Capabilities to Measure Communities: Employing the Capability Approach to Redistrict North Carolina

December, 2025

Abstract

When redistricting, North Carolina requires that Communities of Interest (COI) are considered. COI are broadly defined as groups of individuals with similar socioeconomic characteristics and policy concerns and thus would benefit from being grouped together for representation. To assess if recent district maps preserve or divide COI, we developed a mathematical model using Principal Component Analysis and the capabilities approach to identify and define COI, and evaluate how well recent North Carolina district maps preserve these COIs. We constructed a composite *Grand COI Score* for each county, capturing multidimensional well-being across nine key determinants. Using Spatially Constrained Hierarchical Clustering, we grouped counties into contiguous COI regions and analyzed how the 2020, 2023, and 2025 congressional maps aligned with these communities. Our results show that the 2020 map generally maintained COI integrity, while the 2023 and 2025 maps increasingly split and aggregated COI, reducing representation consistency. Spatial regression models confirm that county-level characteristics are significantly associated with COI scores, and point process analysis highlights clustering patterns of higher education institutions that correspond with COI distributions. These findings suggest that recent redistricting efforts may compromise fair representation of socioeconomically cohesive communities and demonstrate the utility of quantitative methods for evaluating COI preservation in electoral map design.

1 Introduction and Background Context

In this section, we provide an overview of the current congressional divide in North Carolina. We then look at communities of interest and gerrymandering. Finally, we propose Amartya Sen's Capability Approach as a framework for identifying communities of interest.

1.1 Current Politics and Policies

In the 2024 general election, 51.0% of the North Carolina population voted Republican and 47.8% of the population voted Democrat [1]. In a fairly drawn map, we would expect to see North Carolina's congressional district map reflect a similar percentage. However, the 2024 congressional district map has 10 Republican of 14 total seats (71.4%), leaving Democrats 4 out of 14 seats (29.6%).

On October 22, 2025, the North Carolina House of Representatives approved a remapping of the state's congressional district map aimed at securing the 11th Republican seat in the state. The map is planned to be enacted for the 2026 midterm elections.

The North Carolina State Constitution [2] outlines certain requirements for congressional districts. In short, they must:

1. be contiguous
2. have equal populations
3. be geographically compact
4. preserve county boundaries as much as possible

1.2 Gerrymandering

Gerrymandering is defined as "the practice of dividing or arranging a territorial unit into election districts in a way that gives one political party an unfair advantage in elections." [3] Two common techniques for gerrymandering are packing and cracking. Packing aims to cram as many members of minority voters into the smallest number of districts in order to dilute their influence in surrounding districts. Cracking aims to disperse members of minority voters to prevent them from being to influence any one district and elect representatives that advocate for them. [4]

Compactness is a common metric for gerrymandering. The most widely used measure for compactness is the Polsby-Popper score which assumes the ideal district shape is a circle. The Polsby-Popper score is given by the following equation, where a perfect circle earns a score of 1, and a highly contorted region earns a score close to 0 [5]:

$$\frac{4\pi \times \text{Area}}{(\text{Perimeter})^2}$$

The Reock score is another measure of compactness. The Reock score measures dispersion around a central point. Like the Polsby-Popper score, the ideal district shape is assumed to be a circle. A perfectly compact region (a circle) earns a score of 1, and a highly dispersed region earns a score close to 0. [6] The equation is as follows:

$$\text{Reock Score} = \frac{Area_{\text{District}}}{Area_{\text{Minimal Circumscribing Circle}}}$$

However, compactness as a measure has drawbacks. Firstly, it is possible for a compact region to be gerrymandered, and for a non-compact region to be fair. Secondly, region geography often renders nice shapes impossible (eg. along coastlines).

1.3 Communities of Interest

Communities of Interest (COI) are broadly defined as groups of individuals with similar socioeconomic characteristics and policy concerns and thus would benefit from being grouped together for representation. North Carolina court provisions state that "communities of interest should be considered in the formation of compact and contiguous electoral districts". This provision does not provide a definition or criteria for what a community of interest is, and thus many different methods have been used in the past to determine these communities.

