These fights can expose the worst of the political system, and have driven many researchers to try to find ways to prevent these power grabs. Researchers have suggested ways of measuring the districts to detect Gerrymandering by looking at the boundaries of the districts for irregularity, or by looking at the results over time to detect a bias towards one party, or even ways to generate the maps automatically, free of human intervention. In this thesis, I give an overview of some of these approaches, how to implement them using the R mapping packages, and the innate limitations of the analysis.

TABLE OF CONTENTS

Page

List of Figures ii

Chapter 1: A Simplified Model 1
1.1 The Simplified Model 1
1.2 Optimal Gerrymander 5
1.3 Measuring Outcome 7
1.4 Conclusion 9

Chapter 2: Convexity Measures 11
2.1 Measuring Convexity 11
2.2 Area Ratio 11
2.3 Path-Based Measure 12
2.4 Limitations 15

Chapter 3: Automatic Districting 19
3.1 Uses of Algorithms 19
3.2 Districting Algorithms 20
3.3 Further Research 22

Appendix A: Software and Data 24
A.1 File Types 24
A.2 Software 25
A.3 Data Sources 25

Appendix B: R Code 27
B.1 Working with Shapefiles 27

Appendix C: Convexity Measures, Alphabetical 31

Bibliography 47

LIST OF FIGURES

Figure Number — Page

1.1 A simplified model of a state with 30 percent of the population Blue, 70 percent Red, and three natural ways to divide the state. 2

1.2 Left: A picture of a fitted seat-vote line [23]. The slope, $\hat{\beta}$, is termed 'swing ratio' or 'responsiveness.' Right: A seat-vote line applied to a simple state with vertical districting and 8 districts. 4

1.3 Two different Districting of a state, plotted against the seat-vote histogram. The x-axis is proportion of the vote for the Blue party, and the y-axis is the number of districts. 4

1.4 Graphs and R replication code from "Visualizing US House Results with a Seats-Votes curve" by Jason Holt at OffensivePolitics.net. See his blog for excellent articles on using R to visualize election data. 5

1.5 The optimal map to maximize the number of Red congressional seats. Pack the blue votes into just a few districts, and then spread the remaining blue votes evenly throughout the remaining districts. 6

1.6 The maximum number of seats a party can win using gerrymandering versus how many they would win proportional to the vote in a state with 10 districts. 8

1.7 Over representation by state in the 2012 election. Calculated as the difference between the number of seats a party won, versus how many they would have won proportional to the percentage of votes that party received. California has 5 extra democratic representatives, Ohio has 4 extra Republicans. 9

2.1 Randomly generated paths. All of the paths in Carolina's 12th pass out of the district, where as only 3 of the paths pass out of Washington's 4th. 13

2.2 The districts of Maryland. The ratio convexity measure from section 2.2 compares the ratio of the district with its convex hull. The Path-based measure from section 2.3 takes the probability that a random path in the district will not pass through a different district. Displayed is a Monte Carlo estimate of the path based measure, and 95% confidence interval. 16

2.3 The map of the congressional districts of Maryland for the 113th Congress . . 17

B.1 Plots produced by Code Chunk 8 30

ACKNOWLEDGMENTS

I would like to express sincere appreciation to my book group, Gavin, Kay, Thuy, Ben, Eli, Angela, Lauren, and Vlad, that has provided much needed escape from the academic bubble over the course of the past couple years. I would also like to thank Sayan Banerjee for all of the support and tea. Finally, she would like to offer the deepest appreciation to her advisor Sara Billey. You helped give a focus and purpose to my graduate study that without you I'm not sure I would have achieved. She would like to thank her for her time and energy, it made my last year worth something.

Chapter 1

A SIMPLIFIED MODEL

For this thesis I will be focusing on the US House of Representatives, although much of the analysis could apply to any system dominated by two parties. There are 435 voting members of the House of Representatives, and every state is guaranteed at least one representative. From there, the number is determined by the population of the state, where California has the most with 53 representatives, and 7 states have only 1 representative. Each state divides itself into congressional districts of approximately equal population, currently about 1 congressional member for 700,000 voters, more than triple that of the 1910 census[8]. However, the way in which the congressional boundaries are drawn is left up to the individual state. In some states the districting plan is done by a non-partisan committee, but in most states the map proposals and final decision is left up to the state legislature [3].

1.1 The Simplified Model

A very simplified model for an election is a state with uniform population distribution, where the left side of the state is dominated by the Blue party, and the right side of the state is dominated by the Red party. Thus, if a quarter of the population votes Blue, then the left quarter of the state could be colored Blue, and the right three quarters would be colored Red. There are a few obvious ways to partition this state into d districts: with horizontal districts, or vertical districts, or slanty districts, but the only stipulation is that all of the districts must be contiguous (connected) and of approximately equal area (since population is assumed to be uniformly distributed). The winner of each district can be computed by determining the percentage of its population which is Red or Blue

There are different goals one might want to achieve in designing a district. One goal is to make the districts reflect natural divides in the population, grouping similar groups

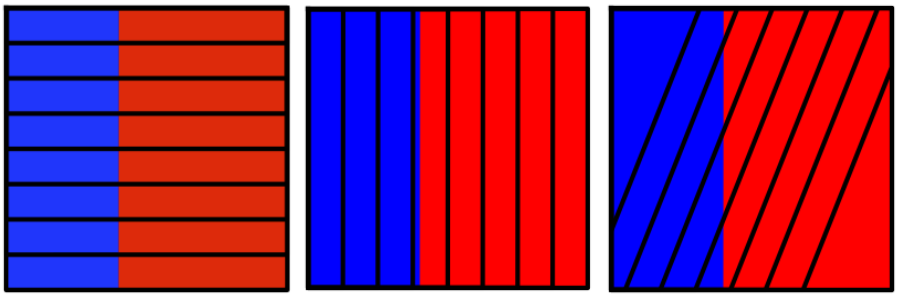


Figure 1.1: A simplified model of a state with 30 percent of the population Blue, 70 percent Red, and three natural ways to divide the state.

together. The goal can be phrased as minimizing dissatisfaction amongst the voters; that is, minimizing the number of voters in districts represented by a representative not of their own party. By this measure, the best plan would either have no blue districts with red voters in them, or no red districts with blue voters in them. Otherwise, if there exists a blue district with a red voter and a red district with a blue voter, the voters could be switched, resulting in a higher satisfaction score. Of the plans above, the vertical districting plan is the best.

An objection to this districting plan is that districts comprised of a very uniform population will result in highly polarized districts, unable to compromise with each other. After all, part of the goal of creating a republic was to control the rise of factions harmful to the good of the union by joining people of diverse interests together. In the Federalist 10, James Madison wrote that, "Among the numerous advantages promised by a well constructed Union, none deserves to be more accurately developed than its tendency to break and control the violence of faction." In this case, a goal should be to increase the amount of diversity within each district, in order to promote compromise within the district itself, and to keep the representative of that district attentive to the needs of both parties. As Madison wrote, "In an equal degree does the increased variety of parties comprised within the Union, increase this security (against factions)." [16] However, to reduce the number of polarized districts, one has to increase the number of dissatisfied voters, because it is only by grouping people together with people they wouldn't strongly identify with that you break up the factions. Therefore, these are inherently competing desires.

Another goal is to have the outcome of the election match the vote totals as closely as possible. That is, if the Red Party receives 10 percent of the vote, then they should also receive approximately 10percent of the seats. This would not be the case in the horizontal districting choice above, because if one party receives even slightly more than 50 percent of the votes, they will carry every single district. On the other hand, the vertical districting choice will match the vote very closely. If there are d districts, then there will be at most a $1/(2d)$ difference between the fractions of districts a party carries versus the fraction of votes they receive. This concept can be captured in a fitted seat-vote line. In the simplified state, the percent of seats a party takes is plotted against the percent of votes all candidates of that party received, creating a step function. A line of best fit is added to the graph, and bias is defined to be the difference between 50 percent and the percentof votes a party needs to carry 50 percent of of the seats.[23] Edward Tufte, in his paper, "The Relationship Between Seats and Votes in the Two Party System," suggests another set of goals: the plan should be responsive, that is a shift in votes should result in a shift in congressional make up, and it should be unbiased, so that if 50 percent of the votes fall for a certain party, then that party should carry 50 percent of the seats. In actual states, this plot can be created by taking a scatter plot of recent elections, or by creating simulated elections based on past data.

For a more refined approach to estimating bias, see Andrew Gelman and Gary King's paper, "A Unified Method of Evaluating Electoral Systems and Redistricting Plans."[12] They update Tufte's approach, as well as improve the statistical interpretation.

Tufte also suggests a way to visualize political diversity over a state with a graph called the seat-vote curve. In a single district, diversity is measured by the percentage of votes each party received. A district in which one party receives close to 100 percent or 0 percent vote would be polarized towards one party, a district with close to 50 percent votes would be a swing district. Then, to display diversity over the entire state, make a histogram of districts that received between 0-10 percent, 10-20 percent, 20-30 percent, etc of the vote.

