

Analyzing the Effects of Partisan Correlation on Election Outcomes Using Order Statistics

Claire Wiebe

April 24, 2019

A thesis submitted to the Department of Mathematics for Graduation with Distinction

Duke University

Durham, North Carolina

Abstract

The legislative representation of political parties in the United States is dependent not only on way that legislative district boundaries are drawn, but also on the way in which people are distributed across a state. That is, there exists a level of partisan correlation within the spacial distribution of an electorate that affects legislative outcomes. This work aims to study the effect of this partisan clustering on election outcomes and related metrics using analytic models and order statistics. Two models of North Carolina, one with a uniformly distributed electorate and one with a symmetrically clustered electorate, are considered both independently and in comparison. These models are used to study expected election outcomes, the proportionality of legislative representation for given state-wide vote fraction, and the sensitivity of vote share to seat share across different correlation length scales. The findings provide interesting insight into the relationship between district proportionality and legislative proportionality, the extent to which the minority party is expected to be underrepresented in seat share for given state-wide vote share and correlation length, and the extent to which the responsiveness of seat share is dependent on state wide vote share and correlation length.

Contents

1	Partisan Clustering and Election Outcomes	6
2	Uniform Distribution	9
3	Symmetric Clustering	17
4	Comparing Models	25
5	The Impact of Correlation Length	30
6	Appendix	33
6.1	Uniform Distribution Formulas and Additional Figures	33
6.2	Symmetric Clustering Formulas and Additional Figures	35
6.3	Asymmetric Clustering Exploration	37
7	References	39

List of Figures

1	Sample District Under Uniform Distribution	10
2	Uniform Distribution District Distributions	11
3	Slope of the District Distribution Plot Under Uniform Distribution	12
4	Safe Districts by Party Under Uniform Distribution	13
5	Expected Number of Republican Seats Under Uniform Distribution	14
6	Proportionality Ratio Under Uniform Distribution	15
7	Vote-Seat Curve Under Uniform Distribution	16
8	Sample District Under Symmetric Clustering	18
9	Symmetric Clustering District Distributions	19
10	Slope of the District Distribution Plot Under Symmetric Clustering	20
11	Safe Districts by Party Under Symmetric Clustering	20
12	Expected Number of Republican Seats Under Symmetric Clustering	21
13	Proportionality Ratio Under Symmetric Clustering	22
14	Vote-Seat Curve Under Symmetric Clustering	23
15	Slope of District Distribution Plot Across Correlation Length	26
16	Average Number of Republican Seats Across Correlation Length	27
17	Proportionality Ratio Across Correlation Length	28
18	Vote-Seat Curve Across Correlation Length	29
19	Probability of Each Party Winning Each District Under Uniform Distribution	35
20	PDF of the Slope Under Uniform Distribution	35
21	Probability of Each Party Winning Each District Under Symmetric Clustering	37
22	PDF of the Slope Under Symmetric Clustering	37
23	North Carolina Precinct Histogram Courtesy of Jay Patel	38

Acknowledgements

I would first like to thank my advisor Professor Jonathan Mattingly and Professor Gregory Herschlag for their guidance and instruction throughout the course of this project. They supported my initial exploration and helped me acquire the mathematical research and computer programming skills I hoped to gain from doing this work. Throughout this project, they helped me think through how to structure my models, focus my points of study, and construct a narrative story to connect and convey my findings.

I would also like to thank Ella van Engen and Jay Patel, my collaborators in the Natural Packing group of the Quantifying Gerrymandering Bass Connections class. Working with Jay and Ella last semester to begin exploring the effects natural packing and different models provided a strong foundation for my research to build off of this semester. They were continuously a useful resource to brainstorm ideas and discuss findings.

Lastly I would like to thank my family and friends for their support throughout this process and for listening to me panic when a formula wasn't working and celebrate with me when it finally did.

1 Partisan Clustering and Election Outcomes

The legislative process in the United States is structured as a system of district-based representation and winner-take-all elections. The representation of political parties is thus dependent on how districts in a state are drawn and how populations are divided across those districts. Within states, people are not evenly distributed: urban areas are geographically compact and dense in population, rural areas are geographically expansive and sparsely populated, and suburban areas are intermediate. These areas also have different partisan compositions with Democrats mostly concentrated in cities and Republicans more evenly distributed in suburban and rural areas such that urban areas are liberal, rural areas are conservative and suburban areas are again intermediate (Chen & Rodden). This non-uniformity introduces a level of correlation between a person and the people around them with respect to political affiliation. Further, the legislation surrounding the drawing of representative districts encourages these correlations to be preserved by mandating districts be compact and continuous. Many states also mandate or encourage the preservation of cities within a district, unless a city split is forced to maintain equal population across districts. For the most part these natural features of human geography and their partisan asymmetries are, from the perspective of legislative map-makers, immutable. Thus, even fairly constructed, non-gerrymandered maps are still constrained and influenced by partisan population clustering. Because people are correlated with the people around them and because districts are drawn to keep people who live near each other together, the way in which people are distributed and clustered across a state affects the composition of legislative districts and subsequently electoral outcomes absent any gerrymandering.

Foundational research into the effect of partisan clustering on election outcomes shows how natural packing introduces bias into the legislative process. In their research, Chen & Rodden explore the divergence between state-wide vote share and seat-share seen in many urbanized states. Using precinct-level voting data from the 2000 Presidential election, they ran simulations drawing districts with a neutral algorithm to produce an ensemble of

possible fair and non-Gerrymandered outcomes. While there was no partisan bias in the drawing of districts, there was a partisan asymmetry in population clustering that created an inherent bias in favor of Republicans. Specifically, Democrats were inefficiently clustered in homogeneous neighborhoods in dense cities or small agglomerations such that a Democrat precinct was more likely to be near a Democrat precinct than a Republican precinct was to be near a Republican precinct; Democrats were more highly spatially correlated. Chen & Rodden showed how Republicans were still able to capture a seat share higher than their state-wide vote share due to partisan clustering and continuous and compact districts. This effect gives each state a different baseline partisan seat distribution.

Similar results were also shown by Cottrell using a numeric model in which symmetric and asymmetric clustering were introduced via Shelling's agent-based model of segregation and districts were drawn using a neutral k-means++ algorithm. Sampling from this numeric model under conditions of no clustering, symmetric clustering, and asymmetric clustering showed that clustering flattens the vote-seat curve, with the effect most pronounced under asymmetric clustering. The vote-seat curve plots vote share against seat share so the flattening of this curve means that seat share is less responsive to vote share such that a party can increase their state wide vote share without winning any more seats. Similarly, comparing across the same vote share clustering reduces the seat share again with asymmetric clustering having the greatest effect. Because of asymmetric clustering, the less clustered party, the Republicans, were able to win a majority of the seats without a majority of the votes and increasing the vote share of the the more clustered party, the Democrats, did not translate into a proportional increase in seat share.

Clustering, however, does not have universally negative effects and Democrats are not always disadvantaged by natural packing. For state legislative elections, rather than Congressional elections, states are divided into a greater number of smaller districts so smaller Democratic pockets become relatively large enough to dominate the surrounding Republican areas in the district. Eubank & Rodden showed that spacial inefficiency is a result of

cities being too large or too small relative to the size of districts so under certain conditions, clustering can be efficient. They further show that there are instances in which clustering can be beneficial for and result in overrepresentation of the clustered party. In conservative states, Democratic clustering in cities can produce more Democrat districts than would be expected with a low state wide vote share. The effects of clustering are dependent on the scale of districts, scale of cities, and the state-wide vote share.

Often, when an election results in non-proportional representation for a state, the results are attributed to partisan gerrymandering, an intentional manipulation of what would otherwise be a fair outcome. However, because people are distributed across states in such a way that there exists a partisan spatial correlation between people and the people around them, and because districts are drawn in such a way that preserves these correlations, elections can generate non-proportional results due to natural, immutable features of human geography. Understanding how these partisan distributions operate within districts, and the effects that this produces in legislative representation gives insight into the natural biases that exist in our electoral system. As mentioned, prior research into the effects of clustering has drawn from real world election data or sampled from numeric models, but this paper will examine correlation effects using analytic models rooted in mathematical statistics, and with specific focus on North Carolina.

North Carolina is a state that receives a lot of attention for its highly gerrymandered legislative districts. Famously, Representative David Lewis, a Republican leading the re-districting efforts in North Carolina, said that the only reason maps were drawn drawn to give Republicans ten seats and Democrats three seats was because they couldn't find a way to draw an eleven-two map. North Carolina is a swing state so one would expect the two parties to be similarly competitive in legislative representation. Because North Carolina is so highly gerrymandered and because the two parties are relatively close in vote share, it is particularly interesting to study the underlying geographic features of North Carolina to explore how gerrymandering may be building off of an existing partisan bias that creates dis-

proportional representation. The analytic models in this analysis apply the metrics of North Carolina under no clustering and symmetric clustering conditions to generate theoretically predictive results on election outcomes and related measures including the distribution of districts, legislative proportionality, and vote-seat curve.

