District Bias in North Carolina:
Introducing a Novel Approach to Identifying Gerrymandered Districts Using Geospatial Data

Nikhil Arayath (na214), Daniel Ehrlich (dje26), Caleb Kiprono (ck297), Pierre Quereuil (pq10)

Part 1: Introduction and Research Question

Gerrymandering is the practice of redrawing electoral district boundaries to give one political party an unfair advantage over its opponents. By manipulating district lines, politicians can create "safe" districts where their party is almost guaranteed to win, distorting the principle of fair representation. This practice has sparked widespread controversy, as it undermines democratic accountability and often results in election outcomes that do not reflect the true will of the voters. In North Carolina, gerrymandering has been particularly contentious, with the state's legislative maps challenged in multiple lawsuits. Recently, in Bard v. N.C. State Bd. Of Elections (2024), a series of plaintiffs sued the NC State Board of Elections for the 2023 congressional and state legislative district lines. Although ultimately dismissed, this case serves as an example of the urgent need for solutions to ensure electoral fairness and equal representation.

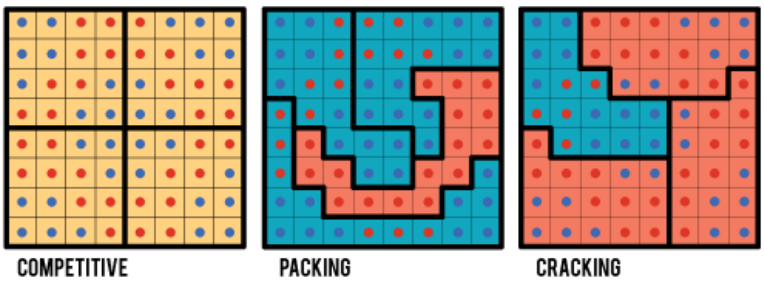
In this project, we aim to answer the following research question:

1. Can we identify gerrymandered districts in North Carolina using existing metrics (compactness, mean-median difference, and efficiency gap) and a new custom score based on voting habits between current and "ideal" districts?
2. Does smoothing districts into circles produce districts that better represent statewide age proportions and demographic distributions? How does our custom metric compare to other commonly used measures of gerrymandering?

By considering multiple modes of evaluating gerrymandered districts (including geographic shape, compactness measures, and voter turnout), our project analyzes the extent of gerrymandering in North Carolina. This is an ongoing issue highlighted by events like the recent national debate on redistricting and electoral reform. Additionally, just in the past election, North Carolinians voted to elect their House, Senate, Attorney General, Lieutenant Governor, Governor, and the US President. Our work could contribute valuable insights to help inform policy decisions aimed at ensuring more equitable representation in the future. We also believe this was worthwhile to our time in the course since we applied our CS216 knowledge to the influential real-world situation that gerrymandering is.

There exist many ways to gerrymander a district, including altering compactness, playing with demographic distributions, and coordinated multi-district manipulations. No one other than those directly involved in the redistricting process will ever fully understand the motivations and justifications behind why districts look the way they do. Given the multitude of levers that can be pulled, we decided to focus on two of the most common gerrymandering tactics when developing our customized score: packing and cracking. To better understand how these processes work, see Figure 1. Imagine you wanted to ensure a statewide republican advantage. To achieve this goal, you could either draw lines to pack Democrat votes into one single district - allowing a Republican majority in other ones - or spread Democrat votes into multiple districts - allowing Republican advantages in all of them.

Figure 1: Packing and Cracking. Obtained from *Exploring Pennsylvania's Gerrymandered Congressional Districts, by Daniel McGlone*



Part 2: Data Sources

Shapefiles of Census Blocks / Districts: After going through many different shapefile publications, we settled on https://www.census.gov (The United States Census Bureau) for census blocks and https://www.ncleg.gov (North Carolina General Assembly) for district boundaries. These are both direct data sources and in our view the most trustworthy (coming from official governmental organizations).

Voter Data per Census Block: To accumulate voting data (registered voter party, demographic, and voting information), we used https://www.l2-data.com (L2 data), the most used and trusted source for aggregated voter records. While second-hand, the source enables easy processing and is widely trusted.