This plot can be smoothed by using kernel density estimation methods, creating the Seat-Vote Curve. Treat each district as coming from a distribution, such as a Gaussian,

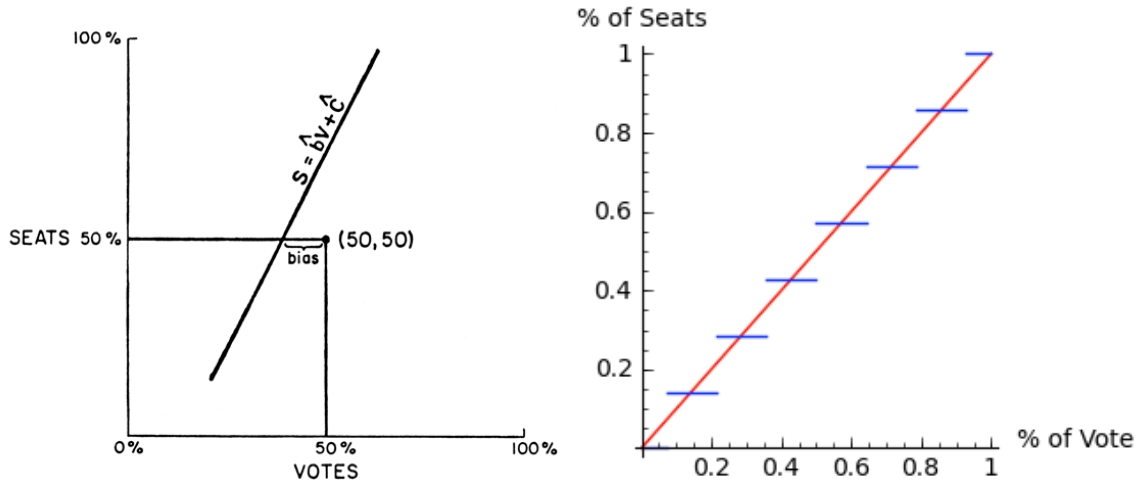


Figure 1.2: Left: A picture of a fitted seat-vote line [23]. The slope, $\hat{\beta}$, is termed 'swing ratio' or 'responsiveness.' Right: A seat-vote line applied to a simple state with vertical districting and 8 districts.

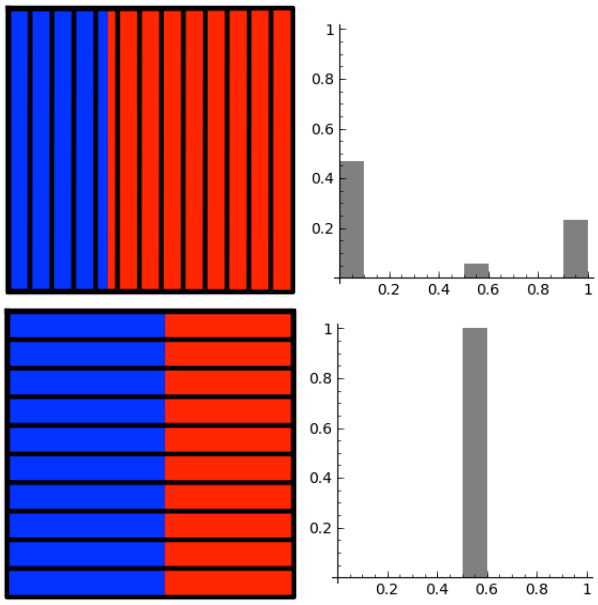


Figure 1.3: Two different Districting of a state, plotted against the seat-vote histogram. The x-axis is proportion of the vote for the Blue party, and the y-axis is the number of districts.

with mean equal to its observed value. This determines a kernel function $K(x) = \frac{e^{-t^2/1}}{\sqrt{2\pi}}$. Then add each of these distributions together, where the width of the distribution, called the smoothing window, is dependent on the number of data points (in our case the number of districts). The final function is written as

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

where h is the smoothing window parameter.

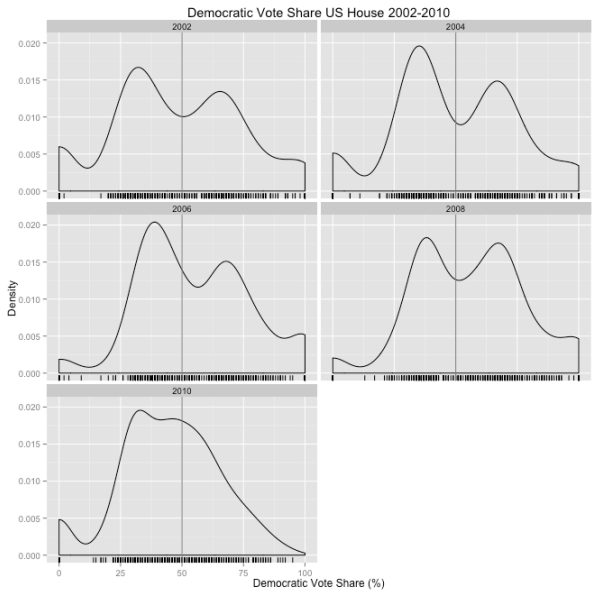


Figure 1.4: Graphs and R replication code from "Visualizing US House Results with a Seats-Votes curve" by Jason Holt at OffensivePolitics.net. See his blog for excellent articles on using R to visualize election data.

1.2 Optimal Gerrymander

From the previous discussion, it's clear that determining a perfect district plan is in essence a political, not mathematical, question because there are conflicting interests at play. However, what if the question was reversed: how would we choose a district plan that would be as biased towards one party as possible? For instance, say that the Blue party has complete

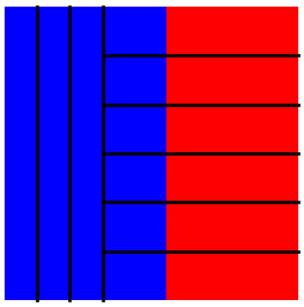


Figure 1.5: The optimal map to maximize the number of Red congressional seats. Pack the blue votes into just a few districts, and then spread the remaining blue votes evenly throughout the remaining districts.

control over the district plan, and wishes to draw the districts in a way that will result in as many Blue congress people as possible.

Assuming a simplified state such as this, it is possible to determine a bound for how much advantage a party can receive through gerrymandering.

Lemma: If p is the percentage of voters in the state for a party, n the total number of districts, then the maximum number of districts G that the party can carry by a margin of m or greater is

$$G = min\left(n, \left\lfloor \frac{np}{.5+m} \right\rfloor\right)$$

Proof. Certainly $G \leq n$. If the party wins a district by a margin of m, then $.5+m$ percent of the district is from the winning party. Each district will hold $1/n$th of the electorate, so this accounts for $(m+.5)/n$ percent of the entire state. Therefore, if G districts are won by a margin of m or more, then the percentage of the electorate accounted for is at least $G(.5+m)/n$. This must be less than p, so we solve the inequality

$$\frac{G(.5+m)}{n} \leq p \quad \Rightarrow G \leq \frac{np}{.5+m}.$$

Further, in an ideal state, the bound can be achieved. If $p \geq .5+m$, then spread the voters evenly through n districts, and the party will carry all of the districts by a margin of m or greater. If $p < .5+m$, then spread the voters evenly through $\left\lfloor \frac{np}{.5+m} \right\rfloor$ districts, leaving

the remaining districts entirely dominated by the other party. Dividing p by the number of districts gives

$$\frac{p}{\frac{np}{.5+m}} = \frac{.5+m}{n}.$$

Since each district has $1/n$th of the population, this means that the party wins in each remaining district by a margin of at least m in the $\left\lfloor \frac{np}{.5+m} \right\rfloor$ districts, so the bound is achieved. □

Theorem: If P is the number of seats a party would receive by a proportional vote and G is the number of seats a party would receive by gerrymandering the votes in their favor, then G is at most $2P$.

Proof. Let p is the percent of the vote a party has, n the number of districts, and P the number of votes the party would receive proportionally, $P = [n * p]$. By the previous lemma,

$$G \leq \left\lfloor \frac{np}{.5} \right\rfloor = \lfloor 2np \rfloor$$

If $\lfloor 2np \rfloor$ is even, then $2np = 2k + \alpha$ for some $\alpha < 1$, so that $np = 2k + \alpha/2$, and $[np] = \lfloor np \rfloor$.

If $\lfloor 2np \rfloor$ is odd, then $np = k + \alpha$ for some $\alpha > .5$. In this case, $\lfloor 2np \rfloor = 2k + 1 < 2(k+1) = 2\,[np]$, so

$$G \leq = \lfloor 2np \rfloor \leq 2\,[np] = 2P.$$

□

This bound is not achieved in any states with a large number of districts, but in Idaho the bound is achieved because proportionally there would be one Democratic and one Republican representative, but instead there are 2 Republican representatives.