2 Uniform Distribution

First, the condition of no clustering is considered. Distributing people uniformly across a state creates a base case scenario in which all geographic structures, natural packing, and correlations between people have been removed. This model considers a state with n districts and k units within each district. Each unit represents a uniform voting block; every person in the block votes the same way, so units are either 100% Democrat or 100% Republican, and units are of equal size. Regardless of the value of k , districts have a fixed number of people such that when k is small, there are many people within each unit and when k is big, there are few people within each unit. Units can also be thought of as a correlation length scale with smaller values of k meaning there is a larger correlation length, and people are more highly correlated, and smaller values of k meaning there is a smaller correlation length, and people are less correlated. Each unit is Bernoulli random variable that is assigned Democrat with probability p and Republican with probability $(1 - p)$ so the number of Democrat units within each district is a binomial random variable with sample size k and probability of success p . In order to model uniform distribution, large values of k are considered so that there are many small units within each district. This creates a random spacing of Democrats and Republicans across each district and reduces the correlation each person has with the people around them. With these large values of k , the Democrat vote fraction within each district can be approximated using a normal distribution with mean p and variance $\frac{p(1-p)}{k}$. In order for the normal to be a reasonably accurate approximation by the Central Limit Theorem, values of k are taken such that $kp \geq 5$ and $k(1 - p) \geq 5$. This analysis focuses

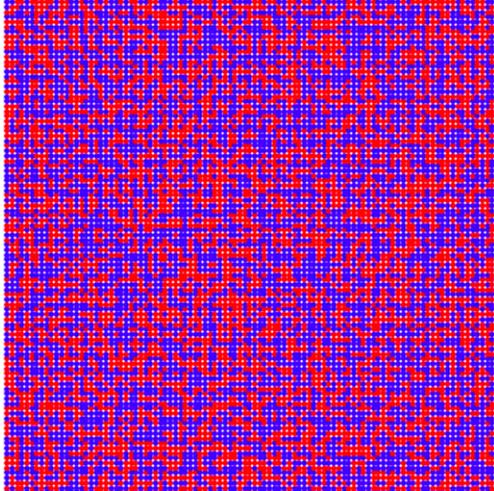


Figure 1: Sample District Under Uniform Distribution

on North Carolina so $n = 13$, the number of Congressional districts, $p = .47$, the state wide Democratic vote fraction in the 2010 election, and $k \geq 11$. Figure 1 gives a sample district under uniform distribution.

In this model, rather than distributing people across a state and drawing districts on top of that, districts are fixed structures and parameters are set to determine the way in which people are distributed within them. Because every unit in the state has the same probability to be Democrat or Republican, this is a reasonable set up and avoids the need to generate algorithmically unbiased districts. Election outcomes and related metrics are then obtained not from running an election simulation and gathering data, but from predictive probabilistic distributions using order statistics. Order statistics are a useful and informative way to model these districts as there are closed formulas for the probability density function, cumulative density function, and joint distribution. Taking the normal approximation to model district vote share, districts are ordered by their Democrat vote fraction with $D_{(1)}$ as the district with lowest Democrat vote and $D_{(n)}$ as the one with the highest, and subsequent results are derived.

To begin understanding what districts look like when the electorate is uniformly distributed and uncorrelated, Figure 2 shows the probabilistic distribution of ordered districts.

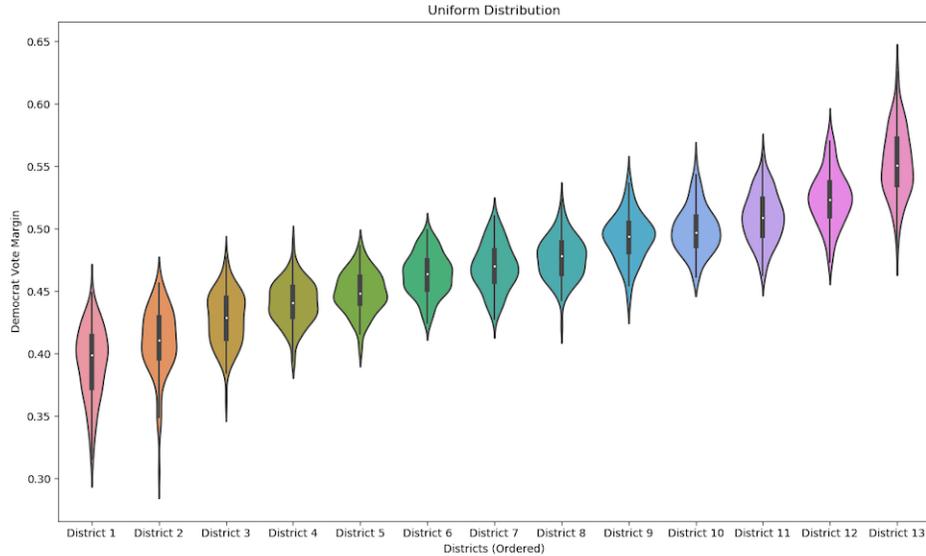


Figure 2: Uniform Distribution District Distributions

This plot was generated with $k = 100$, 100 samples per district, and visualized using a violin plot, which smooths the data via a kernel density estimator. The middle districts are clustered around .47, which is the mean, and have less variance than do the districts at the extremities. Although $k = 100$ and all units in all districts are equally likely to be Democrat or Republican, there is still a bit of variance between each of the districts: the expected value of $D_{(1)}$ is .3867 and the expected value of $D_{(n)}$ is .5532. This variance between districts is a useful study as it indicates the level of partisan spatial correlation, with lower correlation leading to less variation between districts.

To measure and quantify the variation between districts, the slope of the district distribution plot is used. The slope is calculated by taking the expected value of $D_{(1)}$ from the expected value of $D_{(n)}$ and dividing by n . The slope can thus range from $\frac{1}{n}$, when $D_{(n)} = 1$ and $D_{(1)} = 0$, and 0 when $D_{(n)} = D_{(1)}$. Figure 3 plots the slope against k and shows how as k increases, the slope decreases. This is a result of reduced variation of expected outcomes within each district and therefore reduced variation of expected outcomes between districts. Further, the shape of the slope-curve gives insight into the rate at which correlation is removed as k increases. From Figure 3 one can see that there is a decreasing

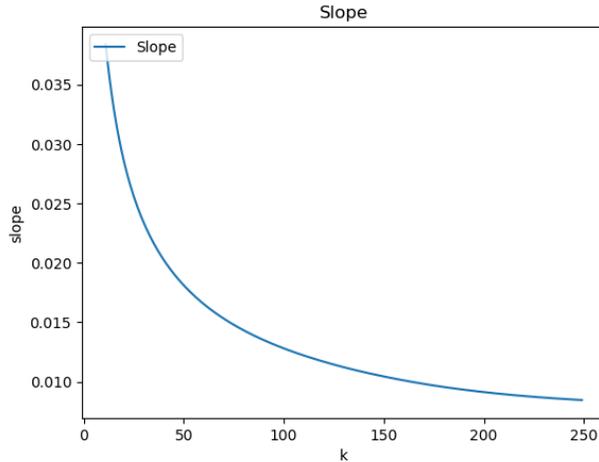


Figure 3: Slope of the District Distribution Plot Under Uniform Distribution

effect with decreasing correlation. That is, increasing from 11 and 12 units per district has a more significant effect on the expected outcome of districts than does increasing from 200 to 201 units. Additionally, as district vote margins grow more similar to each other, they also grow more similar to the mean, which is the state-wide vote fraction p . The way in which correlation length affects the makeup of districts has subsequent implications for election results and legislative representation.

To begin exploring how a uniform distribution impacts election outcomes, the competitiveness of each district's elections is considered. There is an inherent element of fairness to having competitive districts, in which either party has a chance to win, as opposed to safe districts, in which one party is all but guaranteed to win and the other party is shut out. For this exploration, safe districts are defined as those that are predicted to be won by a particular party with over 60% probability. Remaining districts that do not qualify as safe, and therefore have a 40% - 60% of going to either party, are considered competitive. Figure 4 shows the number of safe Democrat and Republican districts taking k from 11 to 1000. At large values of k , the Democrat vote margin of all districts approaches $p = .47$ with lower variance so more districts are expected to be Republican with greater likelihood and eventually all n districts are safe Republican districts. On the converse, this results in no districts that are either safely Democrat or even competitive. This is an interesting effect

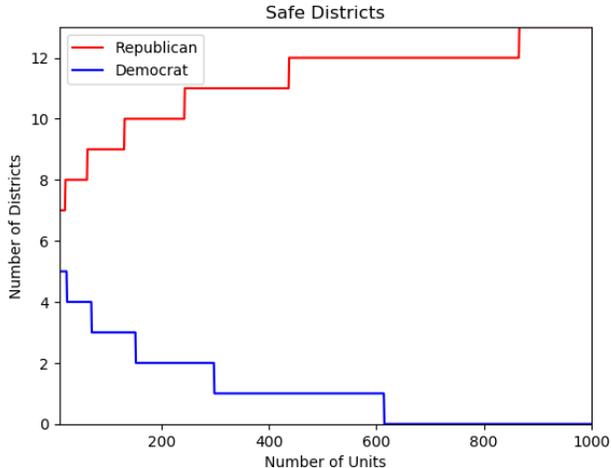


Figure 4: Safe Districts by Party Under Uniform Distribution

of reduced correlation meaning that despite having a seemingly competitive state-wide vote fraction, Democrats are unable to be competitive in any districts and are shut out of any legislative representation.

