Feature Engineering District Maps for Fairness Analysis

Betsy McCall

GEO 590 – UB – Spring 2021

Final Submission

Abstract:

This paper will examine the relationship between district maps and electoral fairness from geometric and population perspectives. This paper will serve as "proof of concept" for a larger analysis of district data nationwide. Here, we will examine the district maps of three states, Ohio, New York and Maryland, chosen for the differences in party control of the government at the time of redistricting after the 2010 census. The analysis will state definitions of fairness, and describe feature engineering for assessing how the shape and properties of districts may impact fairness, and whether one-party control of the state impacts the likelihood of fair outcomes.

I. Introduction

A. Problem Statement

Representative democracy created a number of new mathematical problems related to voting, weighted voting, apportionment and drawing district maps. Three of these problems (voting methods, weighted voting and power, and apportionment) and the questions of fairness that arise from them have been solved. In the case of voting and apportionment, it has been shown that no voting or apportionment method can be perfectly fair. Every voting method will sometimes violate a fairness criterion, and every apportionment method will sometimes create a paradox or violate the simplest fairness rule. Despite this, the problems are far better understood than they were when the United States was founded, and strategies can be developed to minimize the violations of fairness principles that matter most.[19] The one that remains largely unsolved is the question of drawing fair district maps. This is no doubt in part because the geometry and analysis needed to assess both the problems and solutions are orders of magnitude more difficult. To date, the only well-defined fairness criteria acknowledged by law is the equal population provision of Congressional districts.[16]

In North America, the drawing of district maps predates the Republic. The government of England largely allowed their colonies to govern themselves until shortly before the Revolution with minimal interference. After the adoption of the Constitution, district maps had national and not just local impact. Drawing of district maps to support incumbency and preserve the power of those already in office preceded the invention of the term "gerrymander". This means that district map makers have had more than two-and-a-half centuries to perfect their methods to create unfair district maps.[16] Creating fair district maps has been understudied. While various methods have been proposed in reaction to gerrymandering techniques, the research to support these methods, research to show that they do indeed produce fair maps is largely lacking.[7]

The open question then remains: What properties of district maps make such maps fair? And more fundamentally, what do we mean by fair?

B. Objectives to Achieve

Since the United States, from which the data for the analysis will be drawn, has operated under a primarily two-party system since the election of 1800, the following definition of fairness will be adopted, drawing on the definition of the Quota Rule from apportionment:

A district map will be assessed as fair under the following conditions:

i. *All the votes for a given party will be summed across all districts and a statewide percentage for that party in the legislative chamber or Congressional delegation will be found.*

ii. *That percentage will be multiplied by the number of seats in the chamber or Congressional delegation.*

iii. *The number of seats awarded should be equal to either the floor of the number found, or the ceiling to be considered fair. If more than the ceiling, or less than the floor, is awarded in the election, then the district map will be considered unfair.*

This definition essentially aligns with proportional representation, which is commonly practiced in parliamentary systems. It is not entirely clear whether the system in the United States is compatible with achieving proportional representation routinely, but this is nonetheless a good place to begin.

Alternative definitions of fairness are possible. At a minimum, the party that achieved the most votes statewide should have the majority of seats in the Congressional delegation, or the state legislature. In addition, if a greater percentage of people voted for a given party from one election to another, one would want that movement to be reflected in the new awarding of seats: the proportion of seats held by the party with more votes than the last cycle should increase. One might consider this to be a measure of sensitivity to the will of the public. The latter definition would require a comparison of election results in successive cycles.

Demographic features might also be of interest to fairness considerations, given the political history of race in the United States. If one thinks of a district as a community, one might also wish for the parts of the district to be relatively similar, such as on race or population density or other factors, to make the interests of the district easier to represent. However, some of these factors might also be considered features of the districts themselves rather than a fairness criteria per se.

Using the above definition of fairness, the goal of the analysis will be to examine the shapes of districts at the Congressional level (and eventually at the state legislative level). States that have only one Congressional district will be excluded at the Congressional level. Using district shape files, and overlay with Census data, and county and municipal boundary data, various properties of the district can be calculated. The district data will be joined with voting data, and then statewide properties will be obtained. These properties can then be used to build a machine learning model to determine which properties or combination of properties have the best correlation with the proposed definition of fairness.

Ideally, the model developed will be interpretable, in order to aid in the drawing of future district maps. However, models that are harder to interpret but which have high predictive value may still have value since future maps can be run through the model to determine their likely effectiveness, and the model updated with new data for future map-drawing cycles.

The present project will endeavor, at the Congressional district level, using the state maps of Ohio, New York and Maryland, to engineer features that can be used in the machine learning model as a preliminary step to this later analysis. Maps will be used to illustrate the findings. At present, we have an intuitive sense of which districts “look” like they are gerrymandered, and which are not. The project here will be to begin to quantify these intuitions so that they can be tested and modeled.

C. Significance

Much of the previous work in the literature on district maps have centered around technical solutions to produce them using a variety of methods that include genetic algorithms,[6] crowd sourcing,[2] and other optimization methods.[5] But, since at present there are no clear properties to shoot for by which to produce reliably fair maps, these strategies employ methods that don’t have evidentiary support. States have recently resorted to independent commissions to prevent ruling parties from drawing district maps to their liking. Courts have generally been reluctant to step in without clearer standards that can make their rulings enforceable. A method of assessing the fairness of such maps with high predictive power would allow both the public and the courts a basis upon which to make judgments. A model that produces clear properties can likewise provide a basis for algorithms capable of drawing such maps. Such a model could also provide a means of ranking which properties are the most important, and which properties may be traded off.

Fair district maps can ensure that democratic votes can influence the make-up and policy positions of their leaders in response to their vote. Gerrymandering subverts the will of the voters and therefore subverts democracy itself.

II. Background

The literature on redistricting falls into a couple of broad categories, along with a few works that are somewhat out there on their own. Broadly speaking, there are a number of works on the political (science) and legal aspects of redistricting. This is mostly not the focus of the present study, so we will set these aside. A second large group of work focuses on the technical aspects of building district maps, from the Congressional and legislative level, all the way down to the local town council and school board level, in GIS software. We will discuss this part of the literature more below. There are also some isolated works on the history of gerrymandering, and the mathematics of redistricting. These works will play some role in the discussion that follows.

Concerns about the process of redistricting are not new. The term gerrymander itself is a reaction to corrupt redistricting processes early in the nineteenth century. Griffith discusses the history of gerrymandering in the United States (and American colonies) early in the twentieth century.[16] The impact of legal rules on the considerations for redistricting, as discussed by Cannon [25], Buchman[23] and Rush[31], are necessary because they will affect the way that any districts must be drawn and add constraints to any model solution. One important legal consideration is the equal population constraint, which is relatively easy to implement. There are more complex, and even conflicting, legal considerations around race that will ultimately have to be included in any workable model, but which may be worth setting aside in an initial model: not because race is not important, and not because race is not routinely used as a basis of discrimination in redistricting, but because given the conflicting legal landscape, it is worthy of more specialized study.

The literature on the implementation of GIS methods for redistricting tend to focus on the process of drawing new district maps themselves: either as a technical method for constructing the maps, or as part of an optimization system for analyzing constraints in a model. Salling discusses how Ohio, before the 2010 Census, set up a system to allow the public to contribute models to the redistricting process. Weights were assigned arbitrarily in order to assess the maps, and they identified problems with geometric compactness where boundaries followed natural rivers or other irregular features.[2] Joshi and Soh developed a method for designing and optimizing districts boundaries subject to constraints using polygonal clustering.[5] And Xiao likewise used geographical optimization methods using evolutionary algorithms to draw district lines.[6] Bong, Chai and Wong used a multi-objective metaheuristic for the same purposed.[10] Each of these models is forced to contend with many of the same issues that Salling discussed in the Ohio public modeling process. Assessment of the constraints is assigned somewhat arbitrarily, and often not all the constraints can be handled at once. The model to be optimized is based on conjecture rather than historical evidence of effectiveness. GIS software is not designed to handle decision-making processes of the sort required by this type of optimization.[10] As Bowen acknowledges, relatively little research has been to determine which redistricting principles are actually effective in achieving the aims of the model.[4]

The literature does propose various properties that might be effective in combatting gerrymandering techniques. Bowen discusses several, including the value of aligning district boundaries with county and municipal lines for better community cohesion.[4] Suzuki's book on the mathematics of redistricting also includes a discussion of geometric compactness, one of the properties proposed to constrain gerrymandering methods.[7] The specific methods selected for this research will be outlined in the Methods section below. The proposed data mining methodology appears to be absent from the literature.