1.3 Measuring Outcome

One approach to detecting gerrymandering is to try to quantify how a particular map affects the outcome of an election. One way to do this is to compare the number of seats a particular party got to the percent of votes they received. For instance, in the 2012

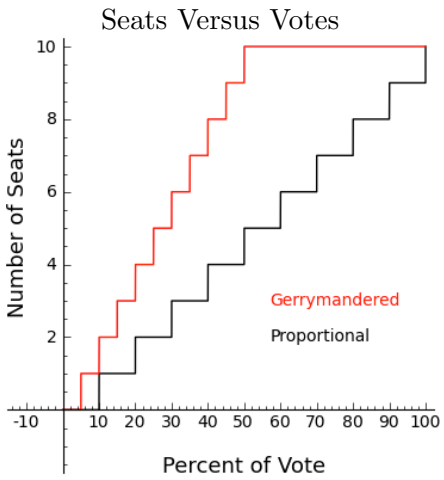
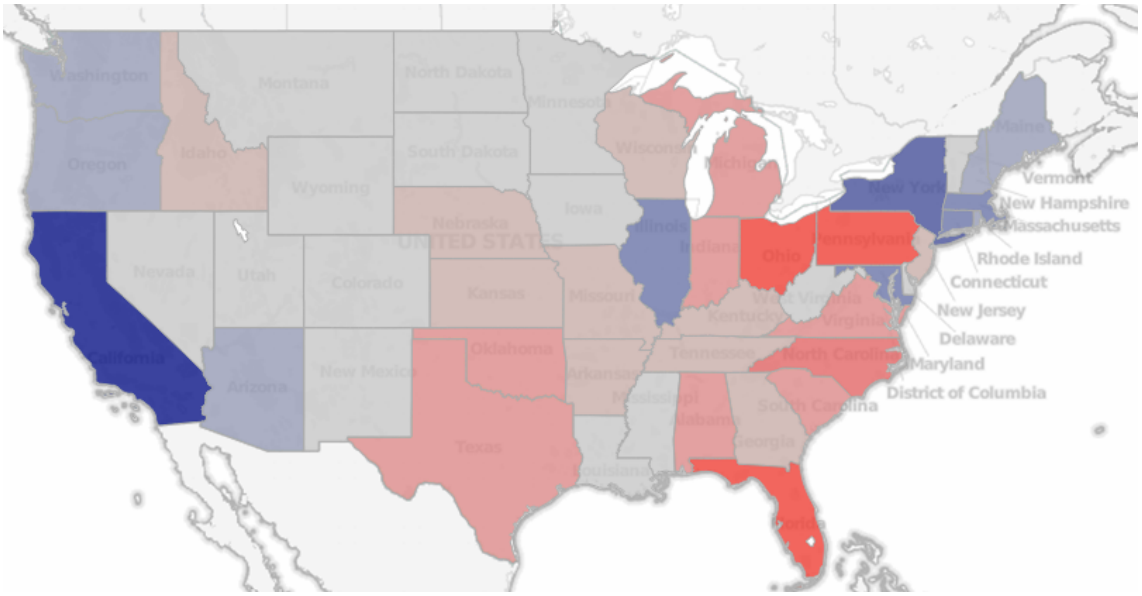


Figure 1.6: The maximum number of seats a party can win using gerrymandering versus how many they would win proportional to the vote in a state with 10 districts.

congressional election, final result showed that while Republicans took 240 of the 435 seats, or 53.8 percent, they only took 48 percent of the popular vote. Even excluding third party votes, they still took only 49.4 percent of the vote [24]. For reference, if a coin is flipped 435 times, there is about a 1.7 percent chance that 240 or more heads would appear. Considering each district as a coin flip, On the face of it, this seems to be a good measure of bias, similar Tufte's concept of bias in the seat-vote line.

However, a recent paper "Unintentional Gerrymandering" by Jowei Chen and Jonathan Rodden challenges this as an accurate measure of gerrymandering. They proposed that even completely absent political influence, the natural geography of the state may contribute to the overrepresentation of certain groups of people. To test this, they took a map of Florida, and used a randomized algorithm to create districts. Then they used historical election data to determine how many congressional seats each party would have won. Out of 100 trials, every simulated district map that they generated was biased towards Republicans[10]. They proposed that in Florida the natural clusters of Democratic voters in cities created naturally packed Democratic votes, and then dispersed the smaller pockets of Democratic votes throughout the state. They also looked at maps proposed by Democratic interest groups during the redistricting process, and not even these maps would have resulted in

Figure 1.7: Over representation by state in the 2012 election. Calculated as the difference between the number of seats a party won, versus how many they would have won proportional to the percentage of votes that party received. California has 5 extra democratic representatives, Ohio has 4 extra Republicans.



Democrats taking the same share of seats as they won in votes[10]. While you could still see eliminating bias as a goal of a redistricting map, the paper complicates its use as an indicator of gerrymandering.

1.4 Conclusion

Mathematics can provide no silver bullet to the redistricting problem. There are too many conflicting interests at play to even compare various maps, let alone create an optimal map. Measuring districts for gerrymandering is difficult, because even if the results end up being biased towards one party, that does not mean that was the intention of those that drew the map. Even randomly generated maps won't eliminate complaints of bias, because even then geographic dispersal of votes can cause a party to get an artificially large portion of the seats.

From the mathematical point of view though, there are a lot of interesting questions raised by the redistricting problem. Further investigation could be done comparing the

different districting algorithms to determine if they produce similar results. Improvements could be made to the algorithms to insure that population is split between the districts more evenly. Also, the relationship between clustered votes and poor representation is worth further investigation. These might not be a panacea for the issue of gerrymandering, but they can help inform the choices that future politicians and cartographers make.

Chapter 2

CONVEXITY MEASURES

2.1 Measuring Convexity

There is no way to decide the perfect district then, but maybe there are ways to detect intentionally Gerrymandered districts. One approach is to try to detect unnaturally odd looking districts. In the simplified state, it was very easy to pack the Blue voters into a few districts, and then spread the remaining voters evenly through the rest of the districts. In an actual state, an evil-minded cartographer might have to make very strange looking districts to accomplish these aims, whereas there would be no reason to do so without any knowledge of the political landscape. There have been many proposals for ways of measuring convexity of a district, some of which are discussed below.

2.2 Area Ratio

A common convexity measure is to look at the ratio of the area to the perimeter. This is a very flawed measure though, because the length of the perimeter of a state depends on how fine grain your measurements are. Also, the measure is not scale invariant, as using units half as long, will double the perimeter, but quadruple the area. A convexity measure should not depend on what units are used.

Another, related class of convexity measures compares the ratio of the area of the polygon to the area of some bounding figure, such as the bounding box, the smallest circumscribing circle, or the convex hull. Of these measures, taking the ratio of the area of the district to the area of the convex hull makes the most sense. Comparing the area of the figure to the area of the bounding box is easy to implement, but it is not rotation invariant. Taking the smallest bounding circle is consistent, but is harder to implement, and is harder to justify. Should a district shaped like a square or a triangle really be rated lower than a circle? To see how Maryland's congressional districts fare by this measure, see Figure 2.3.

Definition 2.2.1. ***Area Ratio Convexity Measure*** *[2]: Let D be a district in a state S. Let $v(D)$ be the ratio of the area of D to the area of the convex hull of D.*

Algorithm 1 Monte Carlo Estimate of Modified Path-Based Measure

```
Maryland.113 <- CleanSubset(subset(Dist.113.shp, STATE == "Maryland"))
Maryland.113$HULL_RATIO <- 0
for(i in 1:8){
    dist = subset(Maryland.113, CONG_DIST == Maryland.113$CONG_DIST[i])
    distHull <- gConvexHull(dist)
    Maryland.113$HULL_RATIO[i] <- gArea(dist)/gArea(distHull)
}
```

2.3 Path-Based Measure

Another convexity measure was introduced by Chambers and Miller in 2012. It looks at the probability that the shortest path joining any two points in the district lies within the district.

Definition 2.3.1. ***Path Based Convexity Measure*** *[9]: Let D be a district in a state S, then define $\chi(D)$ as the probability that the shortest path connecting any two points p_x, p_y within the state S is contained entirely within D.*

Specifying the path between two points be the shortest path connecting two points within the state helps compensate for intrusions in the state's boundary. For instance, the shortest path connecting two points in the state of Maryland may pass through the Chesapeake Bay. This should not count against its convexity score of the district, since it is a feature of the state. It also allowed for extensions to the measure where one compared the ratio of the shortest path within the state to the shortest path within the district. However, the measure is undefined if the district is not contiguous. For example, Washington's Second Congressional District contains islands in the Puget Sound. There is no path within the state, or within the district, connecting a point on an island to a point on the mainland. An alternative introduced by Hodge, Marshall, and Patterson that fixes this problem[13].

Definition 2.3.2. ***Modified Path Based Convexity Measure*** $\widetilde{\chi}(D)$***[13]:*** *Let D be a district in a state S, then $\widetilde{\chi}(D)$ is the probability that the line connecting any two random points in D intersects the boundary of the district D only where it intersects the boundary of the state S.*

Note that by this definition, any states with only 1 congressional district will have a Modified Convexity Measure of 1. This is reasonable, since the district is as convex as it can be.

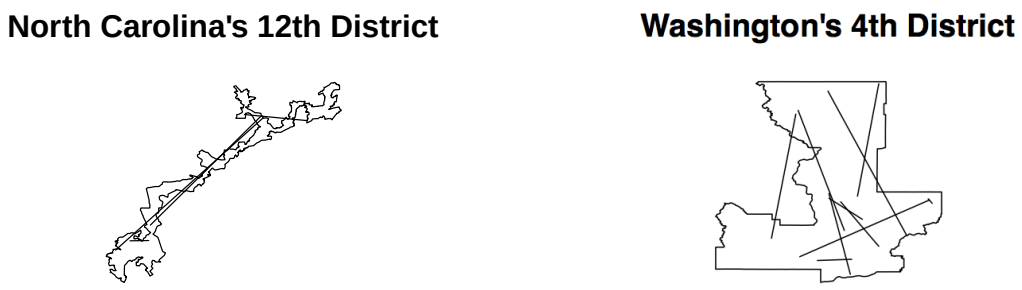


Figure 2.1: Randomly generated paths. All of the paths in Carolina's 12th pass out of the district, where as only 3 of the paths pass out of Washington's 4th.

This measure, even as modified, would be hard to compute explicitly, as shown by Hodge, Marshall, and Patterson [13]. However, a Monte Carlo approximation is easy to implement. To do so, let X be the random variable that returns 0 if the line connecting two points chosen uniformly at random from K passes through another district in the state, and 1 if it does not. This is a Bernoulli($\widetilde{\chi}(K)$) variable, so $\widetilde{\chi}(K)$ can be approximated to arbitrary precision using standard Bernoulli trial statistics. The R code for this is below, as well as some sample results.