Another way to consider legislative representation under a uniform distribution, rather than defining districts as either safe or competitive, is to take the expected number of seats won at low correlation lengths. For the purposes of this exploration, this is done by calculating the expected number of Republican seats won using the probability of Republicans winning i seats for $i \in n$. Note, that because this model considers a two-party system in which constituents vote either for Democrats or Republicans, it is not substantively important whether this is examined as a metric of the Democrats or Republicans as one is the opposite of the other. The probability of the Republicans winning i seats is the probability that $D_{(1)}, \dots, D_{(i)}$ have Democrat vote fraction less than .5 and $D_{(i+1)}, \dots, D_{(n)}$ have Democrat vote fraction greater than .5, which is calculated using the joint density function between $D_{(i)}$ and $D_{(i+1)}$. Figure 5 plots the expected number of Republican seats won across $11 \leq k \leq 800$ on a linear plot (a) and a linear-log plot using $-\log(\frac{1}{k})$ (b), which expands the graph around smaller k values and helps to make clear the relationship between correlation and expected vote share. First, it is clear that as correlation length decreases, the expected number of Republican seats increases to 13. In a winner-take-all election, if every district has a uniformly

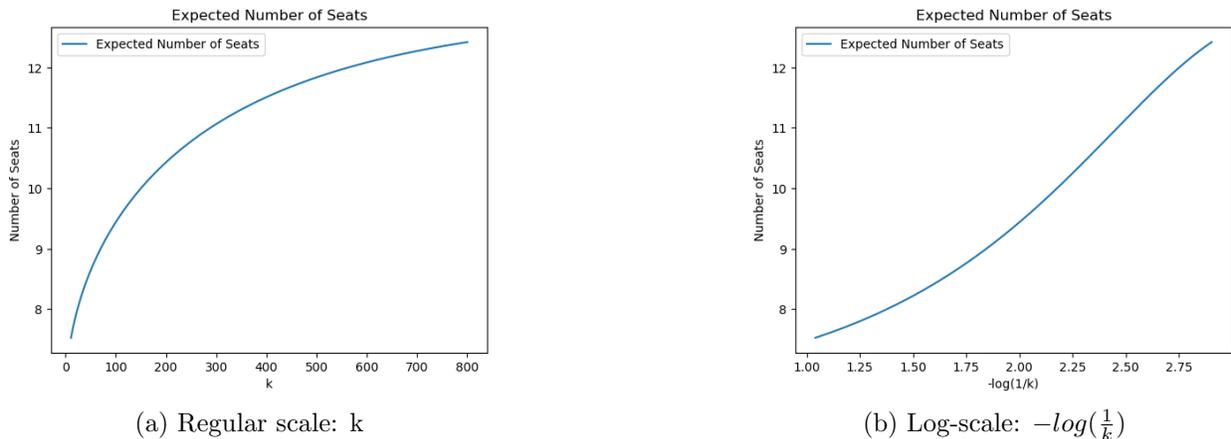


Figure 5: Expected Number of Republican Seats Under Uniform Distribution

distributed electorate then every district will be a majority of the majority party and majority party will win every seat. Therefore, as partisan clusters in districts become smaller, the Republicans are able to win 100% of the seats with only 53% of the vote. Additionally, the shapes of the plots in Figure 15 show the larger effect of higher correlation lengths; in plot (a) there is a much steeper slope at low k values translating to a nearly linear plot in (b). As the spatial correlation between people in each district is removed, Republicans have an increasingly smaller gain in seat share. Examining the average number of seats won again shows the dichotomy between district proportionality and legislative proportionality, as perfectly proportional districts lead to completely uniform legislative representation.

To further understand and quantify proportionality under a uniform distribution, the ratio of seat share to vote share is considered. Taking $11 \leq k \leq 200$, the joint distribution is used to calculate the average number of Democrat seats won and then that is taken as a proportion of the n total seats and divided by the state-wide Democrat vote share p . This gives a ratio of proportionality where if vote share translates perfectly proportionally to seat share the ratio is 1, if vote share translates into disproportionately low seat share the ratio is less than 1, and if vote share translates into disproportionately high seat share the ratio is greater than 1. This analysis is done considering Democrat proportionality, which again is just opposite of Republican proportionality; if Democrats have disproportionately

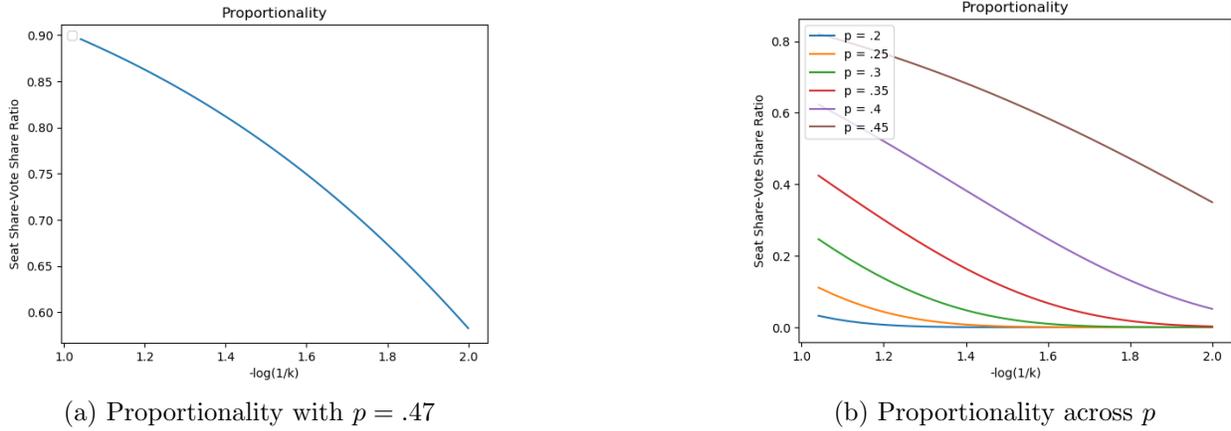


Figure 6: Proportionality Ratio Under Uniform Distribution

low seat share compared to vote share then Republicans have disproportionately high seat share. Figure 6 plots the proportionality ratio using a linear-log plot for North Carolina vote share $p = .47$ (a), and for varied levels of state-wide vote share (b). Across both plots, one can see that as k increases, the proportionality ratio decreases such that reduced clustering in districts leads to increasingly disproportionately low seat share for the minority party. Further, looking at plot (b) it is evident how much of an impact state-wide vote share has both on the starting level of proportionality at $k = 11$ and at the rate at which the proportionality ratio decreases. When there is a large differential in the state-wide vote fraction between the two parties, low values of p , the minority party starts with highly disproportionately low seat share and is quickly shut out of any legislative representation as districts become less clustered. Changing the state-wide vote split between the two parties has a notable impact on legislative outcomes and is thus an important point for further consideration.

Most of the analysis so far has been based off of the North Carolina Democrat vote share from the 2010 election, $p = .47$. However, vote share is highly susceptible to current events, political tides, and specific candidates in the race, so it is important to consider how responsive seat share is to changing vote share. That is, in a spatially uncorrelated distribution, how does increasing Democratic vote share affect Democratic seat share. This

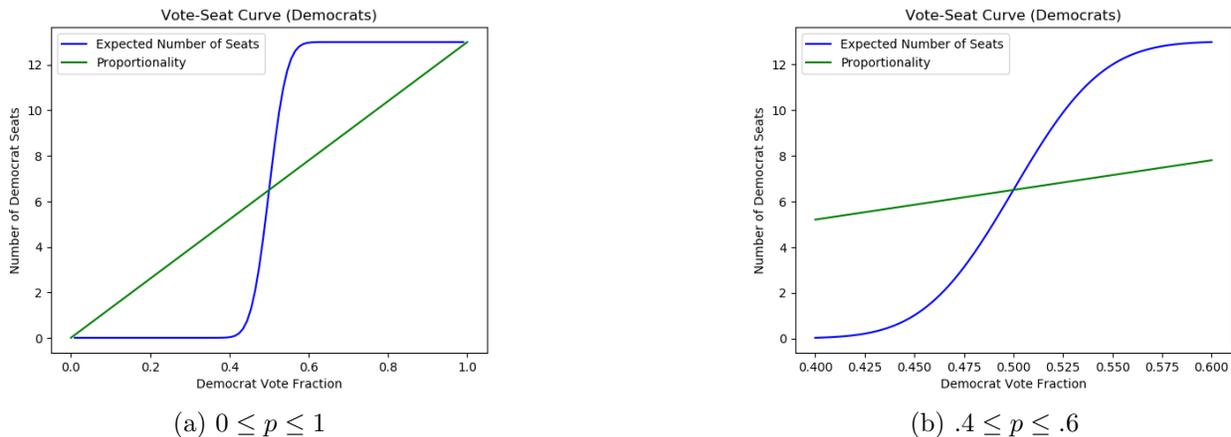


Figure 7: Vote-Seat Curve Under Uniform Distribution

is done using a graph known as the vote-seat curve that plots the expected number of Democrat seats against the Democrat vote fraction. Figure 7(a) shows the vote seat curve, with $k = 200$ and $0 \leq p \leq 1$, and a line of perfect proportionality, and Figure 7(b) shows the same curve zoomed in to $.4 \leq p \leq .6$. Under a uniform distribution, when Democrats have less than 40% of the vote share they are shut out of any legislative representation, and when they have over 60% of the vote share they are over-represented and expected to win every seat. In the range between 40% and 60% vote share, the curve is steep so seat share is highly elastic and highly responsive to changes in vote share. Notably, under a no-correlation model, either party needs 50% of the vote to be expected to win 50% of the seats, and a party cannot win a majority of the seats without a majority of the vote.

In the uniformly distributed model, there is no concept of spatial geography or partisan clustering and there is little correlation, with respect to political affiliation, between a person and the people around them. It acts as a base case where naturally occurring factors, like clustering and non-uniform population distribution, that are immutable features in the real world, have been removed. As k increases, the variance of expected outcomes within each district decreases so the vote margin within each district approaches perfect proportionality at the mean, there is increasingly little difference between the districts, and more districts become safely Republican. Notably, because the United States operates under a

winner-take-all system, as k increases, perfect proportionality within districts translates to complete uniformity in legislative representation as the majority party is able to win every district with only a small electoral majority. Further, when the vote share split between Democrats and Republicans is close, seat share is highly sensitive to changing vote share and a party needs 50% of the vote to win 50% of the seats or cannot get a majority of the seats without a majority of the vote. The uniform distribution model serves as a useful base case in which natural immutable facets of human geography and partisan clustering have been removed. Next, the case of symmetric clustering is considered by implementing the same model framework but with higher correlation lengths, so there are fewer units per district and therefore large correlated voting blocks.