The purpose of the present research will be to fill in some of the noted gaps above, and to provide support (or lack of it) for the constraints and weights for other optimization or district assessment approaches.

The data to be used for the present study comes from three different sources. The shapefile data for the 2012 election of the 113th Congress was originally obtained from the Census TIGER shapefile, but here was directly sourced from Lewis, et al. via UCLA, who have collected not just the recent shapefiles available from the Census, but also shapefiles extending back to the first Congressional district maps.[39] While these earlier maps will not be used in the present analysis, they may eventually be employed in the larger study.

A second set of data for the results of the 2012 elections for the 113th Congress was obtained from the MIT Election Lab site.[41] This data is aggregated by congressional district. The data is available for several recent decades.

The third set of data to be employed here is Decennial Census data from 2010 on population and race obtained from the Census API.[42] Since the districts maps in 2012 are based on the data from the 2010 Census, the match between the maps and the population should be the closest match to assess the intentions of the mappers. The race data will be used to assess discriminatory effect and community cohesion. One goal of this work will be to produce visualizations, in addition to the calculations, to examine these features. At the state level, it is necessary to use census tracts, rather than block-level data, since blocks are not available at that level of geometry. For the larger project, if block-level data is required, it will have to be acquired county-by-county for each state.

Some features will not be considered here. USGS data will be required to provide information on state, county and municipal boundaries.[43] This, too, can lead to an analysis of some definitions of community cohesion, and some geometric assessments of gerrymandering: whether the boundaries align with political boundaries or not. District boundaries that cut through communities may be a symptom of cracking. Decennial census data on overall population will not be employed here, but will be required for the more complete analysis.

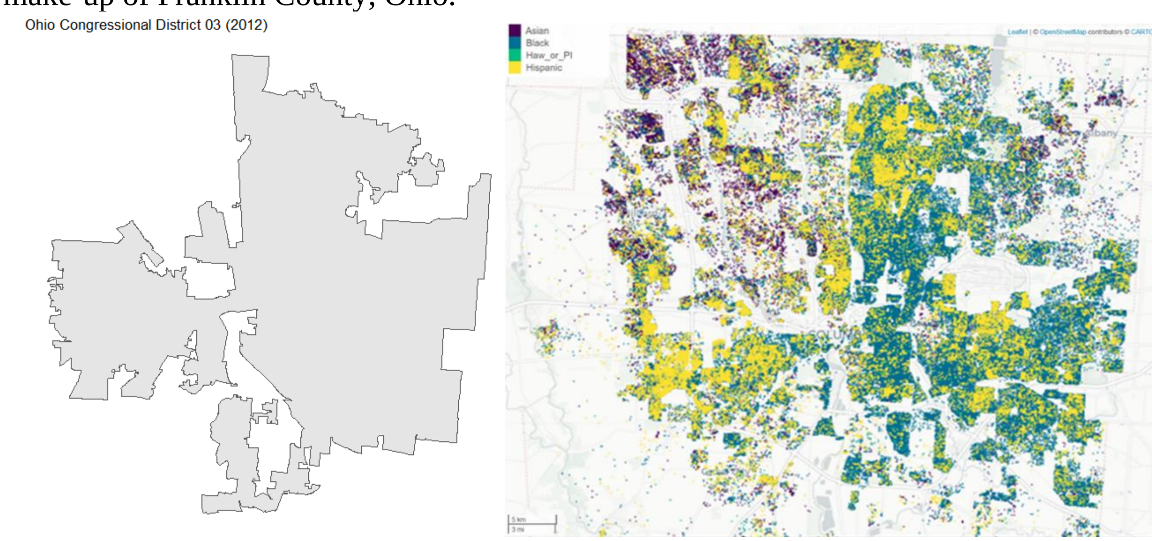
III. Methods

A great deal of written analysis has been put toward describing the methods used to create gerrymandered district maps. The primary methods among these are called cracking and packing, and they are usually used in concert with each other. The method of packing is a technique where the minority party voters are packed into a small number of districts in which they are able to run up the score for their party candidate. The majority party essentially concedes these districts to the minority. Cracking is a method of dividing up voters for the minority party into minority groups within districts dominated by the majority party, thus minimizing their ability to win the district outright without a wide swing in voter preferences. These districts may be won by the majority more narrowly than in the packed districts, but the margin is meant to be large enough to make these districts secure votes for the majority.[7] By these methods, the majority party at the time the district maps are made are thus able to secure their power for the next decade. Moreover, this sort of rigging of the district maps contributes to the rise of extremism and polarization because elected officials become more concerned about being primaried, than they do about the voters in the general election.[30]

The intersection with race here lies in two factors: a historical effort to pack racial minorities into a single district ensures two things: the minimization of the impact of those racial minorities on state and national politics, and the increased likelihood that a member of the racial minority may be elected to office.[11] Because explicitly gerrymandering on race has attracted legal criticism, more recent efforts have focused on cracking and packing by party, with the clear understanding that since the Civil Rights era, African Americans in particular, vote overwhelmingly for one party, making it an easy proxy for racist gerrymandering.[13][14] It is for this reason that the present study will focus on the distribution of political power by party. While there is some research suggesting that political party affiliation may not be as robust as some other factors in the past,[8] in the present era of political polarization, it seems to be a reasonable place to begin.

Figure 1 shows the comparison of the 3rd Congressional District Map of Ohio with the racial make-up of Franklin County. As is clear from the map, black and Latino residents are grouped into a single district (packing), while the Asian population around Ohio State University (top left) is extracted into a neighboring, mostly rural district (cracking).

Figure 1. A comparison of Ohio District 3 (centered on Columbus, Ohio) and the racial make-up of Franklin County, Ohio.



Racial data processed in R, obtained from the 2010 Decennial Census.

For the purposes of the present study, I have defined in the Objectives section a working definition of fairness to be deployed for this analysis. Certainly, other definitions could be proposed—and should be proposed—in order to analyze the same features for their relationship to better district maps. Certainly, other kinds of demographic factors could be taken into consideration (such as race, urban/rural, wealth, etc.) could be the basis of representative considerations. Other types of Census data may need to be employed for such analysis. It may be that different properties of districts lead to better outcomes in one area than they do under another model of fairness, as we see in voting methods. It is certainly possible that the features that are beneficial to fairness of any one kind may change over time as political dynamics change. The proposed study would not attempt to account for every possible factor, but the hope is to develop a framework for further exploration.

A. Summary

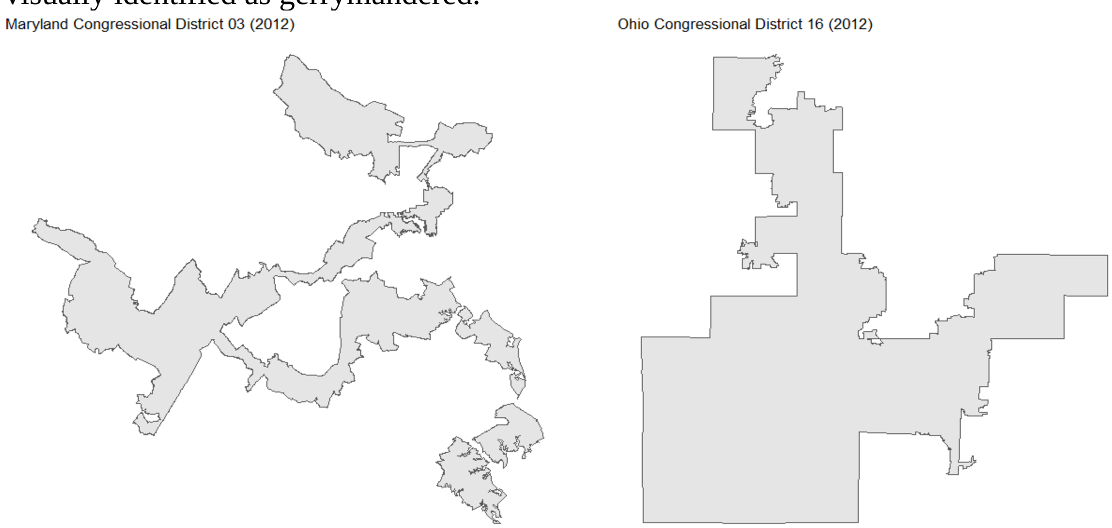
While a full analysis of district maps at all levels of government should be undertaken, the present study intends to focus on the analysis of Congressional districts. The scale of the districts and their purpose may impact the results, so the scope of analysis will be limited initially. Features will be calculated using R, and then combined with Census data and voting results. The resulting dataset will then be processed by machine learning in the same software.