Another modification discussed by Chambers and Miller is to sample points based on population instead of sampling uniformly. This can be accomplished either by using a population density map to create a sampling density over the state, or by considering the

Algorithm 2 Monte Carlo Estimate of Modified Path-Based Measure

```
#creates n random points within a figure
randomPointsInFig <- function(figure,n){
    pts <- SpatialPoints(cbind(0,0))
    #creates n random points contained in the district
    while(length(pts) < n){
        #Creates Random points within bounding box of figure
        x <- runif(n*5, min = bbox(figure)[1,1], max = bbox(figure)[1,2])
        y <- runif(n*5, min = bbox(figure)[2,1], max = bbox(figure)[2,2])
        pts = SpatialPoints(cbind(x,y))
        #subsets the points only within the figure
        pts <- pts[figure]
    }
    return(pts[1:n])
}
#creates n random lines within a figure. Returns as SpacialLines object
RandomLinesInFig <-function(figure, n){
    pts <- randomPointsInFig(figure,2*n)
    i = 1
    lns = list(rep(0,n))
    while(i <=length(pts)/2){
        lns[i] <- Lines(Line(coordinates(pts[(2*i-1):(2*i)])), paste0("a",i))
        i <-i+1
    }
    return(SpatialLines(lns)) }
```

Algorithm 3 Monte Carlo Estimate of Modified Path-Based Measure

```
pathMeasure <- function(file.shp, state, distnum){
    n <- 1000
    dist <- subset(file.shp, STATE == state & CONG_DIST == distnum)
    complement <- subset(file.shp, STATE == state & CONG_DIST != distnum)
    complement <- gUnaryUnion(complement)
    splns <- RandomLinesInFig(dist,n)
    containsLines <- rep(0,n)
    for(i in 1:n){
        containsLines[i] <-!gCrosses(complement,splns[i])
    }
    k <-sum(containsLines)
    p <-k/n
    return(c(p, sqrt(k * (n - k)/(n * n * (n - 1))) * qt(.025, n - 1)))
}
```

state as a collection of census blocks. Then, instead of sampling points in the state at random, the algorithm would sample census blocks at random, and calculate wheather the centroid of the census blocks are connected.

2.4 Limitations

One problem with this line of investigation is that it is difficult to hold up this evidence in court, as there could be other motives that would create artificial looking districts. For instance, below left is Maryland's 2nd Congressional District in the 113th congress, and right is New York's 10th district from the 113th congress[19]. Both districts look odd, and both would score poorly on convexity measures. However, Maryland's 2nd boarders the Chesapeake Bay, which accounts for a lot of the squigglyness. On the other hand, North Carolina's 12th congressional district was created in an odd shape in part as an attempt to give black voters in the state more representation[11]. The percent of people

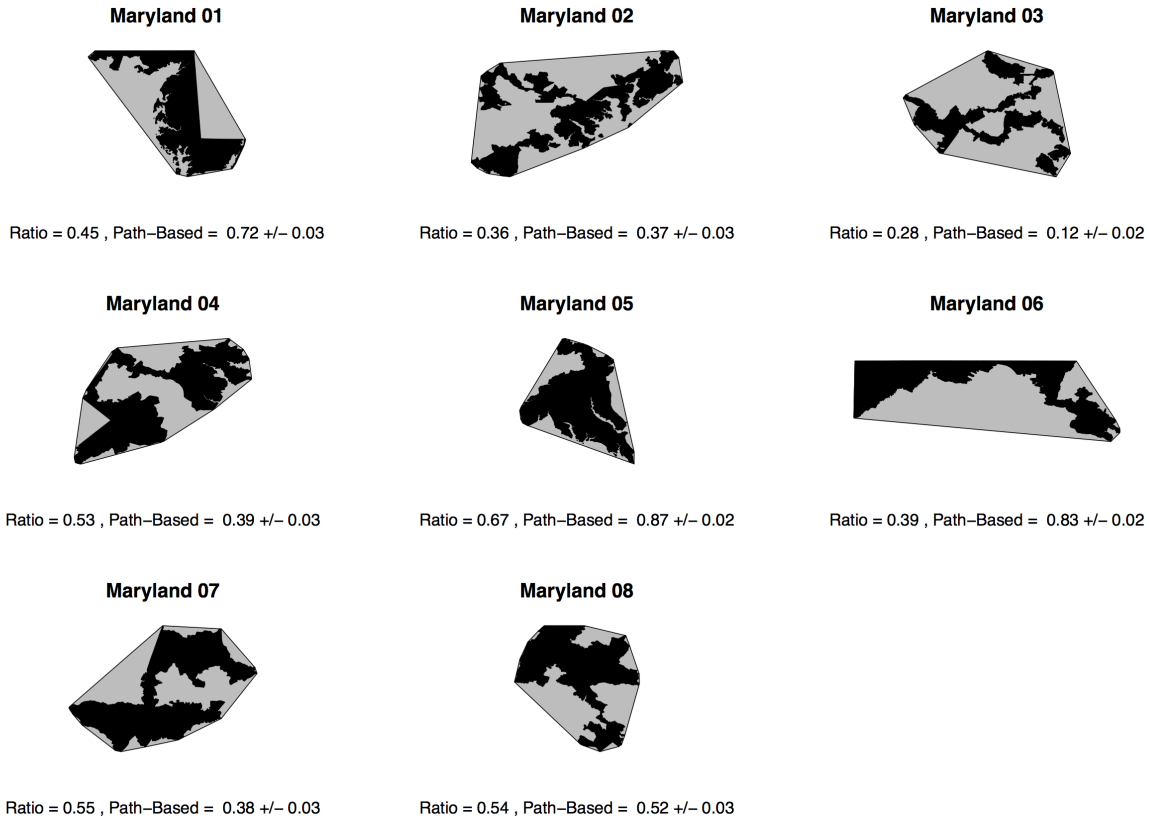


Figure 2.2: The districts of Maryland. The ratio convexity measure from section 2.2 compares the ratio of the district with its convex hull. The Path-based measure from section 2.3 takes the probability that a random path in the district will not pass through a different district. Displayed is a Monte Carlo estimate of the path based measure, and 95% confidence interval.

identifying as Black or African American in North Carolina is approximately 22 percent, which would suggest that at at least 2 of the 13 congressional districts should be carried by black representatives (approximately 15 percent of the seats).

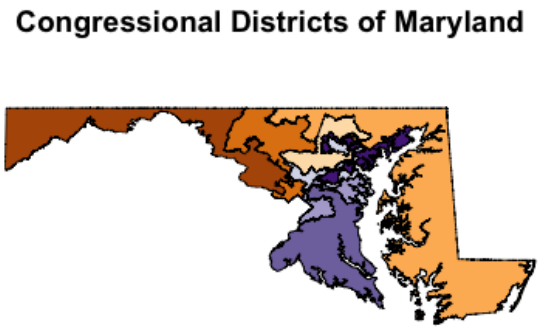
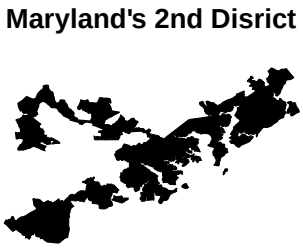


Figure 2.3: The map of the congressional districts of Maryland for the 113th Congress



There are ways to accommodate geographic features interfering with the boundaries of districts, but that still does not address the situation in North Carolina's 12th district. There, the boundaries of the district are manipulated, but for arguably acceptable reasons. There is no way to tell just by looking at the shape of a district what were the intentions of those creating the districts. All that the convexity analysis can do is indicate when a district is likely based on non-geographic considerations, and then leave it up to the courts to decide what the intentions where, and whether those intentions are acceptable.

Another example of the conflict that comes from measuring manipulation, Washington state gained a congressional representative after the 2010 census. Some of the district maps

proposed to create a majority-minority district. However, there have been accusations by Republicans that this is actually motivated by Democrats attempting to spread the largely Democratic Seattle vote into two separate districts, instead of one district as it is now [5], whereas some Democrats are accusing similar plans as being efforts by conservatives to pack democratically voting minorities into a single district, so as to dilute their influence elsewhere [21]. Measuring the convexity of the district is no help in this case, because math alone can not conclude which motivations are acceptable or not.

Chapter 3

AUTOMATIC DISTRICTING

3.1 Uses of Algorithms

Another approach to preventing gerrymandering is to remove the human element from the districting process entirely, and let computer algorithms draw the districts. Computers are becoming more and more powerful, as is the geographic analysis software, so this is a realistic option now. A challenge to this idea is that the districts are not just random divisions of a state, but are supposed to represent actual natural divisions within the state, and that some respect should be paid to previous congressional districts. In this case, there are also algorithms that could modify current districts for equal population constraints.

Even using automatic districting, the question arises how to choose an algorithm. The parties involved might just fight for the algorithm that best favors their interests. Ideally, the parties would argue over algorithms unaware of which system favors their own party, as "it should not be possible for political actors to deduce the results of the redistricting goals over which they bargain"[2]. One step toward accomplishing this it to use an unpredictable algorithm, meaning that one should not be able to determine the result of the algorithm before it is run. This can be accomplished by having the algorithm be highly sensitive to some initial condition, and have that initial condition be randomized at every trial, or by having some randomization within the algorithm. The results of different Algorithms may still vary though, so it is important to fully analyze different districting algorithms.

Another proposed use of districting algorithms is as a standard to compare the chosen districting plans against. In section 1.3 we compared election results against what the results would be if the vote was proportional, but there is no a priori reason to believe that election results should end up being proportional. Looking at the results that are generated by automatic districting plans should indicate whether there are geographic factors contributing to any observed bias. If 1000 plans are generated, then using bootstrap methods,

a distribution of election results could be approximated. Then, the actual result could be compared to the distribution, giving a sense of how abnormal the result is. This was a tactic used in 2004 by McCarty et al [17], and more recently by Chen and Rodden [10].