3 Symmetric Clustering

Under symmetric clustering there exists an increased correlation, with respect to political affiliation, between a person and the people around them, and correlation occurs equally for each party. This model has the same basic framework of the previous model; it considers a state with n districts, k units within each district, and districts are fixed structures with parameters set to determine the way in which people are distributed within them. Each unit is a uniform voting block that is a Bernoulli random variable assigned Democrat with probability p and Republican with probability $(1 - p)$, so the number of Democrat units within each district is a binomial random variable with sample size k and probability of success p . Now, to model symmetric clustering, only high correlation lengths, small values of k , are used; having these larger uniform voting clusters increases the correlation between a person and the people around them. Units have the same structure and function as in the uniform distribution case but now there are a few large clusters within each district rather than many small clusters. Because the number of Democrat units in a district is modeled using a discrete binomial distribution, the Democrat vote fraction within each

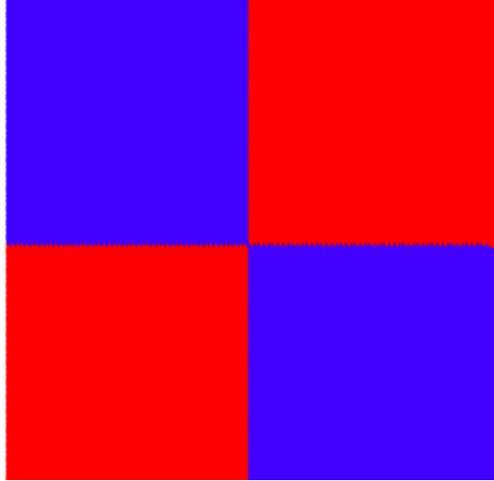


Figure 8: Sample District Under Symmetric Clustering

district only takes on discrete values. So, for $0 \leq m \leq k$ if there are m Democrat units in a district than the Democrat vote share of that district is $\frac{m}{k}$. The average vote share of each district is p and the variance is $\frac{p(1-p)}{k}$. Predicted election outcomes and related metrics are obtained through order statistics, but with modified formulas to account for the possibility of ties between districts that can occur due to the discrete nature of the underlying binomial random variable. The explorations run with this model are again based off of North Carolina 2010 election data: $n = 13$, $p = .47$, and $1 \leq k \leq 10$. Figure 8 shows a sample district under the symmetric clustering condition.

To first explore the symmetric clustering model, the distribution of the districts is considered. Figure 9 shows the distribution of each of the ordered districts taking $k = 5$, 100 samples per district, and visualized using a violin plot. In this plot, instead of seeing the smooth bell curve shape like in the uniform distribution, there are several smaller peaks, which is a result of using a discrete binomial distribution to model district vote share. The bars in the distribution curves correspond to the discrete values obtained and the length of the bars corresponds to the likelihood of seeing each value. Further, while the middle districts are again clustered around the mean Democrat vote share, $p = 47$, under symmetric clustering there is a much wider range of values that the districts can take on. Specifically, the expected value of $D_{(1)}$ is .1218 and the expected value of $D_{(n)}$ is .8286, a much larger

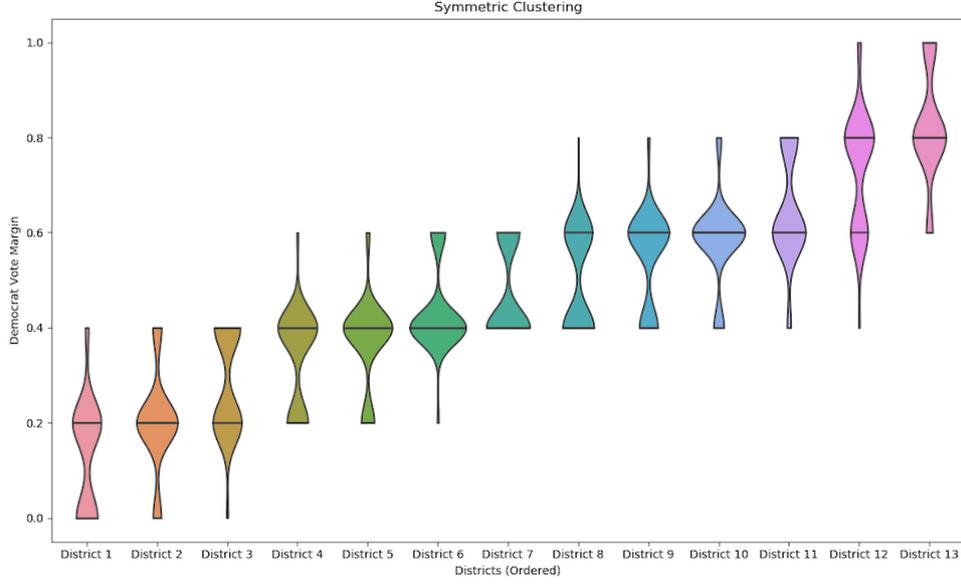


Figure 9: Symmetric Clustering District Distributions

spread than the .3867 - .5532 seen in the uniform distribution model. This is an expected result considering the variance of outcomes within each district. Under symmetric clustering, there are fewer trials per district, or small values of k , so there is much more deviation from the mean vote share and therefore a wider range of potential outcomes per district.

Measuring and quantifying the variation between districts is again done using the slope of the district distribution curve. Based on the district distributions in Figure 9, one would expect to see a large slope as a result of the large variance within and between districts resulting from low values of k . The slope can again range from $\frac{1}{n}$, if $D_{(n)} = 1$ and $D_{(1)} = 0$, and 0, if $D_{(n)} = D_{(1)}$. Figure 10 plots the slope of the symmetric clustering distribution curve using the average value of $D_{(n)}$ and $D_{(1)}$ at high correlation lengths and validates the expectation for a large slope. At $k = 1$ the slope reaches its maximum value of $0.0769 = \frac{1}{n} = \frac{1}{13}$, and then as k increases the slope decreases at a relatively consistent rate. The effect of increasing k is fairly similar across the values of k considered because the structure of this model only supports a relatively small set of low k values as symmetric clustering. It is next worthwhile to consider the implications of the more varied districts that result under symmetric clustering on election outcomes and related metrics.

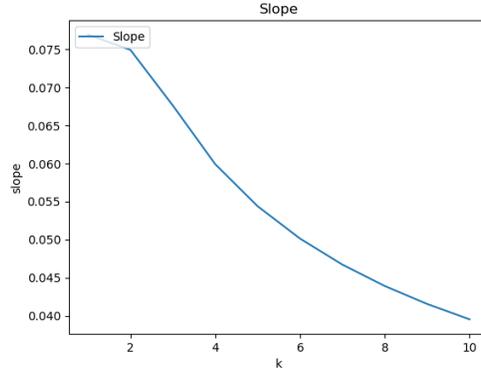


Figure 10: Slope of the District Distribution Plot Under Symmetric Clustering

To begin to analyze the effect of symmetric clustering on election outcomes and legislative representation, the metric of safe districts is examined using the same definitions and 60% cutoff for safe and competitive districts as in the previous section. Figure 11 plots the number of safe districts for Democrats and Republicans across high correlation length values considered as symmetric clustering. Initially, when $k = 1$, there are 6 safe Democrat districts and 6 safe Republican districts. This makes intuitive sense given that there is one unit per district, which means each district is either 100% Republican or 100% Democrat, and each unit is almost equally likely to go to either party. Thus, it is expected about half of the districts will be safely Democrat, half of the districts will be safely Republican, and the final district will be Democrat half the time and Republican half the time and is therefore is a competitive district. As k increases, the plot has a jagged shape which is a result of the

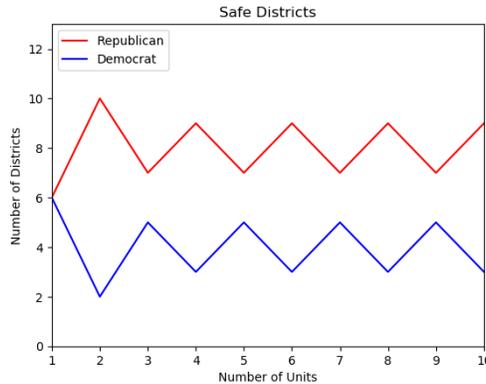


Figure 11: Safe Districts by Party Under Symmetric Clustering

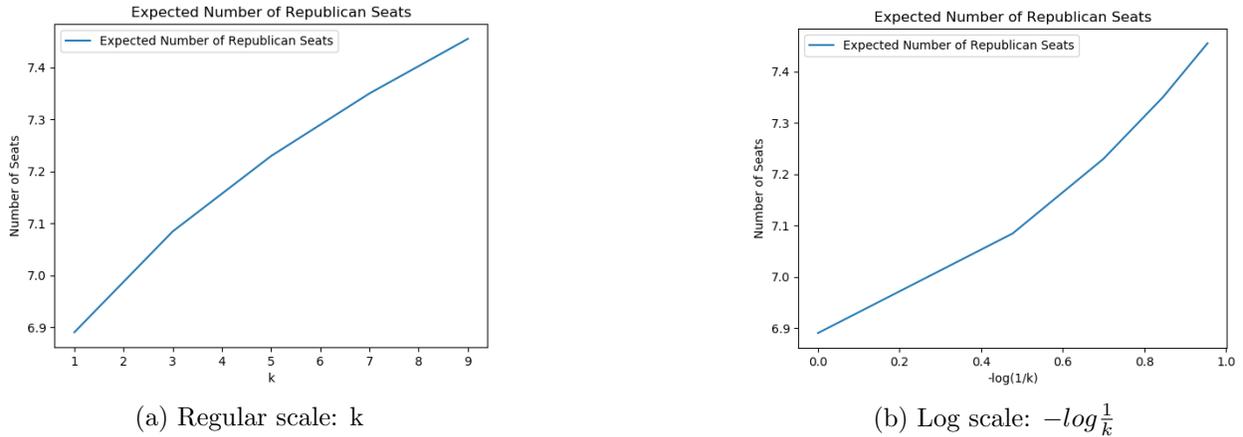
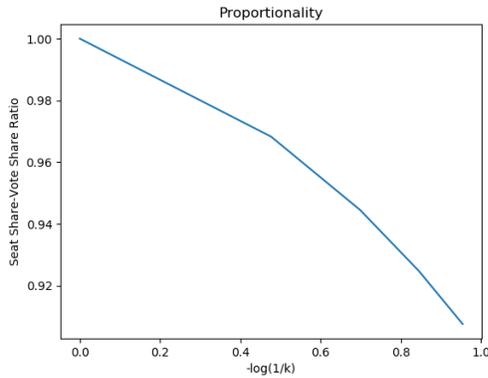