B. Processes and Features

Each feature that is extracted from the data sets above will require different processes or combination of processes. Therefore, listing each of them will be a challenge. The examples below may not be a complete set of all possible reasonable features, will serve as a basis of the present analysis. Features will initially be applied to individual districts. To map to the fairness criterion, which is a state-level analysis, the features will then be summarized to obtain a statewide figure.

A mathematical definition of compactness has been proposed as one possible way of thwarting cracking and packing, since it seems to restrain the most egregious snake-like districts. Figure 2 shows some examples of “snake-like” districts from Maryland and Ohio. When people think of gerrymandering, this is the kind of visual that comes to mind.

Figure 2. Examples of Congressional districts from Ohio and Maryland that could be visually identified as gerrymandered.



Maps created from UCLA district shape files.

Geometric compactness can be defined by the ratio of the area of a shape to its perimeter. The most compact shape is a circle.[7] However, there are several problems with traditional geometric compactness that have been noted in the literature. The value of the ratio depends on the size of the object. A circle of radius 1, has an area of π, and a perimeter of 2π, making the ratio of area to perimeter equal to ½ . But, if the circle has a radius of 2, the area is 4π, and so is the perimeter, making the ratio equal to 1. As the radius increases, so does the value of the ratio. A modified definition of compactness, perhaps using the square root of the area, may be a better measure for shapes of different sizes. From the shape files, we would want to calculate both the area and the perimeter of each district. These figures will be combined to obtain the mathematical compactness value and the modified compactness value.

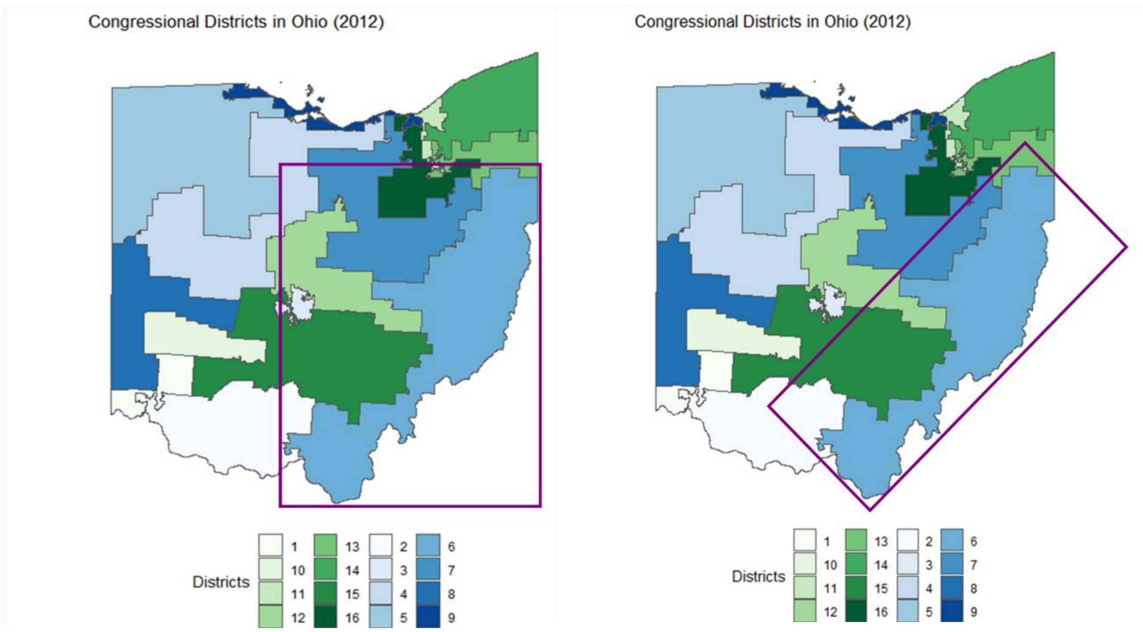
Another known potential problem with the perimeter component of compactness is that some district boundaries are irregular not because of gerrymandering, but because they follow natural boundaries, such as rivers, lakes, or oceans. Salling noted that some boundaries had to be "smoothed" to better compare them with the straight-edged boundaries in the interior of the state.[2] His article did not specify the method of smoothing, but there are a couple of possibilities available using the data proposed. Buffering could be used on the border one kilometer wide, or one meter wide depending on the desired level of smoothing. Then a calculation of area included in the buffer would be proportional to a smoothed perimeter boundary. Another alternative would be to measure the boundary by counting Census blocks that intersect the border and using the total number of census blocks in the district as a measure of the area. Census blocks would absorb any irregularity along the border by adjusting the size or shape of the block. An even more simplified measure could be block groups or census tracts. They could be converted to distances by measuring their widths. There are also simplification methods for polygons in R, however, these methods would have to be used with care since we want to retain the complexity of any interior boundaries and only smooth boundaries that follow state lines or possibly other natural features. These features will require additional shape files for states and counties and so will be left for a future analysis.

The literature also discusses the value of community cohesion in drawing district lines.[4] A measurement of the boundary that would consider this factor would weight boundaries by the extent to which they followed existing state, county, and municipal boundaries. In particular, boundaries that follow state lines might be given little or no weight (penalty), boundaries that followed county lines given a low weight, boundaries that follow town boundaries given a medium weight, and then boundaries that cut through towns given a higher weight. Thus, there would be benefit to aligning district boundaries with other geopolitical boundaries. Census and USGS data overlaid with the Congressional districts could calculate which part of the district borders aligned and which did not. These calculations would then be used in compactness calculations.

Another possible feature to consider might be a bounding box. We could use a feature that was the area of the district in a ratio with the area of the bounding box. This might be a way to control for the size of the district more easily than geometric compactness. A ratio of the district area to the bounding box area may be another way to reveal "snake-iness" of a district since if there is a lot of space in the bounding box not used up by the district, this could be considered non-compact, and therefore, gerrymandered. Ideally, we would want to account for state borders as well, but since we are dividing area by area, the resulting ratio would be unitless, and so should be effective for both large and small districts.

One complication of this method is illustrated in Figure 3. Bounding box dimensions typically follow latitude and longitude lines and so are orientated in an upright orientation. However, districts that follow state boundaries may appear to be oriented diagonally, and so rotated bounding boxes may be more desirable, and more accurate for determining this type of compactness. It is not yet clear to me how to easily obtain theses rotated orientations, so they will not be considered in the present analysis. Bounding boxes in R are reported as locations of the four corners: xmin, xmax, ymin and ymax. These coordinates will have to be converted to polygons (like the districts themselves), or the distances across the rectangles will have to be calculated from their degree measures and converted to meters.

Figure 3. Bounding Boxes are typically calculated in the upright orientation, as shown on the left, but it may be fairer and more accurate to allow for rotated bounding boxes, shown on the right.



Congressional district maps of Ohio taken from UCLA shape files.