These sources are either first-hand or trustable government sources, and contain the relevant data in identifying gerrymandered districts. As we narrowed down our research goals, we were able to reduce the number of databases necessary for the above three sources. Using the Geopandas library, we projected these shapefiles onto a uniform coordinate scaling system. GEOIDs were then used to merge our voter data with these geolocation information tables, allowing us to have extremely fine-grained voter data (there are 14 districts containing 236,638 census blocks).

Part 3: What Modules Are We Using?

Module 4: Data Wrangling

In this project, we applied the concepts of data wrangling from Module 4 to clean, organize, and prepare our data for analysis. The justification for this step is that raw data often comes from different sources with varying formats and standards, making it difficult to combine and analyze. In our case, we used Geopandas, an extension of pandas designed for geospatial data, to handle the district boundary data. Geopandas provided the necessary tools to process spatial information and ensure compatibility across datasets, which was essential for analyzing district shapes and areas. One major challenge was inconsistent geolocation standards across datasets, which required extensive cleaning to ensure compatibility. The cleaned dataset served as a foundation for our subsequent analyses and visualizations.

Module 6: Combining Data

After data wrangling, merging was essential to combine voter data with district shapefiles. We needed to link electoral data to district boundaries for accurate analysis. Traditional merging based on column equivalency was not suitable, so we used Geopandas to perform a spatial intersection, ensuring that census blocks were correctly matched to their respective districts. This spatial merge was necessary because district and voter data were tied by geography rather than shared identifiers.

Module 3: Visualization

We used the concepts from this module to create several visualizations as part of our data analysis to explore potential gerrymandering in North Carolina's districts. For our customized gerrymandering metric, we transformed each district into a circular one corresponding to the minimum bounding circle of the original district. It enabled us to compare each district's geometry to an ideal compact shape and identify deviations that might indicate gerrymandering. We then mapped voter data from census blocks within each circle, visualizing these circular areas and their corresponding census blocks in Figure 2. Our goal was to create an intuitive, standardized shape that could reveal irregularities in district design. Additionally, we created two visualizations comparing the % of voters leaning Democrat in each district and the circle corresponding to that district (figures 3 and 4), as well as two more comparing the difference in demographic distributions (e.g. age) between districts and the whole.

Module 7: Statistical Inference

After visualizing positive results with the previous module, we also conducted statistical analysis through t-testing of our district-wide metrics, comparing Polsby-Popper, efficiency gap, and our custom score. These analyses enabled us to compare the performance of our custom measure of gerrymandering to

methods validated by existing literature. These comparisons informed several of our final conclusions and elucidated areas for further improvement/development (see limitations and future directions).

Part 4: Methods and Results

Accessing Our Implementation: All code, data, and visualizations are accessible on GitLab through the following link: https://gitlab.oit.duke.edu/pq10/cs-216-gerrymandering.

Methodology: To assess gerrymandering in North Carolina, we began by collecting and preparing data from three main sources: the U.S. Census Bureau for census block shapefiles, the North Carolina General Assembly for official congressional district boundaries, and L2 for voter registration and voting data by census block. We used Geopandas to project all shapefiles onto a unified coordinate system, ensuring accurate spatial comparisons across datasets. Using the unique GEOID identifiers, we joined voter data with spatial data, creating a highly detailed combined dataset that covered all 14 districts and 236,638 census blocks. Data cleaning was critical for ensuring accuracy and consistency across the merged datasets. First, we standardized spatial data by aligning the coordinate reference systems of all shapefiles to a common EPSG code, enabling precise spatial overlays. Using GEOID as a linking key, we joined voter data with geospatial information, stripping any extraneous characters and converting data types as necessary to maintain consistent formats across all datasets.

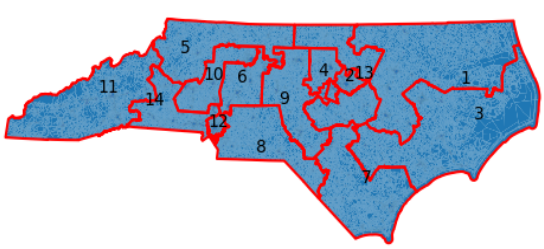
For our gerrymandering analysis, we selected four metrics: Polsby-Popper compactness, efficiency gap, mean-median difference, and a customized voter distribution score. For our customized score, we replaced district lines with minimum bounding circles centered at each district's geographic centroid. Using spatial joins, we aggregated votes from census blocks within each circle to obtain a voting profile for each circular district. Visualizations of these circular districts and their voting patterns revealed shifts in party advantage, suggesting potential gerrymandering (Figures 4-5).