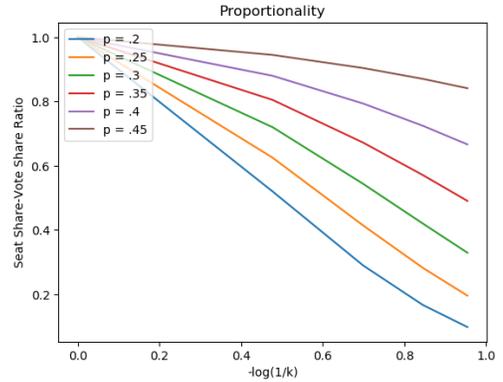
Figure 12: Expected Number of Republican Seats Under Symmetric Clustering

underlying discrete binomial random variable and set threshold cutoff for what is considered a safe versus a competitive district. However, it is more important here to consider the general trend of safe and competitive districts in symmetric clustering rather than the specific number of districts. In Figure 11, there is not the same transfer to n safe Republican districts and no safe Democrat or competitive districts as occurred in the case of uniform distribution. Instead, under symmetric clustering there are both more competitive districts and a more equitable distribution of safe Democrat and safe Republican districts.

The impact of symmetric clustering on legislative representation can also be examined using the expected number of seats won. This is done as a metric of the Republicans, which is again just the inverse of considering the metric for Democrats, by using the probability of Republicans winning i seats for $1 \leq i \leq n$ as the probability that $D_{(1)}, \dots, D_{(i)}$ have Democrat vote fraction less than .5 and $D_{(i+1)}, \dots, D_{(n)}$ have Democrat vote fraction greater than .5. To avoid the case of a tie, which can occur with even k due to the use of the discrete binomial random variable, only odd values of k are considered to evaluate the general trends and outcomes under symmetric clustering. Using the discrete joint probability density function between $D_{(i)}$ and $D_{(i+1)}$, Figure 12 plots the expected number of Republican seats won against k using a linear scale (a) and a linear-log scale (b). It is most notable from these graphs how proportional the seat share is under symmetric clustering. The state-wide Re-



(a) Proportionality with $p = .47$



(b) Proportionality across p

Figure 13: Proportionality Ratio Under Symmetric Clustering

publican vote margin is $(1 - p) = .53$ so proportional seat share is $.53n = 6.89$. At $k = 1$ the expected number of Republican seats is 6.89, so the total uniformity within districts translates to perfectly proportional legislative representation. Even as the correlation length decreases to $k = 10$, representation stays relatively close to proportional, especially considering that actual election outcomes are discrete. That is, if proportionality is 6.98 districts then either 6 or 7 districts could be considered proportional. When $k = 10$, the expected number is 7.46 translating to actual outcomes of either 7 or 8 districts, which is close to the proportional range. Under symmetric clustering, there is expected to be a more equitable distribution of legislative seats between Republicans and Democrats as high levels of partisan spatial correlation within districts translates to proportionality in legislative representation.

To further explore the notion of proportionality in representation under symmetric clustering, the proportionality ratio of seat share to vote share is considered. Taking $1 \leq k \leq 10$, the joint distribution is used to calculate the average number of Democrat seats won taken as a fraction of the total number of seats n and divided that by the state-wide Democrat vote share p . Again this metric is just the inverse of the Republican proportionality ratio so it is not substantively important which party is considered. Figure 13(a) plots this ratio using the standard $p = .47$ and 13(b) plots this ratio across values of $p \in [.2, .45]$, both on

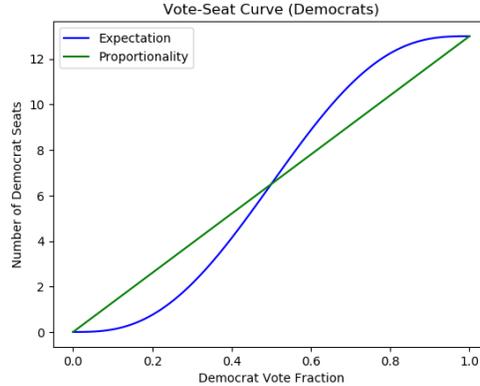


Figure 14: Vote-Seat Curve Under Symmetric Clustering

a linear-log scale. Under symmetric clustering, at any state-wide vote fraction when $k = 1$ there is perfect proportionality in legislative representation. As correlation length decreases, proportionality decreases with vote share translating into disproportionately low seat-share for the minority party. This effect is most pronounced for low values of p ; as k increases to 10, when $p = .47$ the proportionality ratio decreases towards .9, but when $p = .2$ the ratio decreases towards 0. The proportionality ratio gives useful insight into the relationship between proportionality, spatial correlation, and state-wide vote share: when there is complete district-wide clustering, there is always perfect proportionality in legislative representation but when there is within district symmetric clustering, only if the vote margin between the parties is close is there relative proportionality, otherwise, if there is a large disparity in vote margins between the parties, there is quickly a loss of proportionality.

Lastly, the vote-seat curve is considered to examine how the expected number of seats changes as the state-wide vote fraction p changes and measure how responsive the symmetric clustering model is to changes in vote share. Figure 14 plots the symmetric clustering vote-seat curve taking $k = 5$ and Democrat vote share p ranging from 0 to 1 against a line of perfect proportionality. Here, the vote-seat curve is relatively consistently shallow across p and seat share is responsive to vote share but with a weak elasticity. This curve shows that for nearly all values of p , the minority party cannot capture their proportional number of seats while the majority party is able to capture more than their proportional number of

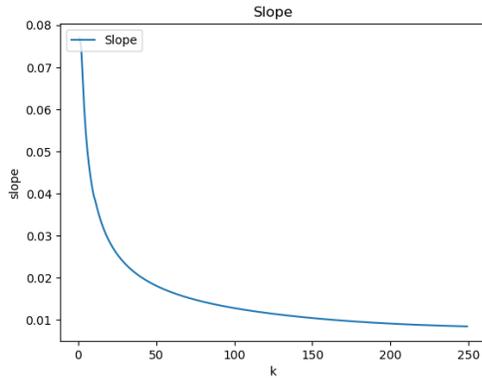
seats. However, while there does exist an element of disproportionality, in the range around $.2 \leq p \leq .8$ the minority party is not shut out of legislative representation all together and is still able to capture some seats. It is further notable that despite this disproportionality, each party cannot capture a majority of the seats without having a majority vote share. That is, either party requires at least a 50% vote share to capture at least a 50% seat share.

The symmetric clustering model introduces concepts of spatial geography, partisan clustering, and high levels of correlation, with respect to political affiliation, between a person and the people around them. In this model, immutable features of the natural world have been manipulated such that they occur equally across districts and between parties. Under symmetric clustering, there is a much wider variation in the vote margin between districts. The greatest variation exists when clustering occurs at the district level, $k = 1$, and as there begins to exist a few large clusters within each district, districts become slightly more similar to each other. Additionally, when considering the North Carolina case of $p = .47$ under district-wide clustering, there is expected to be perfect proportionality in legislative representation and even as there are multiple clusters within each district, representation is fairly proportional. When there is a larger disparity between the vote margin of each party, however, there is only proportionality under district-wide clustering, and as number of clusters within each district increases, the proportionality quickly falls. Further, as the Democrat vote margin changes from 0 to 1, seat share is weakly sensitive to vote share and there exists a level of disproportionality such that the minority party is underrepresented and the majority party is overrepresented. However, the minority party is rarely completely shut out of legislative representation and either party still needs at least 50% of the vote to win 50% of the seats and cannot get a majority of the seats without a majority of the vote. The symmetric clustering model serves as a useful manipulation of the uniformly distributed base case that introduces some simple geography, clustering, and correlation. Next, the two models are compared directly to more fully explore the effect of decreasing correlation length and unpacking clusters as symmetric clustering transitions into uniform distribution.

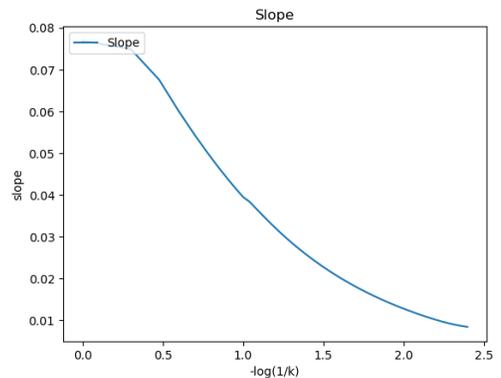
4 Comparing Models

Now that the uniform distribution and symmetric clustering models have been explored in depth, the models can be taken together and compared across key metrics to examine the transition from symmetric clustering to a uniform distribution. The cutoff between models was set at $k = 10$ so that a binomial distribution could be used for symmetric clustering and a normal for uniform distribution, but rather than assigning values of k as corresponding to one model or the other, $k \geq 1$ can be taken to explore the effects of decreasing correlation length across the entire range. Because the two models have parallel structures, it is useful and informative to perform a direct comparison of their outcomes. These analyses consider an increasing number of smaller uniform voting blocks per district, starting with each district as its own block, and changing only the underlying probability distribution used to model the expected district Democratic vote fraction. Every district in every trial and model has mean Democratic vote margin p and variance $\frac{p(1-p)}{k}$.