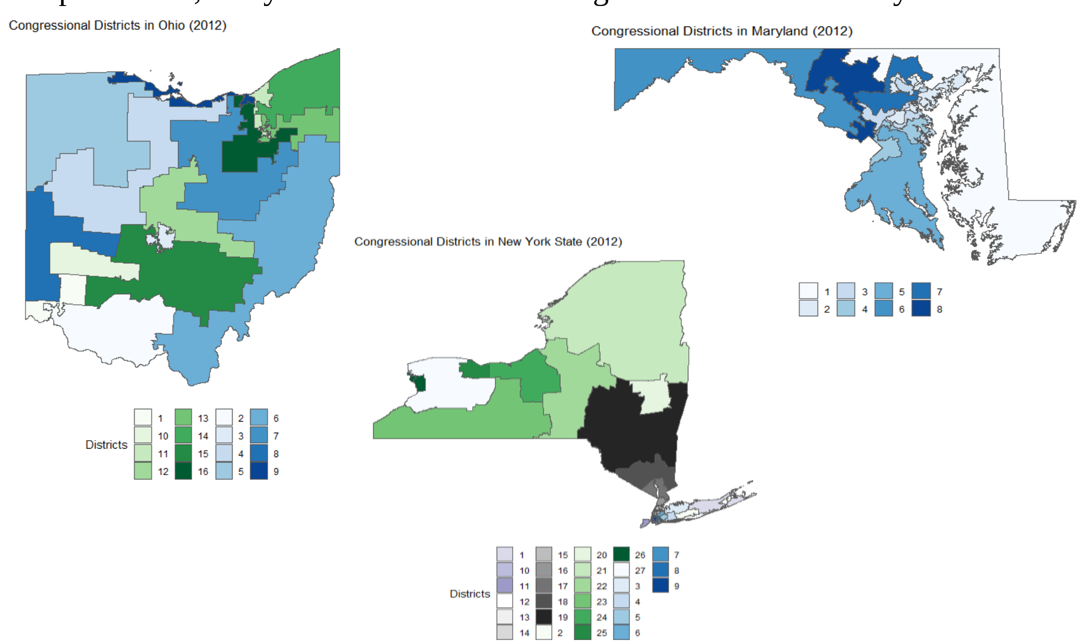
All districts nationally should have approximately the same size, and even more so at a state level, of about 750,000 people. Population density could be a consideration. If a district is be cohesive in terms of urban/rural or other cultural factors, one consideration may be average population density in the district, or using Census blocks/groups/tracts to find the range of density in the district: smaller ranges of density by block would mean the district was more cohesive and similar, while a larger range in density would signal a district with highly dense urban areas together with less dense rural areas in the same district (which in turn could be a sing of cracking). Indeed, the Hoover index could be a useful measure of this. To find the densities, we would need to intersect the Census geometry with the district shapes and then find the maximum and minimum densities for the second measure. The average density could be obtained by the population divided by the area. The present computational problem appears to be that it may need to be done district-by-district in order to associate appropriate population data with districts. More than mean population density is beyond the calculation power of my home computer.

Census data can provide information on racial makeup of a district (as seen in Figure 1 for Franklin County), or other demographic factors such as median housing price. These could be compared across the district to find the range of values by block across the district. I was able to obtain data on racial makeup for six different racial groups across the three states in this analysis. Similar to population density, the similarity of racial makeup across the district across census geometries could be measured. Cracking might be discerned by extreme differences of racial makeup from one end of the district to the other. Packing might be detected by the noting high density of racial minorities in fewer districts. The outcome of this analysis might be especially interesting, since this is where court rulings are currently in conflict. However, processing this data for higher density counties is at the limits of my current computational power.

Because this type of analysis has not been done before, a wide variety of features, including variations on features, should be considered for inclusion, including redundant alternative definitions of the same feature, like compactness. Statistical analyses can be done during the data mining process to determine which are most effective, which are redundant, and can be excluded from the final analysis. Spatial statistics might also suggest additional features to consider.

IV. Results

Figure 4. Maps of Ohio, Maryland and New York Congressional districts analyzed.



Data from UCLA Congressional shape files.

To begin, let’s examine the maps of the three states color coded by congressional district, shown in Figure 4. These states were selected because at the time the districts were drawn, they had three different types of party control. Ohio was controlled by Republicans at both the state legislative level and had a Republican governor. Maryland had the same situation but Democratic control. New York had divided control. Ohio had 16 Congressional districts, New York 27, and Maryland 8. Maryland’s relatively smaller number of districts will make this initial examination more difficult, but when taken together with full national data and multiple elections, should not be as difficult as states with even fewer districts. Examining the maps, some of the gerrymandered districts are readily noticeable in Ohio. Some can also be seen in Maryland. New York initially looks relatively un-gerrymandered, however, because New York City is so population dense, many of the districts can’t be seen on this map scale. In the Appendix, I have included maps of individual districts that I selected that appears to be the most gerrymandered in each state. I selected the fewest from New York, however, from Ohio, I ended up leaving out several very gerrymandered districts because the list was getting too long.

Before examining the features that were calculated, let’s examine the overall fairness considerations of each of the states. The data is summarized in Table. 1

Table 1. Comparison of voting percentages to Congressional district allocation.

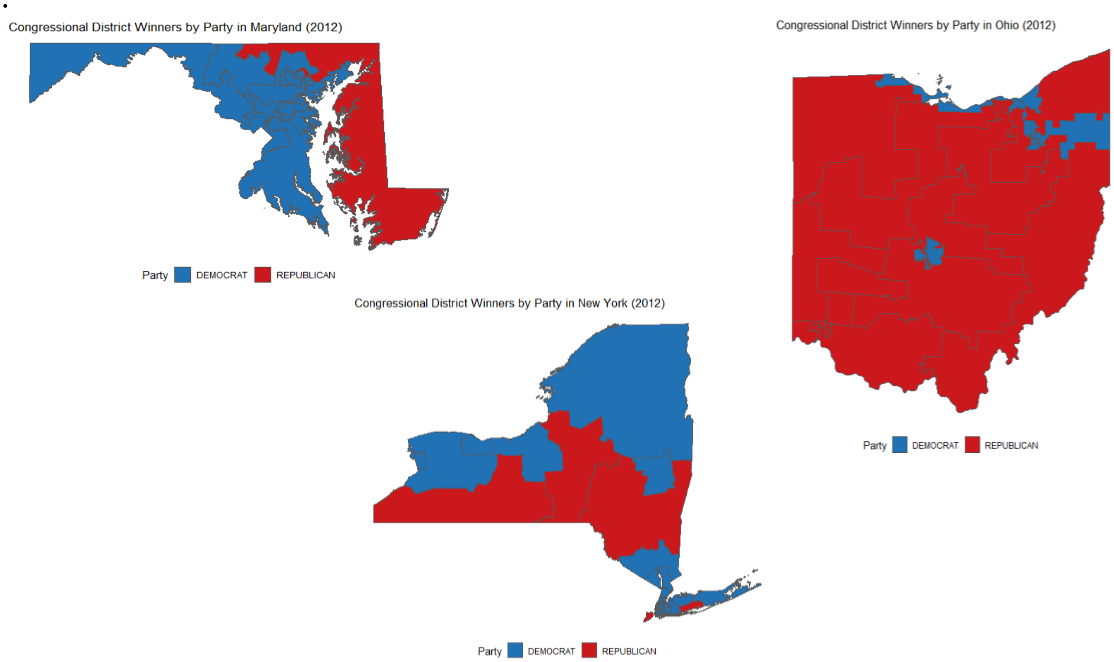
	Voting Percentage of Total Votes Cast Statewide for Congressional Candidates in 2012 (113th Congress)		
	Maryland	New York	Ohio
Democratic Party	62.92%	58.00%	46.91%
Republican Party	33.20%	31.65%	50.96%
Gap	29.72%	26.35%	4.05%
	Expected Number of Seats based on Statewide Results		
	Maryland	New York	Ohio
Democratic Party			
Mean	5.03	15.66	7.51
Maximum	6	16	8
Minimum	5	15	7
Actual	7	22	4
Amount Outside Range	+1	+6	-3
Republican Party Mean	2.66	8.55	8.15
Maximum	3	9	9
Minimum	2	8	8
Actual	1	5	12
Amount Outside Expected Range	-1	-3	+4

	Percent of Seats Won		
	Maryland	New York	Ohio
Democratic Party	87.5%	81.48%	25%
Republican Party	12.5%	18.52%	75%
Gap	75%	62.96%	50%
Ratio of Actual Gap to Voting Gap	2.52	2.39	12.35

Based on these results, both Maryland and New York and strongly Democratic states. The gap between the percent of voters for Democratic candidates and the percent of voters for Republican candidates is over 25 percentage points in both cases, and in Maryland, nearly 30 points. So, while the percentage of Congressional seats awarded to Democrats is greater than that, it is qualitatively different than what is going on in Ohio. The margin of victory was extremely narrow in Ohio, in favor of Republicans, but the number of seats awarded to Republicans was much larger than expectations. These results suggest that Maryland may be the closest to proportional representation (they differ by only one seat from the expected number). New York is more complicated by having several third-party candidates perform better than in either Maryland or Ohio. All three states in this analysis do, at least, match party control to the votes cast statewide in this election: the party that had the most votes statewide, did get awarded the most seats.