Using voter registration data from our original and circular districts, we assigned a gerrymandering score to each district. This score measured the extent to which Democrats' and Republicans' vote share changed between the circular and original district and was calculated using the following equation:

$$\text{score} = [(\%\ \text{Dem votes})_{\text{CIRCLE}} - (\%\ \text{Rep votes})_{\text{CIRCLE}}] - [(\%\ \text{Dem votes})_{\text{ORIGINAL}} - (\%\ \text{Rep votes})_{\text{ORIGINAL}}]$$

We then performed a series of Welch's t-tests to compare the distribution of gerrymandering scores across North Carolina's districts obtained using our custom score, Polsby-Popper, and the efficiency gap metrics (Table 1). We excluded the mean-median difference from our analysis because this metric only provided us with one value, representative of partisan bias across the state of North Carolina, instead of a series of district-specific gerrymandering scores.

Figure 2 (left): Visualization of Districts (Outlined in Red) and Census Blocks within Each District
Figure 3 (right): Adjusted Circular Districts (Outlined in Red) and the Corresponding Original Districts (Outlined in White)



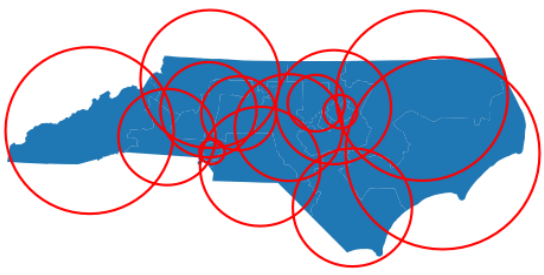


Figure 2 shows the 2020 congressional district lines in North Carolina. A brief visual analysis highlights the prevalence of abnormally shaped districts, which appear to deliberately avoid certain areas on the map. These geometric irregularities could be indicative of "packing and cracking". Figure 3 depicts our adjusted circular districts. These circles' areas are roughly proportional to the areas of the underlying districts in an attempt to conserve population distributions and develop "ideal" districts that minimize partisan bias from potential gerrymandering.

Figure 4 (left): Percentage of Democratic Voters per District
Figure 5 (right): Percentage of Democratic Voters per Adjusted Circular District

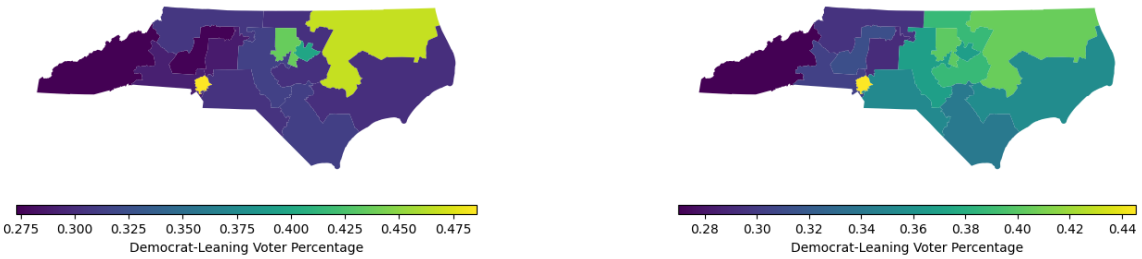
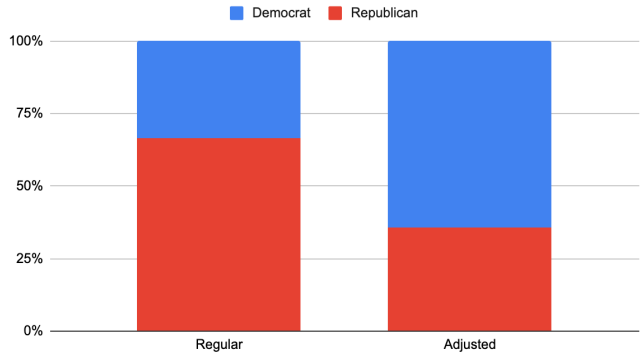