First, the slope of the predicted district distribution curve is considered taking $1 \leq k \leq 250$. As discussed in previous sections, the slope of the district distribution curve provides a measure of the rate at which districts approach the mean, $p = .47$, and each other. This is a proxy measure for the rate of decorrelation; as a person become less correlated with the people around them, there is less variance within districts, less variance between districts, and a smaller slope. The slope can again range from $\frac{1}{n}$, if $D_{(n)} = 1$ and $D_{(1)} = 0$, and 0, if $D_{(n)} = D_{(1)}$ and looking at the slope curve plotted on a linear scale in Figure 15(a), this is exactly what happens. When $k = 1$ and each district is its own uniform voting block and the slope is $\frac{1}{n}$, and as k increases to 250 the slope decreases towards 0. Now, to understand the rate at which decorrelation occurs, the slope curve is plotted on a linear-log scale seen in Figure 15(b). In plot (a) there is a significantly sharper decrease in slope for smaller values of k and a flattening of this rate as k increases. However, because plot (b) is not linear, the decreasing of the slope is close to but not quite a perfectly exponential relationship. Regardless, this points to an increased effect at and around the transition from



(a) Linear scale: k



(b) Linear-log scale: $-\log(\frac{1}{k})$

Figure 15: Slope of District Distribution Plot Across Correlation Length

symmetric clustering to uniform distribution, and less of an effect as correlation continues to be removed. As the number of uniform voting blocks within each district increases, and the correlation between people and the people around them decreases, there is an increasingly decreasing effect as districts grow closer to the mean and to each other.

Next, the expected number of seats won is considered across k to understand how removing partisan correlation within the electorate impacts electoral outcomes. Figure 16 shows the expected number of Republican seats won taking $1 \leq k \leq 800$ on a linear scale (a) and a linear-log scale (b). Plot (a) shows the transition from district uniformity and legislative proportionality at $k = 1$ to district proportionality and legislative uniformity at $k = 800$. This is an interesting dichotomy that highlights the difficulty of balancing competing ideas of fairness in elections. When there is high spatial correlation within the electorate of each district, the state-wide legislative representation is proportional, which seems like the fair and correct outcome. However, it seems inherently unfair and it is inefficient with respect to wasted votes to have districts that are entirely composed of one party. On the other hand, when there is little spatial correlation between the electorate, the partisan makeup of districts reflects that of the state overall but this results in complete dominance of the majority party in legislative representation. It is interesting to note that in plot (b) the slope becomes steeper around $-\log(\frac{1}{k}) = 1$, which corresponds to the transition from symmetric

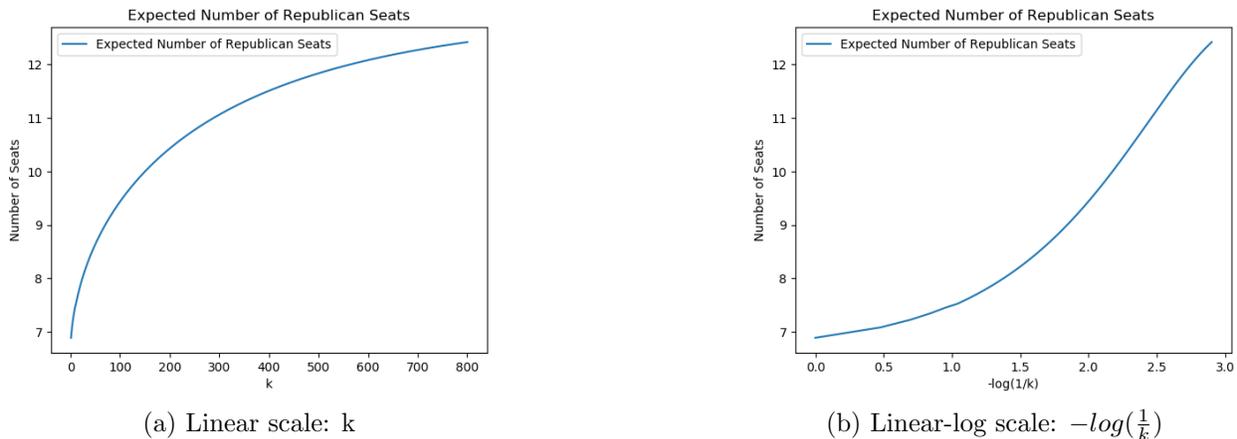


Figure 16: Average Number of Republican Seats Across Correlation Length

clustering to uniform distribution. Additionally, the plots show that a one unit increase in k at lower values results in a greater increase in expected seat share than does a one unit increase in k at high values. Considering expected election outcomes as partisan clusters are unpacked shows the decreasing effect of decreasing correlation in legislative outcomes as district uniformity transitions to legislative uniformity.

The impact of spatial correlation on legislative representation can be studied not just as a raw number of expected seats but as a measure of proportionality using the ratio of seat share to vote share. Figure 17 shows this proportionality ratio across several values of p and taking $1 \leq k \leq 100$. As in previous sections, this analysis considers values of p less than .5 and thus considers the proportionality ratio for the minority party, which is just the inverse of that of the majority party. From the plot one can see that regardless of the state-wide vote fraction, when every district is its own uniform cluster, $k = 1$, seat share is perfectly proportional to vote share and as the number of clusters per district increases, seat share for the minority party is disproportionately low. Note that with $p = .2$ the proportionality ratio decreases to 0 right at the transition from symmetric clustering to uniform distribution at $-\log(\frac{1}{k}) = 1$. For higher values of p , the proportionality ratio decreases as k increases but at a slower rate. Additionally, because the proportionality ratio of the minority party is never greater than 1, the minority party is never able to capture a majority of the seats, regardless

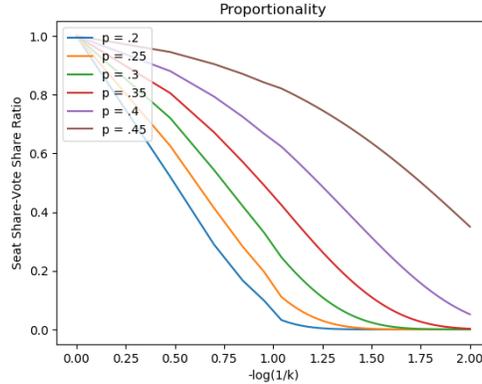


Figure 17: Proportionality Ratio Across Correlation Length

of the level of spatial correlation. The ratio of seat share to vote share reveals the extent to which the minority has increasingly disproportionately low seat share as spatial correlation is removed, and how this effect is impacted by the state-wide vote distribution.

Lastly, the vote-seat curve is applied as a measure by which to evaluate the sensitivity of seat share to vote share at varying levels of correlation. Figure 18 shows the expected number of Democrat seats taking $0 \leq p \leq 1$ and $k = 3, 5, 50, 500$ against a line of perfect proportionality. As k increases, there is a notable steepening of the vote-seat curve around $p = .5$ and a flattening of the curve at smaller and larger values of p . This means that when there is an uncorrelated electorate, the minority party is completely shut out of legislative representation unless they are very closely competitive with the majority party, in which case seat share is highly sensitive to vote share. Further, when there is a symmetrically clustered electorate, the minority party almost always has some legislative representation, even if it isn't proportional. Compared to uniform distribution, seat share is much less responsive to vote share around $p = .5$, but because there is a similar level of elasticity across p , seat share is more responsive to vote share at the extremities. Because North Carolina is a swing state in which the two parties tend to be relatively competitive, it is expected to see vote margins closer $p = .5$ where the responsiveness of legislative representation is much greater when there is an uncorrelated electorate. Additionally, it is notable that regardless of the value of k , each party requires at least 50% of the state-wide vote in order to win 50% of the seats.

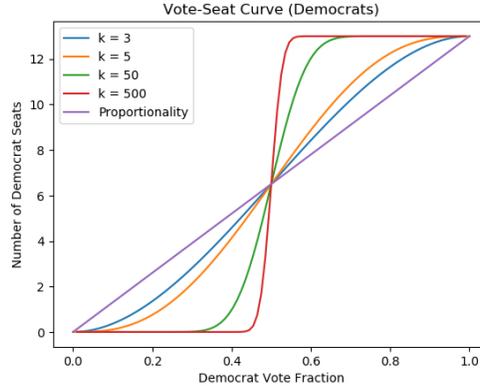


Figure 18: Vote-Seat Curve Across Correlation Length

That is, at any level of correlation, the minority party cannot capture a legislative majority. It is important here to note that all of these calculations have been done using averages, so while it may be possible to have a sampling of the distribution that results in the minority party capturing a majority of the seats, it is not the expected outcome. Similarly, the state-wide vote share is also taken as an average. Rather than taking an actual realization of the vote fraction across the state, based on a particular sampling of the determination of each of the units, it is assumed that the state-wide vote fraction is equal value of p used for the unit determination.

Comparing key metrics of election outcomes across the high correlation lengths of symmetric clustering and the low correlation lengths of uniform distribution reveals the impact of removing clustering and decreasing partisan spatial correlation. When there is a high level of correlation between a person and the people around them, uniformity within districts translates to proportionality in legislative representation but as correlation decreases, proportionality within districts translates to uniformity in legislative representation for the majority party. Further, the minority party always receives disproportionately low seat share as correlation decreases, with the effect most pronounced when the minority party has a low state-wide vote share. Notably, regardless of the state-wide vote share or the number of clusters per district, the minority party is never expected to capture a majority of the seats and in order for a party to have majority seat share on average, they need to have majority

vote share.