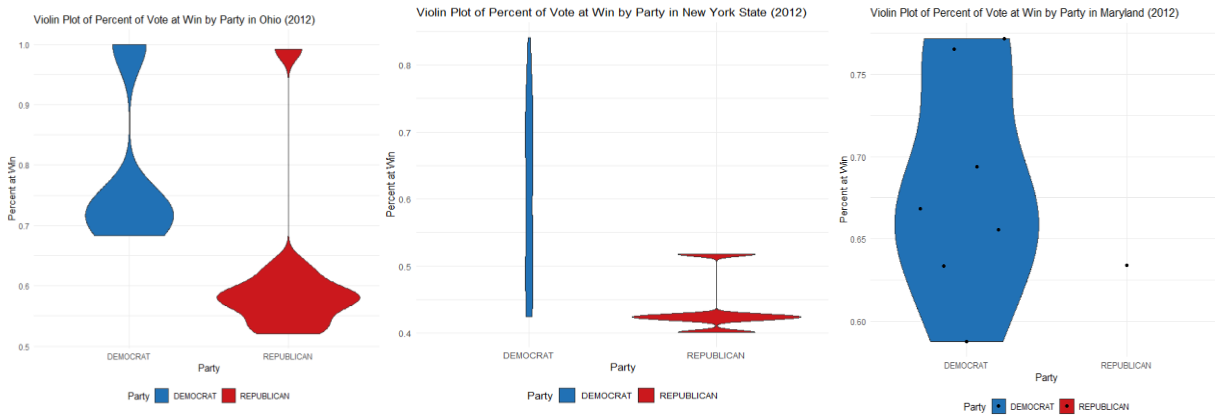
Let's look at these results on a map, and how they differ by party. In Figure 5, the results on the Congressional election, colored by party is shown. In Figure 6, the party winners by percent of vote received in each district is shown in violin plots.

Figure 5. Congressional district maps for Ohio, Maryland, and New York colored by party winner.



District maps from UCLA Congressional shape files. Election data obtained from MIT Election Lab.

Figure 6. Violin plots of election winner voting percentage by party for Ohio, Maryland and New York.



Data calculated from MIT Election Lab data. The Maryland plot is overlayed with a swarm plot since one can't calculate a violin plot from a single value.

What is noticeable from the Congressional district maps in Figure 5 is that the most gerrymandered districts are all Democratic districts (in the minority), while this does not appear to the be the case in both Maryland and New York. In Figure 6, the range of winning percentages differs greatly by party in all three states. While it may be necessary to neglect Maryland since Republicans won only one seat, the range of percentage wins by party differs. The Democratic districts in Ohio are all very lopsided, while the Republican victories are much closer to 50%. In New York, the Republican districts are likewise more competitive, but Democratic districts span the range from competitive to lopsided. One possible interpretation of these graphs is to think that perhaps the difference in ranges by party may be a feature. Another possibility may be the difference in the minimum winning percentages. Since the fairness criteria are all statewide criteria, but the features plotted here are district-level criteria, some sort of summary score will be necessary to map this data onto a statewide feature.

Let us next consider the idea of compactness. First, we can compare two measures of compactness: the traditional one of the ratio of area to the perimeter, and the modified version of this proposed earlier using the square root of the area to make it more size invariant. We will consider Ohio specifically in Figure 7 as a point of comparison as map. Figure 8 will compare the violin plots of both features, by party in each state. Table 2 summaries the compactness scores for both methods.

In both measures of compactness, the smaller the value, the worse the gerrymander. Smaller values indicate that the perimeter is longer than it should be relative to the area of the district. Both maps shown below suggest some of the same (dark blue) districts have problematic compactness values. Districts with less gerrymander (light blue) differ between the two maps according to size. The modified compactness score is not as large due to the geometric size of the district, while traditional mathematical compactness maintains a relationship with a district's absolute size.

Figure 7. Two measures of geometric compactness in Ohio Congressional districts.

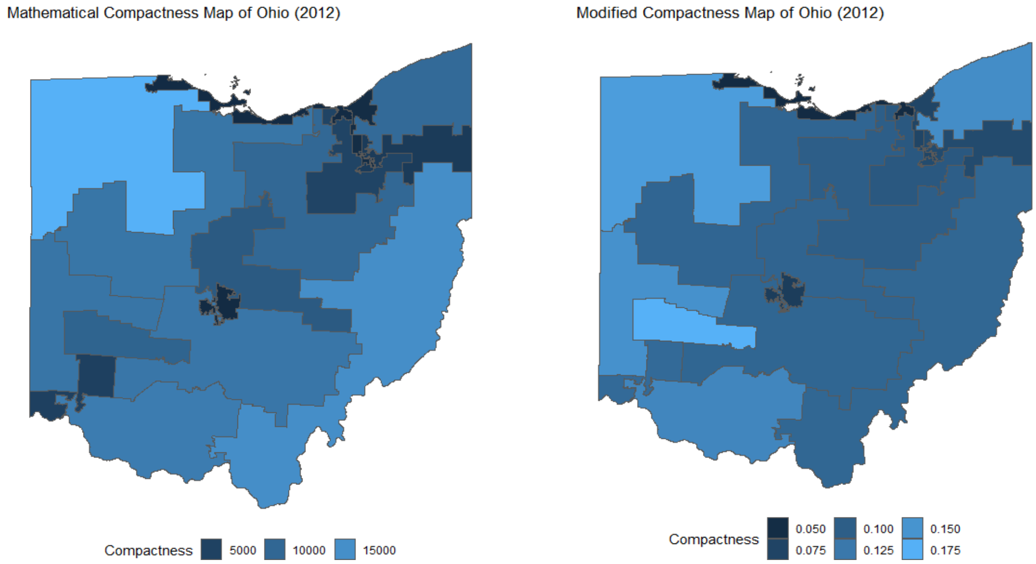


Figure 8. Violin Plots of compactness measures by district, plotted by party.

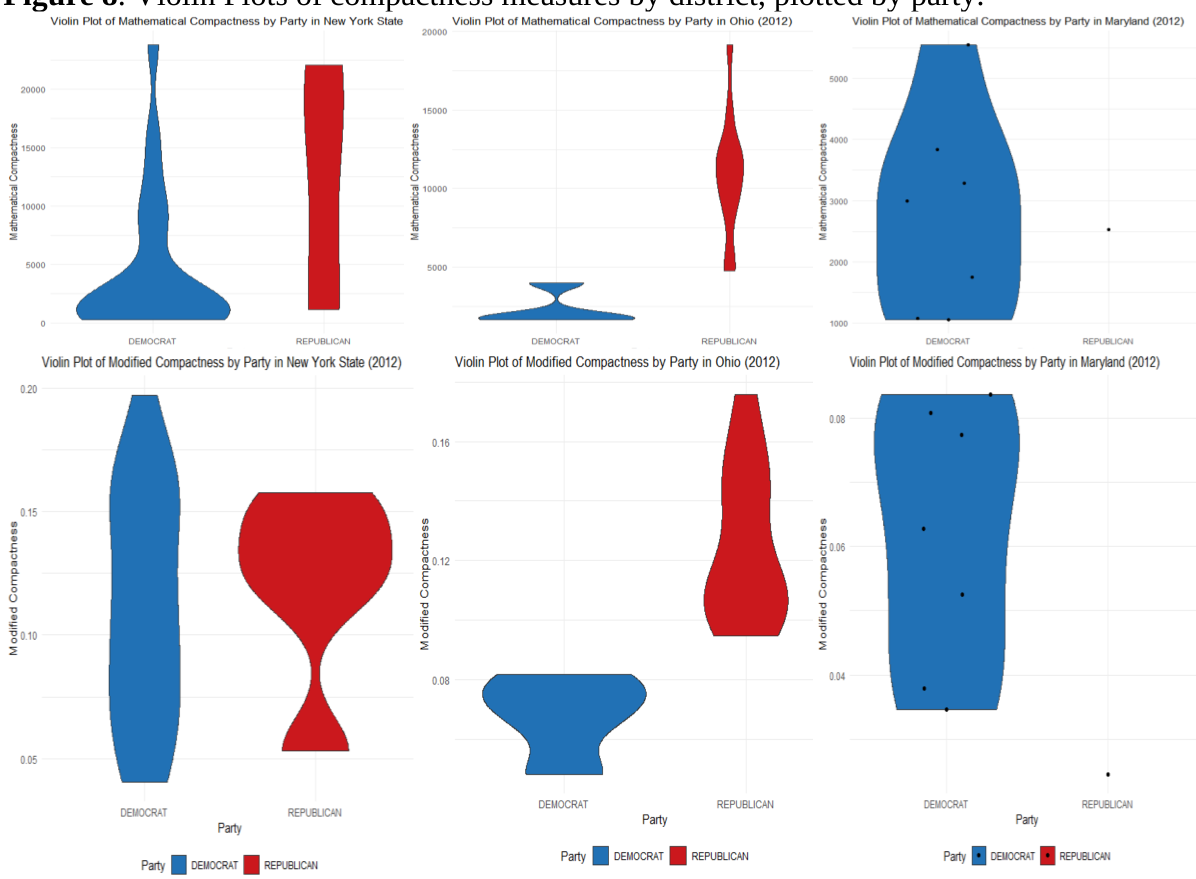


Table 2. Summary table of compactness measures by state.