Figure 6: Partisan Change between Regular and Adjusted Districts



Using our new metric, we identified several potentially gerrymandered districts. Comparing the percentage of democratic voters per original vs circular districts (as visualized in Figures 4 and 5), one observes a notable increase in democratic voter share across almost every district. This is particularly prevalent in eastern rural areas of North Carolina and in the suburbs of Charlotte. The stacked bar charts in Figure 6 directly compare these distributions of democratic and republican-majority districts using the regular and adjusted (circular) borders. Using original district lines, Democrats had a majority in 4 districts and Republicans had a majority in 10. After smoothing districts into circles, Democrats had a majority in 9 districts and Republicans had a majority in 5, resulting in Democrats gaining a statewide majority.

Figure 7: Variations in District Age Group Proportions Relative to Statewide Distribution

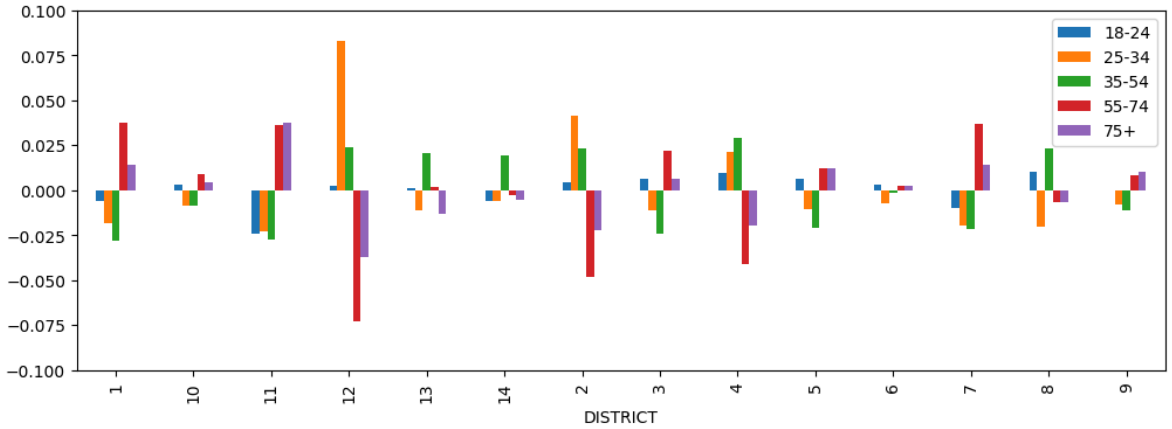
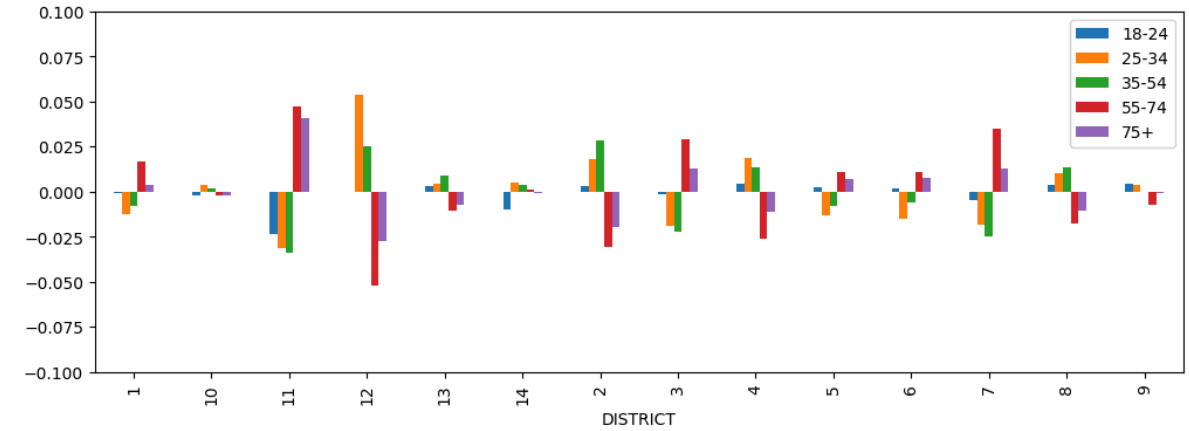


Figure 8: Variations in Circular District Age Group Proportions Relative to Statewide Distribution