5 The Impact of Correlation Length

The primary goal of political parties is to effectively and efficiently translate electoral votes into legislative representation. While for district-based legislatures the conversion from votes to seats is affected by the drawing and manipulation of district boundaries, there are also immutable natural features of human geography that can and do have a notable impact on legislative outcomes. Specifically, there exists a level of correlation, with respect to political affiliation, between a person and the people around them that can introduce natural biases into election outcomes absent any gerrymandering. By modeling a simplified version North Carolina under conditions of uniform distribution and symmetric clustering, order statistics can be applied to gain insight into how the level of correlation within the electorate of each district impacts legislative outcomes and related metrics of proportionality and responsiveness to changing vote share.

It is important here to consider the limitations of these models and their conclusions in how they can be extrapolated to real-world elections. First, a notable simplification in the model is that there is no differentiation between the parameters for each district. That is, every district is constructed the same way with the same number of units and using the same state-wide vote fraction as the probability for each unit to be Democrat or Republican. While in actuality clustering does not happen uniformly across a state or district, this simplification is what allowed for order statistics to be implemented as a method from which to derive conclusions. Order statistics require samples to be drawn from the same distribution and thus each district had to be constrained by the same parameters. While this uniformity across the model state is a notable deviation from reality, it does not render the conclusions drawn from the model obsolete. The purpose of these models is to study the effects of clustering and correlation so simplifying the model by having these correlations occur constantly within

and across districts allows for such effects to be isolated. Actual elections are subjected not only to clustering but to a host of other factors, so while this model does not capture the effects of uneven clustering or the way in which clustering interacts with those other factors, it still provides useful insight into the effects of clustering.

Another limitation with the model setup is that there is no decoupling of districts from the way that people are living. Because districts are defined by a set of parameters that determine the number and size of clusters within them, rather than being drawn on top of a state-wide distribution of clusters, clusters are never split between districts. This is not an issue at large values of k when there is a completely uniform distribution but when there are fewer larger clusters per district, this model assumes geographic distribution parallels district boundaries. While for the most part district boundaries preserve cities, there still exist correlations, clusters, and communities that can and are split across districts. If each of the k units are thought of as being comprised of a certain number of smaller blocks such that at high k values there are few blocks per unit and at small k values there are many blocks per unit, districts could be drawn along these sub-unit blocks rather than along the units. While this moves away from a statistics-based model, a numeric sampling approach could explore how outcomes change now that districts are not in perfect parallel with clustering. Work on this is being done by Ella van Engen, a member of the Natural Packing group in the Quantifying Gerrymandering Bass Connections class and will provide an interesting point of comparison with the results of these models. This cluster-splitting considers a numeric sampling model of a state as a torus so the state is a continuous surface with no boundaries and districts can be drawn at any sub-unit. Early explorations suggest that there are still effects of clustering even when clusters are split across districts. That is, having partisan clusters separated across districts does not result in the same electoral outcomes as having a uniformly distributed electorate. Thus, it appears that having compact and continuous districts enhances the impact that clustering always has on legislative representation.

A final limitation with this analysis is that it only considers symmetric clustering. The

population geography of most states, however, tends to exhibit a partisan tilt in the way in which clustering occurs with Democrats having a higher level of correlation than Republicans. Initial exploration into a statistically-based model of asymmetric clustering began with a modified binomial distribution in which units had Democrat proportion v_1 with probability p_1 and Democrat proportion v_2 with probability p_2 . The system was subjected to various constraints based on a correlation analysis of North Carolina 2010 election data including the state-wide vote fraction, the probability of a Democrat precinct to be near another Democrat precinct, and the probability of a Republican precinct to be near a Republican precinct. These constraints are outlined in greater detail in the Appendix and the correlation analyses were performed by Jay Patel, another member of the Quantifying Gerrymandering Natural Packing group. Using only a binomial determination for the unit vote margins was not sufficient to effectively model the correlation data and satisfy the constraints, so a trinomial determination was considered. In this scenario there is an additional Democrat proportion v_3 that occurs with probability p_3 . To validate the usage of such a trinomial model, we looked at a histogram that plotted the percent of North Carolina precincts that had various Republican vote margins in the hopes of seeing three distinct peaks, to correspond to the three p parameters. The resulting histogram, however, did not support a trinomial determination. At this point, a statistically based asymmetric clustering model has not been reconciled but it is a point for further consideration as it is a logical extension of the current model to better capture natural partisan biases and the way in which clustering impacts election outcomes.

Another limitation and point for further consideration is moving from having uniform correlation length scales within and between districts to having varied correlation lengths. That is, the model is currently structured to have a fixed number of uniform blocks within each district. Instead, there could be several different unit sizes so each district would be comprised of some small, medium, and large blocks corresponding to high, medium, and low correlation length. The distribution of blocks within each district would have to be such

that the population of each district was the same; small blocks have less population than large blocks adding an additional constraint on the model. This extension has not been fully developed but is again a logical point from which the current model could better capture the nuances of real world clustering and subsequent election outcomes.

Overall, implementing a statistics-based model of North Carolina under uniform distribution and symmetric clustering conditions gives insight into the effect of natural packing and partisan spatial correlation on election outcomes and related metrics. Notably, when there is a high level of correlation between a person and the people around them, uniformity within districts translates to proportionality in legislative representation but as correlation decreases, proportionality within districts translates to uniformity in legislative representation for the majority party. Additionally, the minority party is consistently underrepresented in seat share and regardless of state-wide vote fraction or level of correlation is never expected to capture a majority of the seat share. It is clear from this exploration that the level of partisan spatial correlation within the electorate has a substantial impact on electoral outcomes and creates disproportional representation such that biased or unfair legislative outcomes can result from natural features of human geography.

6 Appendix

6.1 Uniform Distribution Formulas and Additional Figures

District CDF:

$$F_N(x) = \Phi\left(\frac{(x - p)}{\sqrt{\frac{p(1-p)}{k}}}\right)$$

District PDF:

$$f_N(x) = \frac{1}{\sqrt{\frac{2\pi p(1-p)}{k}}} e^{-\frac{k(x-p)^2}{2p(1-p)}}$$

CDF of the i^{th} Order Statistic:

$$F_{N, X_{(i)}}(x) = \sum_{j=i}^n \binom{n}{j} (1 - F_N(x))^{n-j} (F_N(x))^j$$

PDF of the i^{th} Order Statistic:

$$f_{N, X_{(i)}}(x) = n \binom{n-1}{i-1} (1 - F_N(x))^{n-i} (F_N(x))^{i-1} f_N(x)$$

Joint Distribution between $D_{(i)}$ and $D_{(j)}$ where $1 \leq i \leq j \leq n$:

$$f_{N, X_{(i)}, X_{(j)}}(x, y) = \frac{n!}{(i-1)!(n-j)!(j-i-1)!} (F_N(x))^{i-1} (1 - F_N(y))^{n-j} (F_N(y) - F_N(x))^{j-i-1} f_N(x) f_N(y)$$

Average Value of $D_{(i)}$:

$$\int_0^1 x f_{N, X_{(i)}}(x) dx$$

Slope Probability where m is the average value of the 1^{st} order statistic and s is the fixed slope value ($0 \leq s \leq \frac{1}{n}$):

$$\int_0^{1-ns} f_{N, X_{(1)}, X_{(n)}}(m, ns + m) dm$$

Expected Number of Republican Seats:

$$\sum_{i=1}^{n-1} (i * \int_{.5}^1 \int_0^{.5} f_{N, X(i), X(i+1)}(x, y)) + n * F_{N, X(n)}(.5)$$

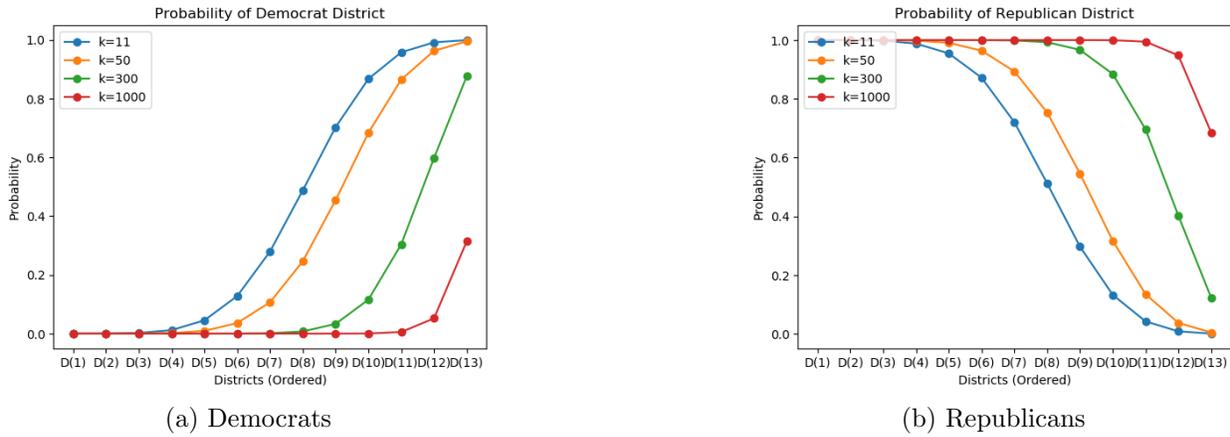


Figure 19: Probability of Each Party Winning Each District Under Uniform Distribution