3.2 Districting Algorithms

In their paper "Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures" Jowei Chen and Jonathan Rodden used a districting algorithm that grouped nearby voting precincts together. A precinct is the finest geographic division at which votes are recorded. In states that have not moved to exclusively mail in ballets, the vote is administered and counted at the precinct level. Each precinct has a polling place, and workers at each polling place record and report the vote for their precinct. Many states release the vote totals at the precinct level, so for analysis that uses both election results and election maps, precincts are often the most convenient unit to think about.

Algorithm 4 Chen and Rodden's Districting Algorithm (Pseudocode)

S = State

D = The set of Districts in S

d = desired number of districts

$p(i)$ = Population of a district or state i

$d(i, j)$ = Centroidal distance between two districts i and j

While $(D) > d$ {

    Select a district i from D

    Select j in D such that j boarders i and $d(i, j)$ minimal

    $D = D\{i, j\}$

    $i = i \cup j$

    $D = D \cup \{i\}$

}

This results in d districts, all contiguous and relatively compact. However, populations of each district may vary dramatically. A hill-climbing algorithm must be employed to

improve population balance. The algorithm Chen and Rodden used is shown in Algorithm 5.

Algorithm 5 Chen and Rodden's Hill-Climb Algorithm (Pseudocode)

S = State

D = The set of Districts in S

P = The set of Precincts in S

$p(i)$ = Population of i

$d(i, j)$ = Centroidal distance between i and j

Until (population of all districts within 5% of $(p(S)/d)$){

Find pair i, j in D such that i, j boarder, $p(i) > p(j)$, and $p(i) - p(j)$ maximal.

Let B_{ij} be the precincts p in i such that p can be switched from i to j without violating contiguity of i or j.

Find p in B_{ij} such that $d(i, p) - d(j, p)$ is maximal

Reassign p to j.

}

This results in fairly compact districts, because both the original algorithm and the hill-climb consider centroidal distance when choosing precincts. They also created districts without choosing the closest precinct, resulting in districts were less compact, but they did not observe significant differences in results. Other modifications that could be used are to use a different distance function, like population-weighted centroid, or combine the original algorithm with a different hill-climb algorithm.

Another approach to districting is to use Voronoi Diagrams. Voronoi Diagrams break up a state by first plotting d points in the state, and then breaking up the state into d regions, where each region consists of all of the points in the state closest to a particular initial point. The placement of the initial points can then be moved based on an algorithm to achieve a more balanced population distribution amongst the districts. This approach was used by Stacy Miller, based on MacQueens K-means algorithm for Voronoi Diagrams[18].

This makes very nice looking districts, but they do not end up having very balanced

Algorithm 6 Miller's Modified MacQueen's Method (Pseudocode)

d = desired number of districts

n = number of times to iterate process (arbitrary).

$z_1, z_2, ...z_d$ = Centroids of largest population centers in state.

$j_1, j_2, ...j_d = 1$

for (i = 1:n){

1. Select a precint t randomly, where the probability of p being chosen is $p(t)/p(S)$.
2. Find closest centroidal point z_i to t.
3. $z_i = \frac{j_i z_i + x}{j_i + 1}$
4. $j_i = j_i + 1$

}

For each precinct p find closest point z_i. Assign p to corresponding district d_i.

populations at the end. Another hill-climb algorithm to balance out population would need to be done after the districts are created.

A third approach by Burden et. al.[6] uses what they refer to as "Voronoiesque Diagrams" to create districts. An equivalent description of Voronoi Diagrams is to take d initial points, and then plot circles around each point with radiuses small enough that the circles do not intersect. Then, allow the circles to grow at a constant rate, but where ever the expanding circles meet they must stop growing. This ends up dividing the state into regions grouped by closest point. Burden et. al.[6] modified this algorithm so that the circles grew at a rate proportional to their contained population.

3.3 Further Research

Chen and Rodden used their algorithm to create 100 simulated districting plans, each with 25 districts, and then used the results of the 2000 presidential election to create fake Congressional elections. The vote was very evenly divided between George Bush and Al

Gore in that election, so proportionally, each party should have won 12-13 representatives each. Of the 100 simulations, not a single one gave even 12 representatives to Democrats. Democrats received between 7-11 districts, with the median being 9. Using these votes, the actual districting plan would have rewarded Democrats 8 representatives. This leans slightly more towards the republicans than the mean, but is within the of the randomized algorithm's standard deviation.

More research needs to be done similar to Chen and Rodden work. It allows an evaluation of proposed automated districting methods, and also acts a better benchmark for gerrymandering than comparing to the proportional result. However, some care needs to be taken when interpreting these results. By comparing the results of a randomized algorithm to the actual plan, one is implicitly treating the actual plan as a random variable X selected from the space of all districting plans, and that the results of the randomized algorithm approximate the distribution of X. However, the algorithms do not sample uniformly from the space of all possible districting plans, nor do they represent the types of plans that the humans making the plans were considering when drawing the districts. Used as hypothesis testing, the most one can say is whether or not it is the type of district that could have been created using the randomized districting method. Thus, as with all of the methods considered in this thesis, automatic districting methods should be considered primarily an indicator of gerrymandering, not as proof.

Appendix A

SOFTWARE AND DATA

To actually implement any of the algorithms in the previous chapter, one first needs the ability to work with geographic data. After all, how can one measure the perimeter of a congressional district without first knowing how to access data on the boundary of the congressional district? In this chapter I hope to outline the software and data sources I used in hopes that it will assist future researchers implement their algorithms.

A.1 File Types

There are two main categories of file types for geographic data: vector files and raster files. Raster files store information as a grid. For example, you may have a map of Washington with each pixel representing the approximate population averaged over 100 square miles, where a color encodes the population. This would be raster data. It is easy to see and interpret visually, there is also software that can use this data for calculations. Vector files store data on geographic features, considering each feature as a geometric object, either a point, line, or polygon. For example, to store population data in a vector file, you would need to choose geographic units to partition the land into, such as counties. The vector file would store the shape and population of each county. Given that congressional districts are an innately geometric question, vector files make more sense for gerrymandering research.

There are several different file types for storing geographic vector data. However, the easiest file type to find data in is the esri Shapefile. While created by esri, a private company, the shapefile is an open standard. This means that the full details of how the data is stored is publicly available, allowing researchers to know exactly what the software is doing, and how the data is being stored. This helps the researcher modify programs to fit their needs, create new programs around the files, and to know when to attribute unexpected results to the software or reality. It also allows us to use the R statistical software to analyze the

data, which is an open source statistical software (see Section A.2). Other geographic vector file types of interest are the Geographic Markup Language (GML) designed by the Open Geospatial Consortium (OGC), and the related Keyhole Markup Language (KML) used in Google Maps. GML and KML are much more recent standards, and are more compatible with html. However, shapefiles are the historical staple of geographic data, so for now most election data you find will be in shapefile format.

The shapefile is actually a set of at least three files: a .shp file, a .shx file, and a .dbf file. All files should be stored in the same folder. The .shp file is the main file, and it stores the list of vertices of each shape in your file. The .shx file stores index information on the .shp file, allowing the software to manage the information stored in the .shp file more smoothly. The .dbf, or dBase, file stores the attributes of each feature. Other files, such as the .prj file and the .sbn file, are optional. The .prj file contains map projection information, and the .sbn file is used only for the esri ArcGIS software. Whatever software you use will work with all the file types simultaneously. As long as all of the files are in the same directory, you should only have to load the .shp file and the software will do the rest.

A.2 Software

R is a free, open source software package for statistical computing and data analysis. I suggest using the RStudio interface, which can be downloaded at rstudio.com[1]. The following is a summary of useful packages for manipulating shapefiles.

A.3 Data Sources

To analyze the shape of the districts, the best place to get data is NationalAtlas.gov. NationalAtlas.gov is led by the United States Geological Survey, and aims to provide public access to the massive amount of geographic data collected for the United States. It does not have precinct data, but it does have maps of all the congressional districts from the 106th congress in 1999 to the present.

Precinct maps and election results are handled on a state-to-state basis. For most states, election results can be found on the Secretary of State's webpage, but not normally in a format that is easy to work with. There is a site called The Harvard Election Data Archive

(http://projects.iq.harvard.edu/eda) that has precinct level voting data for many states, but the data is not well documented and often doesn't load in R.

Appendix B

R CODE

B.1 Working with Shapefiles

Libraries required for my code are rgdal, rgeos, maptools, spdep, spatstat, and RColorBrewer.

After downloading your data, keep the various files in the same directory. There are several equivalent way so loading the file into R: setwd("directorypath") sets your working directory. Then readShapePoints("foo.shp"), readShapeLines("foo.shp"), or readShapePoly("foo.shp") will load the data interpreting it as point, line, or polygon data respectively. The command readShapeSpatial("foo.shp") will also work, and allows R to interpret the data type for you. You can also use file.choose() in place of "foo.shp" to open a file navigation window to select the file. While all the files in the directory will be used and interpreted, only select the main .shp file.

Another command for loading files using the rgdal package is readOGR(directoryname, filename). In this case do not put the .shp at the end of the file name. Example: you wish to load the file foo contained in the folder bar which is in your working directory. Load the data using the command foo $<-$ readOGR("./bar", "foo").

For the following code, I began with the 113th congressional districts downloaded from NationalAtlas.gov [19], and used saved it as Dist.113.shp on my workspace.