Given that age often correlates with voting patterns, we then assessed the distribution of age groups across each of the original and adjusted districts. To do this, we stratified our data on voter registration into 5 age groups (18-24, 25-34, 35-54, 55-74, and 75+). For each district, we calculated the proportion of voters belonging to each age group and compared these numbers to the statewide distribution. We then visualized our results using bar graphs (see Figure 7). Taller bars corresponded to greater deviations between district- and state-wide age distributions. We subsequently repeated this analysis using data from our adjusted circular districts (see Figure 8). When comparing our results, we noted that every adjusted circular district (with the exception of districts 3 and 6) had age distributions that better correlated to the statewide average.

We repeated these analyses to evaluate differences in racial/ethnic composition across districts (see Appendix A, Supplementary Figures S1 and S2). For this study, we focused primarily on the distribution of individuals identifying as Hispanic, African American, or of White European descent. Similar to our evaluation of age groups, we noticed that our adjusted circular districts better represented the statewide demographic composition, however the difference between adjusted and original district distributions was not as dramatic as that observed in Figures 7 and 8.

Table 1: Table of T-Statistics (with P-Values in Parentheses) Obtained by Comparing our Custom Assessment of Gerrymandering to the Polsby-Popper and Efficiency Gap Metrics

	Custom Score	**Polsby-Popper**	**Efficiency Gap**
Custom Score	—	-2.52 (0.02)	-8.53 (2.50×10^{-8})
Polsby-Popper	2.52 (0.02)	—	0.08 (0.93)
Efficiency Gap	8.53 (2.50×10^{-8})	-0.08 (0.93)	—

We then wanted to see how our custom metric compared to other commonly used measures of gerrymandering, particularly, the Polsby-Popper score and Efficiency Gap metric. For this comparison, we performed a series of Welch's t-tests. We did not do a normal t-test because we could not assume that our metric's scores had a similar variance to those obtained using Polsby-Popper and Efficiency Gap methods. For these comparisons, we made two key assumptions. First, we assumed independence between the observations made by each gerrymandering metric. Second, we assumed that all three metrics generated a normal distribution of scores across North Carolina's 14 districts.

The results of these t-tests suggest that there is no statistically significant difference in the distribution of scores obtained using Polsby-Popper and Efficiency Gap metrics. However, the mean scores generated from our custom approach differ significantly from both of the aforementioned metrics. We (optimistically) believe that this difference stems from the fact that our custom score incorporates novel features into its assessment of gerrymandering. By integrating spatial analysis and simulating voting outcomes to reveal deviations from compactness and boundary manipulation, our custom score provides a unique perspective into how district lines may have been drawn to shift party advantages. We are also aware that there exist several limitations to our custom score that may have contributed to its results deviating significantly from those obtained using Polsby-Popper and Efficiency Gap metrics.

In addition to comparing district-specific gerrymandering scores, we wanted to see how North Carolina fared on a state-wide level. To do so, we measured the difference between median and mean Democrat and Republican vote share across all 14 districts. These calculations yielded a mean-median difference of -3.4% for Democrats and + 3.6% for Republicans. A positive the mean-median difference was indicative of partisan bias (potentially due to gerrymandering) in favor of a given party, while a negative mean-median difference was indicative of partisan bias against a given party.

Part 5: Limitations and Future Work

Limitations: Our implementation of circular adjusted districts to assess gerrymandering comes with several limitations. First, our decision to draw circles according to the minimum bounding area for each district fails to take into account demographic factors. Even though our analyses in Figures 7 and 8 and in Appendix A demonstrate that our circular districts better represent statewide age, racial, and ethnic distributions, this could be further optimized through a less arbitrary method of selecting circle radii.

Our adjusted districts are currently centered based on the geographic centroid of each district. This measure of centrality fails to consider critical data on population density distributions within districts. Another core limitation was that our circles exhibit significant spatial overlap (as can be seen in Figure 3). This overlap makes our model prone to double counting, which introduces noise into our analyses and is especially problematic for highly concentrated areas (i.e., Charlotte). Finally, our current model relies exclusively on voter registration data from 2020. The fact that our data comes from only one year prevents us from making any intertemporal generalizations and the fact that we look at voter registration

data (as opposed to data on votes cast in actual elections) means that our results cannot be validated against any real world outcomes.