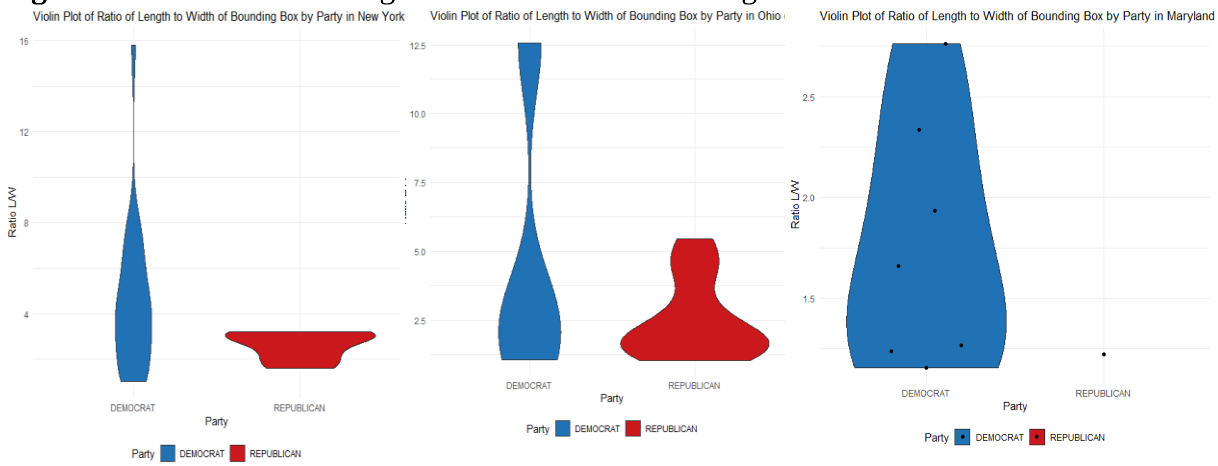
	Compactness Measures by State		
	Maryland	New York	Ohio
Mathematical Compactness A/P			
Minimum	1052	293	1644
First Quartile	1582	785	4571
Median	2759	1636	9816
Mean	2758	5906	8477
Third Quartile	3421	9494	11894
Maximum	5553	23803	19159
Modified Compactness $\sqrt{A}/P$			
Minimum	0.025	0.040	0.048
First Quartile	0.037	0.073	0.091
Median	0.058	0.117	0.108
Mean	0.057	0.113	0.111
Third Quartile	0.078	0.157	0.140
Maximum	0.084	0.197	0.176

Geometries based on NAD83. Smaller ratios imply that the perimeter is larger (possibly more gerrymandered) relative to its area.

Looking at the violin plots for the three states, we see somewhat similar patterns to those we saw in percentage wins. The ranges of compactness scores differ by party. Not by much in New York, which is thought to have less gerrymandering because of the mixed-party control at the time the map was drawn, however, the differences in Maryland and Ohio are more stark. The violin plots in Ohio do not even overlap. Thus, we can expect that differences in ranges, or minimum or maximum values could serve as summary features for the states. Based on these plots, though, it is not clear which method of measuring compactness is to be preferred, since the plots of both statistics appear to be quite similar. The extremely small value for the Republican district in Maryland results from its boundaries being mostly shoreline. Other methods of measuring the perimeter will likely help to bring the value here back into line with the measurements from other districts in the state.

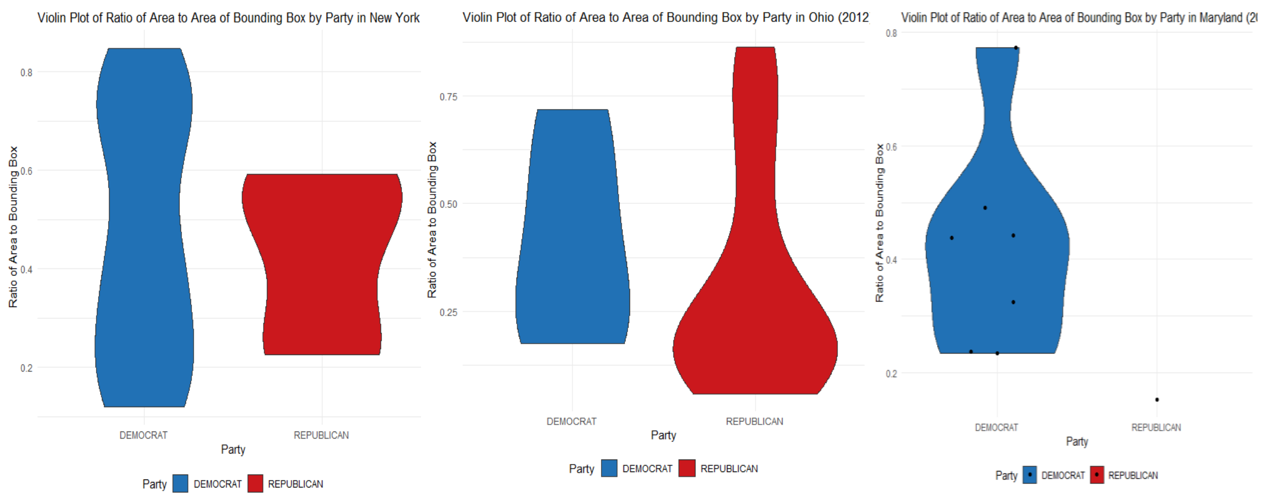
The last set of features I calculated for this analysis were related to bounding boxes. The length and width of the bounding boxes were calculated in meters, and the ratio of the length and width obtained on the idea that long and skinny districts might be more heavily gerrymandered than squarer districts. In addition, from the length and width, the area of the bounding box was estimated and the ration of the district area to the area of the bounding box was obtained on the idea that heavily gerrymandered districts will take up less of the bounding box. One could think of both measures as a kind of compactness measure. Figure 9 shows the comparison by party in the three states of the ratio of length to width, where length is always the longer side. Figure 10 shows the ratio of the area of the district to the area of the bounding box. Here, larger is better.

Figure 9. Violin Plots of Length to Width of the Bounding Box for each district.



Calculated from UCLA Congressional shape files.

Figure 10. Violin plots of the ratio of area of the district to the area of the bounding box by party in each state.



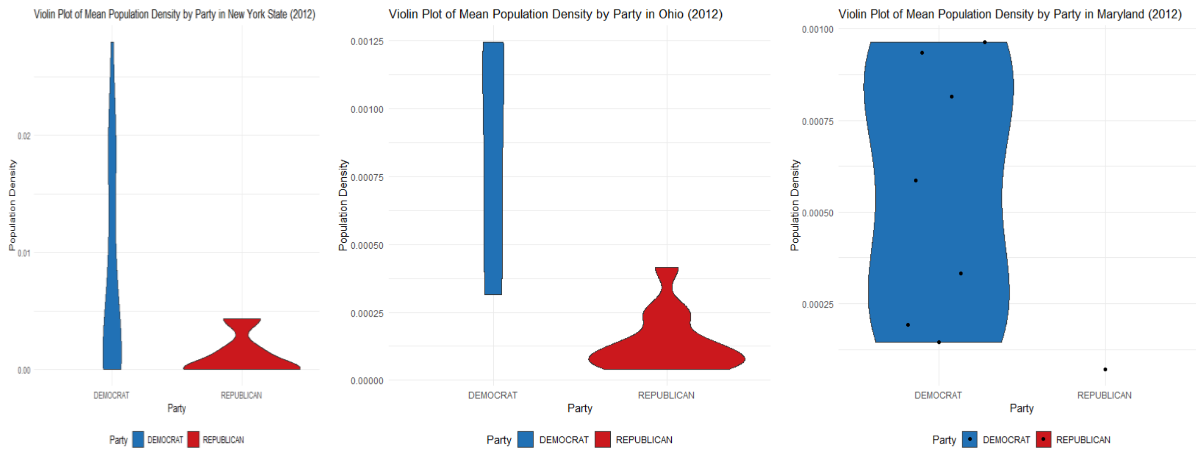
Calculated from UCLA Congressional shape files.