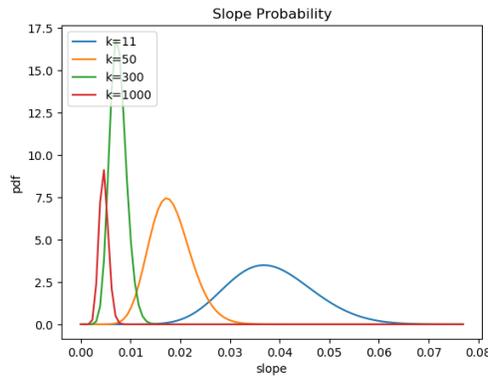


Figure 20: PDF of the Slope Under Uniform Distribution

6.2 Symmetric Clustering Formulas and Additional Figures

District CDF:

$$F_{Bin}(x) = \sum_{j=0}^x \binom{k}{j} p^j (1-p)^{k-j}$$

District PDF:

$$f_{Bin}(x) = \binom{k}{x} p^x (1-p)^{k-x}$$

CDF of the i^{th} Order Statistic:

$$F_{Bin, X_{(i)}}(x) = \sum_{j=i}^n \binom{n}{j} (F_{Bin}(x))^j (1 - F_{Bin}(x))^{n-j}$$

PDF of the i^{th} Order Statistic:

$$f_{Bin, X_{(i)}}(x) = \sum_{j=i}^n \binom{n}{j} ((F_{Bin}(x))^j (1 - F_{Bin}(x))^{n-j} - (F_{Bin}(x-1))^j (1 - F_{Bin}(x-1))^{n-j})$$

Joint CDF between $D_{(i)}$ and $D_{(j)}$:

$$F_{Bin, X_{(i)}, X_{(j)}}(x, y) = \sum_{s=j}^n \sum_{r=i}^s \frac{n!}{r!(s-r)!(n-s)!} (F_{Bin}(x))^r (F_{Bin}(y) - F_{Bin}(x))^{(s-r)} (1 - F_{Bin}(y))^{(n-s)}$$

Joint Distribution between $D_{(i)}$ and $D_{(j)}$ where $1 \leq i \leq j \leq n$:

$$f_{Bin, X_{(i)}, X_{(j)}}(x, y) = \begin{cases} F_{Bin, X_{(i)}, X_{(j)}}(x, x) - F_{Bin, X_{(i)}, X_{(j)}}(x-1, x) \\ F_{Bin, X_{(i)}, X_{(j)}}(x, y) - F_{Bin, X_{(i)}, X_{(j)}}(x-1, y) - F_{Bin, X_{(i)}, X_{(j)}}(x, y-1) + F_{Bin, X_{(i)}, X_{(j)}}(x-1, y-1) \end{cases}$$

Average Value $D_{(1)}$:

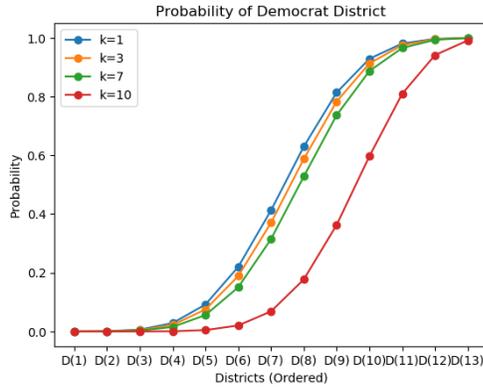
$$\frac{\sum_{x=0}^k (1 - F_{Bin}(x))^n}{k}$$

Average Value $D_{(n)}$:

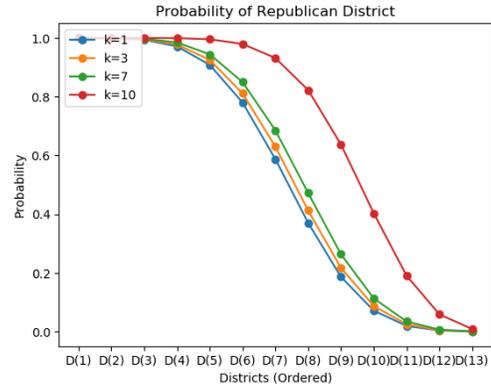
$$\frac{\sum_{x=0}^k 1 - (F_{Bin}(x))^n}{k}$$

Expected Number of Republican Seats:

$$\sum_{i=1}^{n-1} i * \sum_{y=\lceil \frac{k}{2} \rceil}^k \sum_{x=0}^{\lfloor \frac{k}{2} \rfloor} f_{B, X_{(i)}, X_{(j)}}(x, y) + n * F_{Bin, X_{(i)}}\left(\left\lfloor \frac{k}{2} \right\rfloor\right)$$

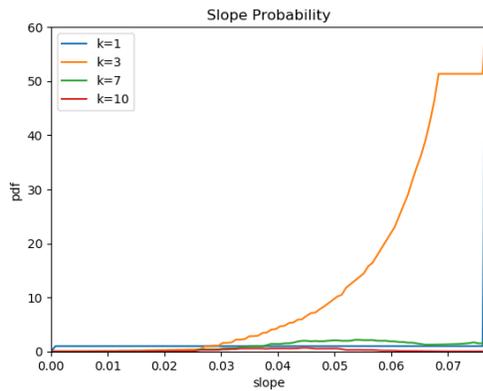


(a) Democrats

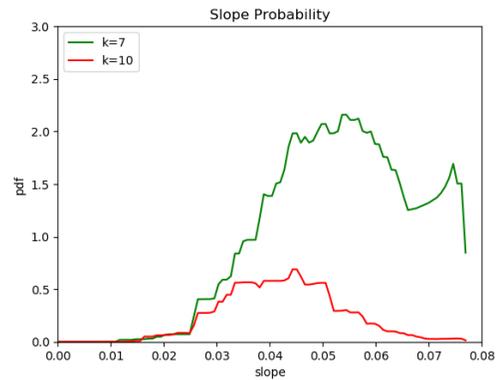


(b) Republicans

Figure 21: Probability of Each Party Winning Each District Under Symmetric Clustering



(a) $k = 1, 3, 7, 10$



(b) $k = 7, 10$

Figure 22: PDF of the Slope Under Symmetric Clustering

6.3 Asymmetric Clustering Exploration

Constraints:

- $p_1 + p_2 = 1$
- $p_1 v_1 + p_2 v_2 = v_{sw}$
- $\frac{p_1 v_1^2 + p_2 v_2^2}{p_1 v_1 + p_2 v_2} = v_D$
- $\frac{p_1(1-v_1)^2 + p_2(1-v_2)^2}{p_1(1-v_1) + p_2(1-v_2)} = v_R$

Where p_1 and p_2 are the probabilities of seeing Democrat vote fraction v_1 and v_2 respectively, v_{sw} is the state-wide Democrat vote fraction, v_D is the probability of a Democrat precinct to be next to another Democrat precinct, and v_R is the probability of a Republican precinct to be near another Republican precinct. The constraints above are for a binomial determination but can be easily extended to a trinomial by adding p_3 and v_3 .

Probability Distributions for a Trinomial:

- Probability of having J_1 v_1 units and J_2 v_2 units: $P(J_1, J_2) = \frac{p_1^{J_1} p_2^{J_2} (1-p_1-p_2)^{k-J_1-J_2}}{Z}$
Where Z is a normalizing factor: $Z = \sum_{J_1=0}^k \sum_{J_2=0}^{k-J_1} p_1^{J_1} p_2^{J_2} (1-p_1-p_2)^{k-J_1-J_2}$
- Corresponding district Democratic vote fraction: $v(J_1, J_2, k) = \frac{J_1 v_1 + J_2 v_2 + (k-J_1-J_2)v_3}{k}$
- District CDF: $P(D_i < v) = \sum_{J_1=0}^k \sum_{J_2=0}^{k-J_1} \frac{p_1^{J_1} p_2^{J_2} (1-p_1-p_2)^{k-J_1-J_2}}{Z} \mathbb{1}v(J_1, J_2, k) < v$
- District PDF: $P(D_i = v) = \sum_{J_1=0}^k \sum_{J_2=0}^{k-J_1} \frac{p_1^{J_1} p_2^{J_2} (1-p_1-p_2)^{k-J_1-J_2}}{Z} \mathbb{1}v(J_1, J_2, k) = v$

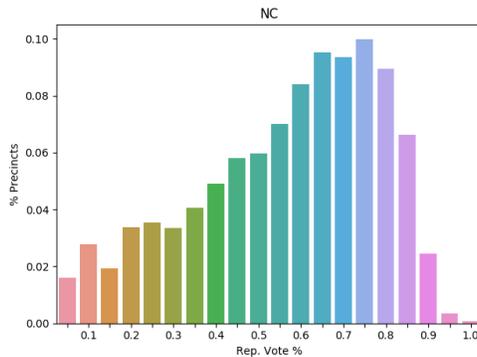


Figure 23: North Carolina Precinct Histogram Courtesy of Jay Patel

7 References

- Arnold, B. C., Balakrishnan, N., Nagaraja, H. N. (1992). *A First Course in Order Statistics*. New York: Wiley.
- Atkinson, K. E. (1993). *Elementary Numerical Analysis*. New York: John Wiley Sons.
- Chen, Jowei & Rodden, Jonathan, 2013. "Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures," *Quarterly Journal of Political Science*, now publishers, vol. 8(3), pages 239-269, June.
- Cottrel, D. (n.d.). *A Geographic Explanation for Partisan Representation: How Residential Patterns of Partisans Shape Electoral Outcomes*. Retrieved from http://www-personal.umich.edu/~dcott/pdfs/Chapter_1.pdf
- Eubank, Nicholas and Rodden, Jonathan, *Who is My Neighbor? The Spatial Efficiency of Partisanship* (August 23, 2017). Available at SSRN: <https://ssrn.com/abstract=3025082> or <http://dx.doi.org/10.2139/ssrn.3025082>
- Pitman, J. (2006). *Probability*. New York, NY: Springer.
- Siotani, M. (1956). Order Statistics for Discrete Case with a Numerical Application to the Binomial Distribution. *Annals of the Institute of Statistical Mathematics*, 8(2), 95–104. doi:10.1007/bf02863574