To simplify your data, you can delete attributes, and select just a subset of the data to deal with. You can also create new attributes for the data set.

Algorithm 7 Examining Data

```
setwd("Directory Name Here")
#Opens a file navigation. Select the .shp file of the file you wish to use.
Dist.113.shp <- readShapeSpatial(file.choose())
#A summary of the file
summary(Dist.113.shp)
#List of attributes
names(Dist.113.shp)
[1] "STATE" "STATE_FIPS" "CONG_DIST" "CONG_REP" "PARTY_AFF" "URL"
[7] "SENATOR_1" "SEN1_PARTY" "SEN_1_URL" "SENATOR_2" "SEN2_PARTY"
"SEN_2_URL"
#Plots all districts
plot(Dist.113.shp)
# Top results in attibute
head(Dist.113.shp$STATE)
Alabama Alabama Alabama Alabama Alabama Alabama
53 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
head(Dist.113.shp$CONG_REP, 3)
[1] Mo Brooks Robert B. Aderholt Terri A. Sewell
435 Levels: Aaron Schock Adam B. Schiff Adam Kinzinger Adam Smith ... Zoe Lofgren
#To get the bounding box of the file
bbox(Dist.113.shp)
min max
x -179.1473 179.77848
y 17.6744 71.38921
```

Algorithm 8 Examining Data

```
# Deletes useless attributes
Dist.113.shp$STATE_FIPS <- Dist.113.shp$URL <- Dist.113.shp$SENATOR_1
<- Dist.113.shp$SEN1_PARTY <- Dist.113.shp$SEN1_PARTY
<- Dist.113.shp$SEN_1_URL <- Dist.113.shp$SENATOR_2 <-
Dist.113.shp$SEN2_PARTY <- Dist.113.shp$SEN_2_URL <- NULL
#Removes some useless information from the file
Dist.113.shp$STATE_FIPS <- Dist.113.shp$URL <- Dist.113.shp$SENATOR_1
<- Dist.113.shp$SEN1_PARTY <- Dist.113.shp$SEN1_PARTY
<- Dist.113.shp$SEN_1_URL <- Dist.113.shp$SENATOR_2 <-
Dist.113.shp$SEN2_PARTY <- Dist.113.shp$SEN_2_URL <- NULL
#Adding an attribute, to be filled in later.
Dist.113.shp$CONVEXITY_MEASURE <- 0
#A function that removes unwanted levels from a district or state, once subsetted
CleanSubset <- function(Dist){
        Dist$STATE <- factor(Dist$STATE)
        Dist$CONG_DIST <- factor(Dist$CONG_DIST)
        Dist$CONG_REP <- factor(Dist$CONG_REP)
        return(Dist)
}
#Creates a shapefile for the state of Maryland, and one for each district.
Maryland.113.shp <- subset(Dist.113.shp, STATE == "Maryland")
Maryland.113.shp <- CleanSubset(Maryland.113.shp)
MD2 <- CleanSubset(subset(Maryland.113.shp, CONG_DIST == "02"))
plotclr <- brewer.pal(8,"PuOr")
plot(Maryland.113.shp, col = plotclr)
title("Congressional Districts of Maryland")
plot(MD2, col = plotclr[2])
title("Maryland's First Congressional District")
```

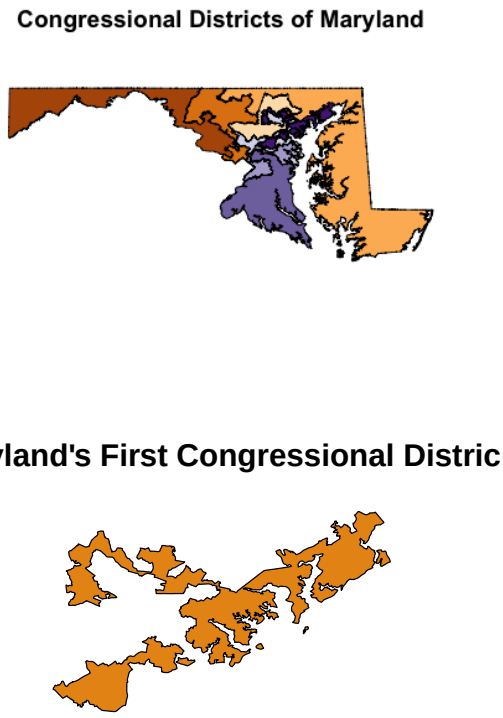


Figure B.1: Plots produced by Code Chunk 8

Appendix C

CONVEXITY MEASURES, ALPHABETICAL

(Hawaii is excluded)

State	District	Hull Ratio	Path Probability	95% confidence error
Alabama	01	0.65	0.69	0.03
Alabama	02	0.74	0.74	0.03
Alabama	03	0.73	0.87	0.02
Alabama	04	0.62	0.64	0.03
Alabama	05	0.78	0.85	0.02
Alabama	06	0.68	0.49	0.03
Alabama	07	0.62	0.8	0.02
Alaska	01	0.07	1	0
Arizona	01	0.74	0.82	0.02
Arizona	02	0.88	0.96	0.01
Arizona	03	0.75	0.83	0.02
Arizona	04	0.62	0.59	0.03
Arizona	05	0.88	0.96	0.01
Arizona	06	0.82	0.92	0.02
Arizona	07	0.84	0.94	0.01
Arizona	08	0.66	0.68	0.03
Arizona	09	0.54	0.54	0.03
Arkansas	01	0.7	0.75	0.03
Arkansas	02	0.71	0.78	0.03
Arkansas	03	0.53	0.44	0.03
Arkansas	04	0.8	0.83	0.02

State	District	Hull Ratio	Path Probability	95% confidence error
California	01	0.87	0.83	0.02
California	02	0.59	0.81	0.02
California	03	0.62	0.75	0.03
California	04	0.82	0.91	0.02
California	05	0.7	0.78	0.03
California	06	0.67	0.72	0.03
California	07	0.89	0.95	0.01
California	08	0.81	0.93	0.02
California	09	0.83	0.93	0.02
California	10	0.76	0.84	0.02
California	11	0.77	0.89	0.02
California	12	0.68	0.88	0.02
California	13	0.81	0.98	0.01
California	14	0.3	0.92	0.02
California	15	0.78	0.92	0.02
California	16	0.74	0.8	0.02
California	17	0.76	0.87	0.02
California	18	0.81	0.84	0.02
California	19	0.78	0.8	0.02
California	20	0.85	0.94	0.01
California	21	0.72	0.78	0.03
California	22	0.56	0.66	0.03
California	23	0.72	0.7	0.03
California	24	0.52	0.92	0.02
California	25	0.68	0.68	0.03
California	26	0.32	0.83	0.02
California	27	0.75	0.86	0.02
California	28	0.71	0.63	0.03

State	District	Hull Ratio	Path Probability	95% confidence error
California	29	0.69	0.8	0.02
California	30	0.66	0.73	0.03
California	31	0.58	0.49	0.03
California	32	0.66	0.73	0.03
California	33	0.42	0.84	0.02
California	34	0.73	0.74	0.03
California	35	0.61	0.7	0.03
California	36	0.94	0.99	0.01
California	37	0.81	0.92	0.02
California	38	0.72	0.79	0.03
California	39	0.75	0.85	0.02
California	40	0.65	0.66	0.03
California	41	0.75	0.73	0.03
California	42	0.62	0.6	0.03
California	43	0.72	0.84	0.02
California	44	0.65	0.82	0.02
California	45	0.86	0.88	0.02
California	46	0.73	0.87	0.02
California	47	0.13	0.69	0.03
California	48	0.68	0.76	0.03
California	49	0.59	0.81	0.02
California	50	0.8	0.89	0.02
California	51	0.74	0.84	0.02
California	52	0.55	0.69	0.03
California	53	0.7	0.63	0.03

State	District	Hull Ratio	Path Probability	95% confidence error
Colorado	01	0.54	0.44	0.03
Colorado	02	0.76	0.81	0.02
Colorado	03	0.79	0.84	0.02
Colorado	04	0.81	0.84	0.02
Colorado	05	0.8	0.94	0.01
Colorado	06	0.56	0.4	0.03
Colorado	07	0.63	0.72	0.03
Connecticut	01	0.67	0.43	0.03
Connecticut	02	0.82	0.93	0.02
Connecticut	03	0.68	0.8	0.02
Connecticut	04	0.66	0.88	0.02
Connecticut	05	0.75	0.71	0.03
Delaware	01	0.73	1	0
Florida	01	0.76	0.99	0.01
Florida	02	0.72	1	0
Florida	03	0.78	0.88	0.02
Florida	04	0.71	0.66	0.03
Florida	05	0.28	0.22	0.03
Florida	06	0.66	0.88	0.02
Florida	07	0.76	0.67	0.03
Florida	08	0.63	0.83	0.02
Florida	09	0.8	0.94	0.02
Florida	10	0.73	0.8	0.02
Florida	11	0.69	0.83	0.02
Florida	12	0.78	0.93	0.02
Florida	13	0.56	0.94	0.02
Florida	14	0.42	0.86	0.02
Florida	15	0.75	0.88	0.02