In Figure 9, in all three states, the ratio of length to width is larger in Democratic districts than in Republican ones, which is itself intriguing. According to Figure 10, in New York and Maryland, the ratio of area of the district to the area of the bounding box spans a larger set of values for Democratic districts than for Republican districts. The reverse is true in Ohio. The difference in the ranges by party might also be a useful statewide feature here.

One potential difficulty with the bounding box calculations is the accuracy of the conversion formulas for the coordinates to meters to the area comparison. Alternatively, it may be easier in a future analysis to convert the bounding boxes polygons similar to the districts so that the same area calculations can be performed on both objects with the same formula. Some initial estimates of the area produced areas of the bounding boxes that were smaller than the district sizes, which is impossible.

The last possible feature I was able to calculate in this analysis was related to population density. All Congressional districts are required to have the same population (and this would still be the same approximately across states). Therefore, I considered mean population density (population per square meter) of each district. Figure 11 shows the violin plots by party for each of the three states.

Figure 11. Population density by Congressional district by party in Ohio, Maryland and New York.



Calculated from UCLA Congressional shape files.

In all three states, the mean population density of Republican districts is less than the mean population density of the Democratic districts. This is understandable given the differences between rural and city-dwelling voters and their particular party leanings. It's possible that the range of population densities could be a feature, however, it does not initially appear to distinguish between these three stats. I think it would probably be more revealing to have within-district ranges to consider.

V. Discussion & Conclusion

The goal for this analysis was to begin to get some insights as to which features could be calculated from Congressional district shape files, together with election and Census data. That goal was achieved at least to some extent. Several candidate features were obtained. However, more are possible and future work will continue to develop methods for calculating these features. Since this analysis has never been attempted before, many possible features will need to be considered to determine which features lead to the best results. It is better to start with more than you need and cutting back after the fact, than having too few to start with.

It does seem from the calculations discussed in this analysis that feature engineering is a useful thought experiment about districts if nothing else. Trying to establish features that map onto badly gerrymandered districts can by themselves lead to better district maps as a way to potentially thwart the most egregious efforts at cracking and packing. Being able to quantify something that has previously seemed so fuzzy—I know it when I see it—is an important step forward that allows these features to be considered not only in my own analysis, but in analyses that approach the problem from a map-drawing perspective.

More work on population features within districts and accounting for state, county and natural boundaries is needed. Some of these feature calculations may require significantly more computing power. For features that were successfully calculated in this project, extending the analysis to all 43 states that had more than one Congressional district in 2012 will need to be attempted as well. Creating more efficient code to accomplish this task is also necessary to continue the work.

Bibliography

[1] M. Eagles, R. S. Katz and D. Mark, "GIS and Redistricting: Emergent technologies, social geography, and political sensibilities," *Social Science Computer Review,* vol. 17, no. 1, pp. 5-9, Spring 1999.

[2] M. J. Salling, "Public Participation Geographic Information Systems for Redistricting: A Case Study in Ohio," *Urban and Regional Information Systems Association Journal,* vol. 23, no. 1, pp. 33-40, 2011.

[3] R. Morrill, "Bushmanders and Bullwinkles: How Politicians Manipulate Electronic Maps and Census Data to Win Elections," *Annals of the Association of American Geographers,* vol. 91, no. 4, pp. 772-773, 2001.

[4] D. C. Bowen, "Boundaries, Redistricting Criteria, and Representation in the U.S. House of Representatives," *American Politics Research,* vol. 42, no. 5, pp. 856-895, 2014.

[5] D. Joshi and L.-K. Soh, "Redistricting Using Constrained Polygonal Clustering," *IEEE Transactions on Knowledge and Data Engineering,* vol. 24, no. 11, pp. 2065-2079, 2012.

[6] N. Xiao, "A Unified Conceptual Framework for Geographical Optimization Using Evolutionary Algorithms," *Annals of the Association of American Geographers,* vol. 98, no. 4, pp. 795-817, 2008.

[7] J. Suzuki, Constitutional Calculus: The Math of Justice and the Myth of Common Sense, Baltimore: Johns Hopkins University Press, 2015.

[8] M. Eagles, R. S. Katz and D. Mark, "Controversies in Political Redistricting, GIS, Geography, and Society," *Political Geography,* vol. 19, pp. 135-139, 2000.

[9] B. Forest, "Information Sovereignty and GIS: the evolution of "communities of interest" in political redistricting," *Political Geography,* vol. 23, pp. 425-451, 2004.

[10] C. Bong, W. Chai and C. Wong, "State-of-the-Art Multiobjective Metaheuristic for Redistricting," *IEEE,* pp. 763-769, 2004.

[11] L. Hardy, A. Heslop and S. Anderson, Eds., Reapportionment Politics: The History of Redistricting in the 50 States, Beverly Hills, CA: Sale Publications, 1981.

[12] S. B. Parsons, W. W. Beach and M. J. Durbin, United States Congressional Districts and Data, 1843-1883, New York: Greenwood Press, 1986.

[13] B. Taper, Gomillion versus Lightfoot: The Right to Vote in Apartheid Alabama, Tuscaloosa: Fire Ant Books, 1963.

[14] T. W. Yarbrough, Race and Redistricting: The Shaw-Cromartie Cases, Lawrence, KS: University Press of Kansas, 2002.

[15] J. G. Behr, Race, Ethnicity and the Politics of City Redistricting: Minority-Opportunity Districts and the Election of Hispanics and Blacks to City Councils, Albany, NY: State University of New York Press, 2004.

[16] E. C. Griffith, The Rise and Development of the Gerrymander, Chicago: University of Chicago Press, 1907.

[17] M. J. Anderson, The American Census: A Social History, 2nd ed., New Haven, CT: Yale University Press, 1988.

[18] L. M. Schwab, The Impact of Congressional Reapportionment and Redistricting, Lanham, MD: University Press of America, 1988.

[19] G. F. Moncrief, Ed., Reapportionment and Redistricting in the West, New York: Lexington Books, 2011.

[20] North Carolina General Assembly Legislative Services Office, Reapportionment and Redistricting Materials, Raleigh, NC: North Carolina General Assembly Legislative Services Office, 1981.

[21] J. Barker, "Supreme Court sidesteps definitive ruling in Maryland congressional redistricting case," 18 June 2018. [Online]. Available: http://www.baltimoresun.com/news/maryland/politics/bs-md-supreme-court-districts-20180618-story.html. [Accessed 2 September 2018].

[22] S. Bickerstaff, Lines in the Sand: Congressional Redistricting in Texas and the Downfall of Tom Delay, Austin: University of Texas Press, 2007.

[23] J. Buchman, Drawing Lines in Quicksand: Courts, Legislatures and Redistricting, New York: Peter Lang, 2003.

[24] C. S. I. Bullock, Redistricting: The Most Political Activity in America, Lanham, MD: Rowman & Littlefield Publishers, 2010.

[25] D. T. Cannon, Race, Redistricting, and Representation: The Unintended Consequences of Black Majority Districts, Chicago: University of Chicago Press, 1999.

[26] O. Guest, F. J. Kanayet and B. C. Love, "Gerrymandering and Computational Redistricting," 13 November 2017. [Online]. Available: https://arxiv.org/abs/1711.04640?. [Accessed 12 August 2018].

[27] G. A. Keith, Ed., Rotten Boroughs, Political Thickets and Legislative Donnybrooks: Redistricting in Texas, Austin: University of Texas Press, 2013.

[28] F. L. Kury, Gerrymandering: A Guide to Congressional Redistricting, Dark Money and the U.S. Supreme Court, Lanham, MD: Rowman & Littlefield Publishers, 2018.