Ethical Implications and Potential Biases: The double counting that accompanies spatial overlap of circular districts disproportionately biases major population centers, which tend to lean democrat. This may, in part, explain why so many districts flipped in our analysis (Figure 6). Our datasets also contain implicit biases. Although the US Census was the most accurate source of population data we could find, factors such as non-reponse, incorrect reporting, and belonging to a "hard-to-count" group (homeless, etc.) may provoke bias towards demographics.

Future Directions: In response to the limitations listed above, there are several modifications that can be made to our current approach for measuring gerrymandering. First, instead of using minimum bounding circles for our adjusted districts, we could adopt a data-driven approach to determine circle radii. To do so, we could cluster districts based on features such as population density and demographic distribution. We can generate a series of objective functions that incorporate one or two of the most prominent factors for each cluster and apply cross validation to generate radii of circular districts that minimize artificial partisan bias. Additionally, we can center each adjusted district using the underlying district's population-weighted centroid (instead of using the geographic centroid), incorporate results from multiple election years, and look at actual votes cast instead of voter registration data. To make our adjusted districts more representative, we could also build a Monte Carlo simulation focused on random district generation. With this approach, instead of having one circular adjustment, we could generate various district plans and compare how election results evolve across these plans. This more generalized approach can provide important insights and suggest approaches to redistricting that improve the fairness and representation of statewide elections in North Carolina.

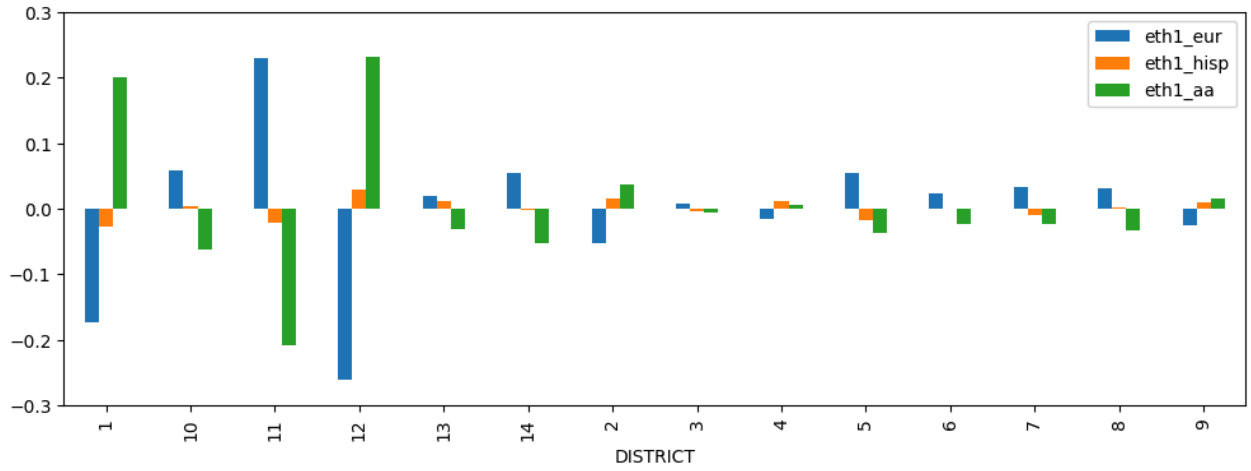
Lastly, we could include additional analyses to measure alternative methods of gerrymandering (beyond just packing and cracking). To do this, we propose representing each district as a node in a weighted graph (edge weights corresponding to geographic proximity and similarities in electoral behavior/partisan bias). A strongly weighted edge would suggest that subtle boundary changes between these two (potentially neighboring) districts may lead to significant consequences. We can then look at metrics such as cut size and betweenness centrality to identify key districts where boundary changes may disproportionately influence statewide electoral outcomes.