State	District	Hull Ratio	Path Probability	95% confidence error
Florida	16	0.7	0.86	0.02
Florida	17	0.82	0.94	0.01
Florida	18	0.78	0.97	0.01
Florida	19	0.48	0.84	0.02
Florida	20	0.74	0.88	0.02
Florida	21	0.61	0.72	0.03
Florida	22	0.43	0.38	0.03
Florida	23	0.46	0.76	0.03
Florida	24	0.69	0.82	0.02
Florida	25	0.73	0.75	0.03
Florida	26	0.25	0.98	0.01
Florida	27	0.44	0.88	0.02
Georgia	01	0.72	0.9	0.02
Georgia	02	0.83	0.92	0.02
Georgia	03	0.79	0.91	0.02
Georgia	04	0.8	0.91	0.02
Georgia	05	0.85	0.97	0.01
Georgia	06	0.67	0.84	0.02
Georgia	07	0.74	0.74	0.03
Georgia	08	0.67	0.62	0.03
Georgia	09	0.84	0.98	0.01
Georgia	10	0.81	0.94	0.01
Georgia	11	0.75	0.82	0.02
Georgia	12	0.68	0.78	0.03
Georgia	13	0.61	0.58	0.03
Georgia	14	0.78	0.76	0.03
Idaho	01	0.74	0.76	0.03
Idaho	02	0.81	0.97	0.01

State	District	Hull Ratio	Path Probability	95% confidence error
Illinois	01	0.63	0.51	0.03
Illinois	02	0.81	0.86	0.02
Illinois	03	0.67	0.74	0.03
Illinois	04	0.42	0.23	0.03
Illinois	05	0.49	0.4	0.03
Illinois	06	0.57	0.37	0.03
Illinois	07	0.5	0.5	0.03
Illinois	08	0.6	0.69	0.03
Illinois	09	0.6	0.53	0.03
Illinois	10	0.63	0.7	0.03
Illinois	11	0.55	0.32	0.03
Illinois	12	0.65	0.78	0.03
Illinois	13	0.56	0.64	0.03
Illinois	14	0.65	0.58	0.03
Illinois	15	0.7	0.78	0.03
Illinois	16	0.67	0.7	0.03
Illinois	17	0.54	0.81	0.02
Illinois	18	0.74	0.55	0.03
Indiana	01	0.76	0.95	0.01
Indiana	02	0.83	0.88	0.02
Indiana	03	0.9	0.98	0.01
Indiana	04	0.82	0.93	0.02
Indiana	05	0.78	0.9	0.02
Indiana	06	0.86	0.96	0.01
Indiana	07	0.93	0.98	0.01
Indiana	08	0.69	0.97	0.01
Indiana	09	0.8	0.88	0.02

State	District	Hull Ratio	Path Probability	95% confidence error
Iowa	01	0.67	0.77	0.03
Iowa	02	0.73	0.86	0.02
Iowa	03	0.83	0.92	0.02
Iowa	04	0.88	0.98	0.01
Kansas	01	0.88	0.85	0.02
Kansas	02	0.74	0.86	0.02
Kansas	03	0.85	0.94	0.01
Kansas	04	0.86	0.83	0.02
Kentucky	01	0.62	0.58	0.03
Kentucky	02	0.64	0.68	0.03
Kentucky	03	0.78	0.98	0.01
Kentucky	04	0.54	0.65	0.03
Kentucky	05	0.78	0.81	0.02
Kentucky	06	0.76	0.83	0.02
Louisiana	01	0.4	0.56	0.03
Louisiana	02	0.38	0.38	0.03
Louisiana	03	0.79	0.96	0.01
Louisiana	04	0.61	0.72	0.03
Louisiana	05	0.57	0.88	0.02
Louisiana	06	0.62	0.32	0.03
Maine	01	0.46	0.7	0.03
Maine	02	0.82	0.94	0.01
Maryland	01	0.45	0.68	0.03
Maryland	02	0.36	0.34	0.03
Maryland	03	0.28	0.13	0.02
Maryland	04	0.53	0.39	0.03
Maryland	05	0.67	0.88	0.02

State	District	Hull Ratio	Path Probability	95% confidence error
Maryland	06	0.39	0.86	0.02
Maryland	07	0.55	0.42	0.03
Maryland	08	0.54	0.49	0.03
Massachusetts	01	0.75	0.81	0.02
Massachusetts	02	0.77	0.85	0.02
Massachusetts	03	0.68	0.81	0.02
Massachusetts	04	0.65	0.67	0.03
Massachusetts	05	0.61	0.71	0.03
Massachusetts	06	0.68	0.84	0.02
Massachusetts	07	0.41	0.35	0.03
Massachusetts	08	0.57	0.47	0.03
Massachusetts	09	0.37	0.91	0.02
Michigan	01	0.43	0.99	0.01
Michigan	02	0.83	0.92	0.02
Michigan	03	0.74	0.7	0.03
Michigan	04	0.81	0.91	0.02
Michigan	05	0.58	0.51	0.03
Michigan	06	0.87	0.99	0.01
Michigan	07	0.7	0.75	0.03
Michigan	08	0.76	0.78	0.03
Michigan	09	0.69	0.6	0.03
Michigan	10	0.85	0.99	0.01
Michigan	11	0.64	0.59	0.03
Michigan	12	0.65	0.52	0.03
Michigan	13	0.62	0.48	0.03
Michigan	14	0.39	0.32	0.03

State	District	Hull Ratio	Path Probability	95% confidence error
Minnesota	01	0.83	0.9	0.02
Minnesota	02	0.74	0.91	0.02
Minnesota	03	0.77	0.85	0.02
Minnesota	04	0.88	0.98	0.01
Minnesota	05	0.83	0.94	0.02
Minnesota	06	0.62	0.75	0.03
Minnesota	07	0.73	0.87	0.02
Minnesota	08	0.75	0.98	0.01
Mississippi	01	0.81	0.82	0.02
Mississippi	02	0.78	0.86	0.02
Mississippi	03	0.63	0.64	0.03
Mississippi	04	0.82	0.96	0.01
Missouri	01	0.72	0.86	0.02
Missouri	02	0.76	0.87	0.02
Missouri	03	0.77	0.88	0.02
Missouri	04	0.68	0.71	0.03
Missouri	05	0.69	0.54	0.03
Missouri	06	0.76	0.83	0.02
Missouri	07	0.82	0.89	0.02
Missouri	08	0.81	0.99	0.01
Montana	01	0.93	1	0
Nebraska	01	0.83	0.85	0.02
Nebraska	02	0.86	0.95	0.01
Nebraska	03	0.83	0.93	0.02
Nevada	01	0.9	0.95	0.01
Nevada	02	0.9	0.97	0.01
Nevada	03	0.9	0.97	0.01
Nevada	04	0.87	0.96	0.01

State	District	Hull Ratio	Path Probability	95% confidence error
New Hampshire	01	0.65	0.66	0.03
New Hampshire	02	0.7	0.64	0.03
New Jersey	01	0.69	0.82	0.02
New Jersey	02	0.75	0.84	0.02
New Jersey	03	0.56	0.66	0.03
New Jersey	04	0.71	0.77	0.03
New Jersey	05	0.59	0.64	0.03
New Jersey	06	0.43	0.57	0.03
New Jersey	07	0.7	0.82	0.02
New Jersey	08	0.47	0.24	0.03
New Jersey	09	0.59	0.59	0.03
New Jersey	10	0.53	0.4	0.03
New Jersey	11	0.73	0.79	0.03
New Jersey	12	0.63	0.71	0.03
New Mexico	01	0.71	0.71	0.03
New Mexico	02	0.85	0.9	0.02
New Mexico	03	0.79	0.88	0.02
New York	01	0.45	0.98	0.01
New York	02	0.64	0.96	0.01
New York	03	0.75	0.94	0.01
New York	04	0.75	0.94	0.01
New York	05	0.56	0.88	0.02
New York	06	0.8	0.83	0.02
New York	07	0.41	0.38	0.03
New York	08	0.42	0.61	0.03
New York	09	0.68	0.87	0.02
New York	10	0.34	0.28	0.03

State	District	Hull Ratio	Path Probability	95% confidence error
New York	11	0.69	0.95	0.01
New York	12	0.66	0.88	0.02
New York	13	0.63	0.67	0.03
New York	14	0.47	0.58	0.03
New York	15	0.86	0.97	0.01
New York	16	0.76	0.9	0.02
New York	17	0.82	0.96	0.01
New York	18	0.72	0.82	0.02
New York	19	0.76	0.82	0.02
New York	20	0.77	0.9	0.02
New York	21	0.91	0.96	0.01
New York	22	0.65	0.8	0.02
New York	23	0.77	0.86	0.02
New York	24	0.75	0.89	0.02
New York	25	0.84	0.98	0.01
New York	26	0.74	0.96	0.01
New York	27	0.77	0.84	0.02
North Carolina	01	0.49	0.48	0.03
North Carolina	02	0.72	0.58	0.03
North Carolina	03	0.5	0.49	0.03
North Carolina	04	0.37	0.18	0.02
North Carolina	05	0.7	0.68	0.03
North Carolina	06	0.75	0.66	0.03
North Carolina	07	0.62	0.59	0.03
North Carolina	08	0.67	0.83	0.02
North Carolina	09	0.51	0.29	0.03
North Carolina	10	0.7	0.81	0.02