[29] T. E. Mann and B. E. Cain, Eds., Party Lines: Competition, Partisanship and Congressional Redistricting, Washington, D.C.: Brookings Institution, 2005.

[30] W. J. Miller and J. D. Walling, Eds., The Political Battle of Congressional Redistricting, New York: Lexington Books, 2013.

[31] M. E. Rush, Ed., Voting Rights and Redistricting in the United States, Westport, CT: Greenwood Press, 1998.

[32] D. Stebenne, "Re-mapping American Politics: The Redistricting Revolution Fifty Years Later," *Origins,* vol. 5, no. 5, February 2012.

[33] J. Winburn, The Realities of Redistricting: Following the Rule and Limiting Gerrymandering in State Legislative Redistricting, Lanham, MD: Rowman & Littlefield Publishers, 2008.

[34] F. Caro, T. Shirabe, M. Guignard and A. Weintraub, "School redistricting: embedding GIS tools with integer programming," *Journal of the Operational Research Society,* vol. 55, pp. 836-849, 2004.

[35] G. Siegel-Hawley, "Colored Boundaries: School Redistricting in a Racially Changing Suburb," in *American Sociological Association*, 2010.

[36] B. Bozkaya, E. Erkut, D. Haight and G. Laporte, "Designing New Electoral Districts for the City of Edmonton," *Interfaces,* vol. 41, no. 6, pp. 534-547, 2011.

[37] P. Miller, S. Kimbrough and J. Schacht, "Simulating Redistricting in the Classroom: A Binding Arbitration Decision Game Using Louisiana Census Data," *American Political Science Association,* pp. 664-668, 2018.

[38] M. Monmonier, Bushmanders & Bullwinkles: How Politicians Manipulate Electronic Maps and Census Data to Win Elections, Chicago: University of Chicago Press, 2001.

[39] J. B. Lewis, B. DeVine, L. Pritcher and K. C. Martis, "United States Congressional District Shapefiles," 2013. [Online]. Available: http://cdmaps.polisci.ucla.edu/. [Accessed 1 December 2020].

[40] Ohio Secretary of State, "ELECTION RESULTS AND DATA," 2020. [Online]. Available: https://www.ohiosos.gov/elections/election-results-and-data/. [Accessed 1 December 2020].

[41] "MIT Election Data and Science Lab," 2 December 2020. [Online]. Available: https://electionlab.mit.edu/data. [Accessed 6 December 2020].

[42] US Census, "The Census Bureau Data," 2020. [Online]. Available: https://data.census.gov/cedsci/. [Accessed 1 December 2020].

[43] USGS, "Data and Tools: GIS Data," 2020. [Online]. Available: https://www.usgs.gov/products/data-and-tools/gis-data. [Accessed 1 December 2020].

Datasets Used

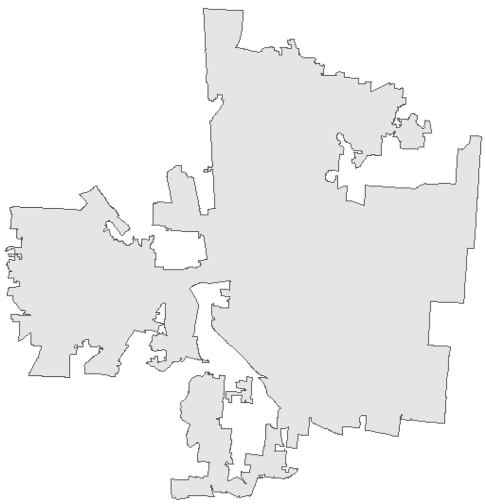
[1] MIT Election Data and Science Lab, "U.S. House 1976–2018," 2017. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IG0UN2. [Accessed 2 April 2020].

[2] J. B. Lewis, B. DeVine, L. Pitcher and K. C. Martis, "Digital Boundary Definitions of United States Congressional Districts, 1789-2012," 2013. [Online]. Available: http://cdmaps.polisci.ucla.edu. [Accessed 2 April 2020].

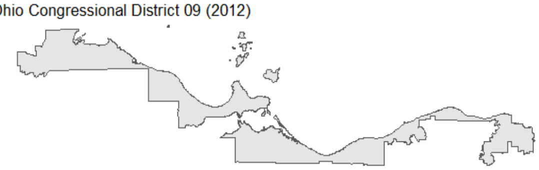
[3] U.S. Census Bureau, "2010 Decennial Census Data on Race," 2011. [Online]. Available: https://data.census.gov/cedsci/table?g=0400000US36.100000&y=2010&tid=DECENNIALSF12010.H8&hidePreview=false. [Accessed 2 April 2020].

Appendix. Maps of the most gerrymandered Congressional districts in 2012, selected by the author, in Ohio, Maryland and New York.

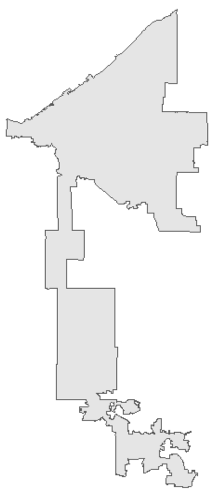
Ohio Congressional District 03 (2012)



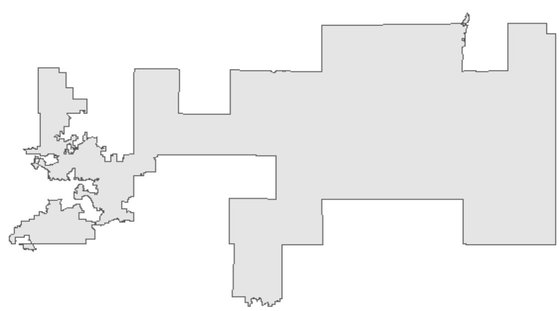
Ohio Congressional District 09 (2012)



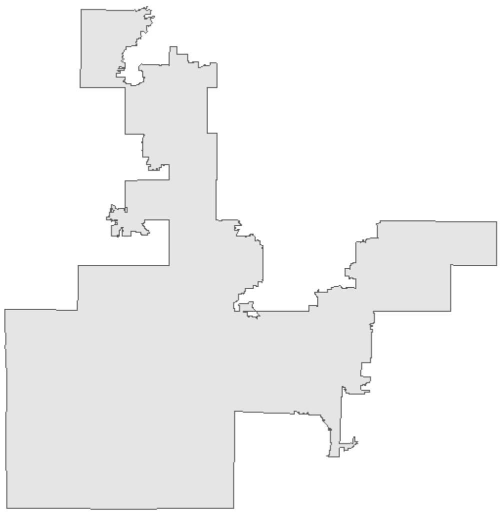
Ohio Congressional District 11 (2012)



Ohio Congressional District 13 (2012)



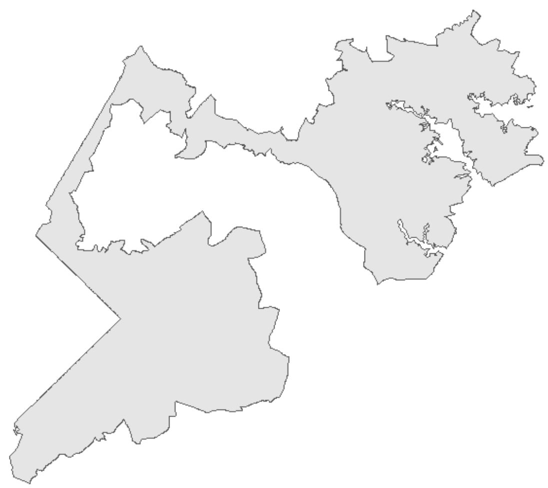
Ohio Congressional District 16 (2012)



Maryland Congressional District 02 (2012)

Maryland Congressional District 03 (2012)

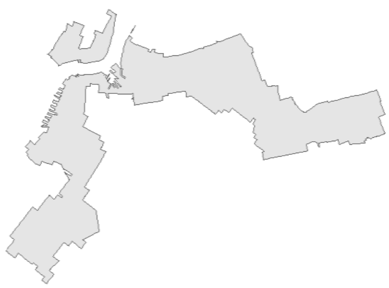
Maryland Congressional District 04 (2012)



Maryland Congressional District 07 (2012)

Maryland Congressional District 08 (2012)

New York State Congressional District 07 (2012)



New York State Congressional District 08 (2012)

New York State Congressional District 10 (2012)