Part 6: Conclusion

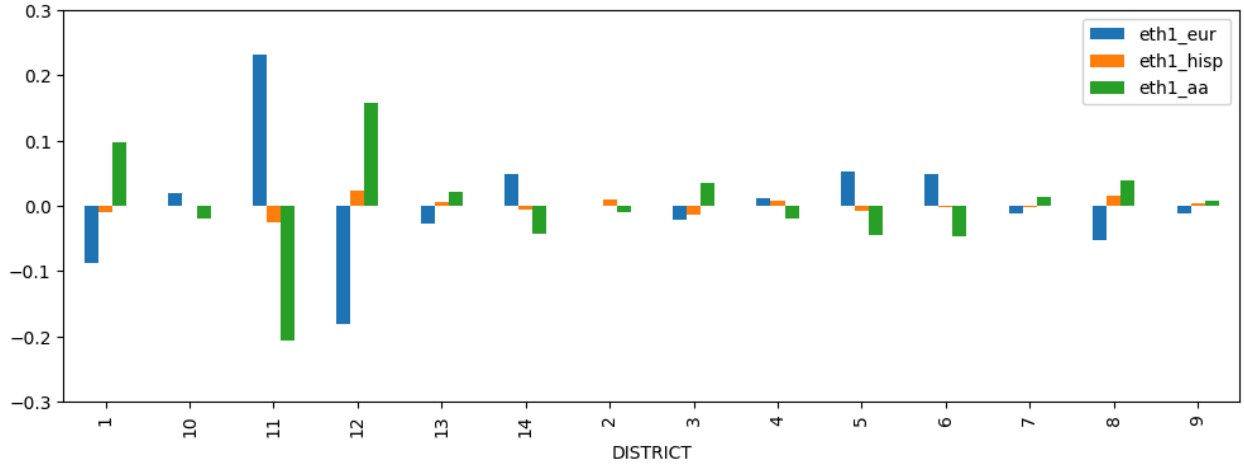
Gerrymandering currently stands as one of the greatest threats to the democratic integrity of American elections. Using geospatial and voter registration data, our project aims to tackle this challenge by proposing a novel approach to measure gerrymandering across congressional districts. By smoothing each district into a perfect circle, we developed a series of artificial districts, with zero compactness bias. Demographic analyses of age and ethnic/racial distributions demonstrated that our adjusted districts overall were more representative of statewide averages than the original districts. Given these findings, we concluded that our artificial districts could be used as a baseline against which we could assess the fairness and partisan bias of current district lines. When comparing voter registration data (% vote share for Democrats vs Republicans) between the original districts and their corresponding circles, we found that 5 districts flipped from having a Republican majority to a Democratic majority. As a result of this change, Democrats won a statewide majority (in our idealized district model), while Republicans had a statewide majority based on current district lines. This finding aligned with results from the Princeton Gerrymandering Project which suggested the prevalence of gerrymandering in North Carolina in favor of Republicans.

Appendix A

Supplementary Figure S1: Variations in Racial and Ethnic Distributions Across Districts Relative to the Statewide Demographic Composition



Supplementary Figure S2: Variations in Racial and Ethnic Distributions Across Adjusted Circular Districts Relative to the Statewide Demographic Composition



Works Cited

Bureau, US Census. “Mapping Files.” *Census.gov*, 24 Oct. 2024, www.census.gov/geographies/mapping-files.2020.List_27529751.html#list-tab-List_27529751. Accessed 7 Dec. 2024.

“Princeton Gerrymandering Project” *gerrymander.princeton.edu*, gerrymander.princeton.edu/.

Kirschenbaum, Julia, and Michael Li. “Gerrymandering Explained.” *www.brennancenter.org*, Brennan Center for Justice, 10 Aug. 2021, www.brennancenter.org/our-work/research-reports/gerrymandering-explained.

“L2 Political.” *L2 Political*, 2020, www.l2-data.com/.

Michael Li, et al. “Anatomy of a North Carolina Gerrymander | Brennan Center for Justice.” *Www.brennancenter.org*, 27 Oct. 2023, www.brennancenter.org/our-work/analysis-opinion/anatomy-north-carolina-gerrymander.

“North Carolina General Assembly.” *Www.ncleg.gov*, www.ncleg.gov/.

Prison Gerrymandering: How One Count Leads to a Decade of Distortion.

“Redistricting - North Carolina General Assembly.” *Www.ncleg.gov*, www.ncleg.gov/Redistricting.

Writer, Juan Siliezar Harvard Staff. “An Algorithm to Detect Gerrymandering.” *Harvard Gazette*, 3 Nov. 2022, news.harvard.edu/gazette/story/2022/11/an-algorithm-to-detect-gerrymandering/.