State	District	Hull Ratio	Path Probability	95% confidence error
North Carolina	11	0.82	0.68	0.03
North Carolina	12	0.25	0.14	0.02
North Carolina	13	0.54	0.29	0.03
North Dakota	01	0.99	1	0
Ohio	01	0.62	0.47	0.03
Ohio	02	0.81	0.92	0.02
Ohio	03	0.66	0.42	0.03
Ohio	04	0.54	0.41	0.03
Ohio	05	0.77	0.82	0.02
Ohio	06	0.6	0.56	0.03
Ohio	07	0.62	0.51	0.03
Ohio	08	0.62	0.64	0.03
Ohio	09	0.27	0.45	0.03
Ohio	10	0.86	0.95	0.01
Ohio	11	0.47	0.38	0.03
Ohio	12	0.62	0.58	0.03
Ohio	13	0.58	0.56	0.03
Ohio	14	0.82	0.85	0.02
Ohio	15	0.69	0.73	0.03
Ohio	16	0.58	0.51	0.03
Oklahoma	01	0.5	0.68	0.03
Oklahoma	02	0.81	0.76	0.03
Oklahoma	03	0.7	0.93	0.02
Oklahoma	04	0.76	0.92	0.02
Oklahoma	05	0.68	0.61	0.03
Oregon	01	0.7	0.72	0.03
Oregon	02	0.87	0.98	0.01
Oregon	03	0.71	0.83	0.02
Oregon	04	0.86	0.96	0.01
Oregon	05	0.6	0.54	0.03

State	District	Hull Ratio	Path Probability	95% confidence error
Pennsylvania	01	0.51	0.35	0.03
Pennsylvania	02	0.87	0.85	0.02
Pennsylvania	03	0.69	0.68	0.03
Pennsylvania	04	0.86	0.93	0.02
Pennsylvania	05	0.73	0.87	0.02
Pennsylvania	06	0.52	0.38	0.03
Pennsylvania	07	0.46	0.22	0.03
Pennsylvania	08	0.77	0.94	0.02
Pennsylvania	09	0.66	0.44	0.03
Pennsylvania	10	0.6	0.53	0.03
Pennsylvania	11	0.56	0.53	0.03
Pennsylvania	12	0.43	0.35	0.03
Pennsylvania	13	0.6	0.43	0.03
Pennsylvania	14	0.62	0.7	0.03
Pennsylvania	15	0.6	0.53	0.03
Pennsylvania	16	0.62	0.6	0.03
Pennsylvania	17	0.46	0.41	0.03
Pennsylvania	18	0.67	0.62	0.03
Rhode Island	01	0.45	0.74	0.03
Rhode Island	02	0.69	0.98	0.01
South Carolina	01	0.46	0.34	0.03
South Carolina	02	0.73	0.75	0.03
South Carolina	03	0.86	0.97	0.01
South Carolina	04	0.8	0.91	0.02
South Carolina	05	0.76	0.84	0.02
South Carolina	06	0.66	0.62	0.03
South Carolina	07	0.78	0.8	0.02
South Dakota	01	0.93	1	0

State	District	Hull Ratio	Path Probability	95% confidence error
Tennessee	01	0.77	0.85	0.02
Tennessee	02	0.55	0.53	0.03
Tennessee	03	0.62	0.36	0.03
Tennessee	04	0.69	0.51	0.03
Tennessee	05	0.79	0.86	0.02
Tennessee	06	0.71	0.84	0.02
Tennessee	07	0.72	0.74	0.03
Tennessee	08	0.8	0.93	0.02
Tennessee	09	0.69	0.75	0.03
Texas	01	0.77	0.85	0.02
Texas	02	0.42	0.31	0.03
Texas	03	0.87	0.96	0.01
Texas	04	0.78	0.89	0.02
Texas	05	0.68	0.78	0.03
Texas	06	0.76	0.85	0.02
Texas	07	0.55	0.5	0.03
Texas	08	0.84	0.91	0.02
Texas	09	0.61	0.56	0.03
Texas	10	0.73	0.87	0.02
Texas	11	0.62	0.62	0.03
Texas	12	0.77	0.88	0.02
Texas	13	0.67	0.93	0.02
Texas	14	0.5	0.54	0.03
Texas	15	0.52	0.42	0.03
Texas	16	0.94	1	0
Texas	17	0.67	0.87	0.02
Texas	18	0.59	0.34	0.03
Texas	19	0.71	0.68	0.03
Texas	20	0.7	0.64	0.03

State	District	Hull Ratio	Path Probability	95% confidence error
Texas	21	0.77	0.86	0.02
Texas	22	0.7	0.64	0.03
Texas	23	0.73	0.93	0.02
Texas	24	0.75	0.84	0.02
Texas	25	0.61	0.7	0.03
Texas	26	0.91	0.96	0.01
Texas	27	0.54	0.65	0.03
Texas	28	0.5	0.47	0.03
Texas	29	0.6	0.31	0.03
Texas	30	0.77	0.82	0.02
Texas	31	0.82	0.94	0.02
Texas	32	0.62	0.61	0.03
Texas	33	0.43	0.26	0.03
Texas	34	0.51	0.42	0.03
Texas	35	0.37	0.42	0.03
Texas	36	0.76	0.86	0.02
Utah	01	0.67	0.76	0.03
Utah	02	0.82	0.88	0.02
Utah	03	0.63	0.61	0.03
Utah	04	0.65	0.7	0.03
Vermont	01	0.83	1	0
Virginia	01	0.67	0.85	0.02
Virginia	02	0.25	0.86	0.02
Virginia	03	0.49	0.61	0.03
Virginia	04	0.75	0.72	0.03
Virginia	05	0.65	0.67	0.03
Virginia	06	0.69	0.78	0.03
Virginia	07	0.61	0.71	0.03
Virginia	08	0.78	0.79	0.03

State	District	Hull Ratio	Path Probability	95% confidence error
Virginia	09	0.67	0.97	0.01
Virginia	10	0.59	0.72	0.03
Virginia	11	0.52	0.37	0.03
Washington	01	0.73	0.89	0.02
Washington	02	0.46	0.8	0.02
Washington	03	0.7	0.86	0.02
Washington	04	0.66	0.65	0.03
Washington	05	0.85	0.81	0.02
Washington	06	0.78	0.97	0.01
Washington	07	0.54	0.83	0.02
Washington	08	0.66	0.7	0.03
Washington	09	0.65	0.75	0.03
Washington	10	0.69	0.84	0.02
West Virginia	01	0.46	0.84	0.02
West Virginia	02	0.51	0.36	0.03
West Virginia	03	0.67	0.71	0.03
Wisconsin	01	0.88	0.94	0.01
Wisconsin	02	0.88	0.91	0.02
Wisconsin	03	0.58	0.65	0.03
Wisconsin	04	0.71	0.79	0.03
Wisconsin	05	0.82	0.9	0.02
Wisconsin	06	0.69	0.71	0.03
Wisconsin	07	0.71	0.91	0.02
Wisconsin	08	0.68	0.92	0.02
Wyoming	01	1	1	0

BIBLIOGRAPHY

[1] JJ Allaire and Hadley Wickman. *RStudio*. RStudio, Boston, MA, 0.97.551 edition.

[2] Micah Altman. *Districting Principles and Democratic Representation*. PhD thesis, California Institute of Technology, http://resolver.caltech.edu/CaltechETD:etd-05192004-142452, 2004.

[3] Micah Altman and Michael McDonald. Public mapping project. http://www.publicmapping.org/what-is-redistricting.

[4] Micah Altman and Micheal P. McDonald. The limitations of quantitative methods for analyzing gerrymanders: Indicia, algorithms, statistics and revealed preference, July 2007.

[5] Jim Brunner. Activists propose 'majority minority' congressional district for washington. *The Seattle Times*, March 11 2011.

[6] Aaron Dilley Burden, Sam and Lukas Svec. Applying voronoi diagrams to the redistricting problem. *MCM Contest*, 2007.

[7] United States Census Bureau. State & county quickfacts: North carolina.

[8] Kristen D. Burnett. Congressional apportionment. United States Census Bureau, http://www.census.gov/prod/cen2010/briefs/c2010br-08.pdf, November 2011.

[9] Christopher Chambers and Alan D Miller. A measure of bizarreness. *Quarterly Journal of Political Science*, 5(1):27–44, 2012.

[10] Jowei Chen and Jonathan Rodden. Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quarterly Journal of Political Science*, 8:239–269, 2013.

[11] Redistricting Task Force for the National Conference of State Legislatures. North carolina redistricting cases: the 1990s.

[12] Andrew Gelman and Gary King. A unified method of evaluating electoral systems and redistricting plans. *American Journal of Political Science*, 38(2):514–54, May 1994.

[13] Emily Marshall Hodge, Jonathan K. and Geoff Patterson. Gerrymandering and convexity. *The College Mathematics Journal*, 41(4):312–324, September 2010.

[14] J. S. Kaastra, F. B. S. Paerels, F. Durret, S. Schindler, and P. Richter. Esri shapefile technical description. pages 155–190. 1998.

[15] Alan M Macearchren. Compactness of geographic shape: Comparison and evaluation of methods. *Geografiska Annaler. Series B, Human Geography*, 67(1):53–67, 1985.

[16] James Madison. The federalist 10: The utility of the union as a safeguard against domestic faction and insurrection, November 1787.

[17] Keith T. Poole McCarty, Nolan and Howard Rosenthal. Does gerrymandering cause polarization? *Social Science Research Network*, 2008.

[18] Stacy Miller. *The Problem of Redistricting: the Use of Centroidal Voronoi Diagrams to Build Unbiased Congressional Districts.* Senior project, Whitman College, May 2007.

[19] National Atlas of the United States. 1:1,000,000-scale congressional districts of the united states 113th congress. http://nationalatlas.gov/atlasftp-1m.html.

[20] National Atlas of the United States. Congressional districts of the united states - 112th congress.

[21] Everett Rummage. Minority-majority redistricting: Empowerment for whom? *Seattlest.com*, September 14 2011.

[22] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

[23] Edward R. Tufte. The relationship between seats and votes in two-party systems. *American Political Science Review*, LXVII(2):540–554, June 1973.

[24] David Wasserman. 2012 national house popular vote tracker. Google Doc.