Generally, there are three categories of criteria that are used to determine communities of interest [7]

1. common interests or characteristics
2. patterns of interaction
3. administrative or geographic boundaries

The provision is also unclear how crucial communities of interests should be in redistricting. Thus, we aim to find a method that considers personal characteristics, social interaction patterns, and geographical boundaries while also remaining flexible.

1.4 The Capability Approach

Amartya Sen proposed the capability approach as an alternative to welfare economics [8]. The capabilities approach not only evaluates whether or not people have the means to achieve their goals, but also whether or not they have the freedom to do so.

The capability approach introduces the ideas of capability, functionings, and conversion factors. Capabilities are what people are able to be and do, and functionings are beings and doings. In other words, capabilities are a person's opportunities to achieve functionings [9]. Conversion factors connect resources and functionings. Conversion factors are defined as a person's ability to convert resources into functionings. They can be broken down into three categories: personal, social, and environmental. These three factors determine a person's capability set – capabilities that a person is simultaneously able to have [9]. Along with these conversion factors, capability, and functionings, the capability approach use resources as inputs, and well-being as the ultimate output. The capability approach is a general framework, and thus can be applied to create specific capability theories.

Traditionally, the capability approach views well-being from the individual perspective. However, there has been exploration into applying the capabilities approach in group settings. Specifically, there is some evidence to suggest that those with lower socio-economic status forming groups can lead to an expansion in their capabilities [10].

We propose the idea of using the capability approach to identify communities of interest and draw district boundaries. The capability approach allows for the consideration of socio-economic and environmental factors that communities aim to fundamentally group by. By grouping those with similar capabilities into and considering their average capabilities to be a representation of the group, we can understand what those communities value and show how effectively they are able to use resources allocated to them, and thus are an effective way to group persons for representation purposes.

2 Data

We consolidated North Carolina data from multiple governmental organizations to round out the list of conversion factors considered. Due to data constraints, we keep the most granular grouping of people to the county level. However, as North Carolina districts are required to avoid crossing county lines as much as possible, the granularity should not significantly impact the applicability of our results. All the data we used and the code used for analysis can be found in the appendix.

2.1 North Carolina Association of County Commissioners

The North Carolina Association of County Commissioners (NCACC), is an advocacy group that works with registered lobbyists to inform North Carolina politicians about county-level issues [11]. The NCACC releases an annual "County Map Book", which contains county-level data regarding factors such as age, education, health, economic and environmental factors, and taxation policies. Through web scraping in Python, we consolidated the presented data into a dataset, organized by county. We chose to gather data from the 2024 County Map Book to account for other databases not having more recent data yet.

2.2 North Carolina Department of Health and Human Services

The North Carolina Department of Health and Human Services (NCHHS) is a state department that focuses on the health- and human-related services across North Carolina, where they also host a robust database on county-level demographic composition. From this database, we gathered the 2022 estimated demographic compositions of each North Carolina county by race [12]. We selected data from 2022 as this was the most recent data that we could find. These races are categorized in accordance to the 2000 US Census Bureau classifications of race.

2.3 Federal Emergency Management Agency

The Federal Emergency Management Agency (FEMA) is a federal government agency concerned with providing aid towards the prevention, support, and recovery of communities affected by/potentially affected by natural disasters. In 2022, FEMA produced a comprehensive and thorough dataset of the environmental "risk factors" of each county across the US, where we collected data on all 100 counties of North Carolina.

As defined by FEMA, the risk factor of a given natural disaster is calculated as follows:

$$\text{Risk} = \text{Expected Annual Loss} \times \text{Community Risk Factor}$$

where "Expected Annual Loss" is calculated to be the projected cost of damages to a given region in US dollars, and "Community Risk Factor" refers to a probabilistic function outlined below:

$$\text{Community Risk Factor} = f(\frac{\text{Social Vulnerability}}{\text{Community Resilience}}), \quad f(\cdot) = \tau(a = 0.5, b = 2, c = 1)$$

where $\tau(\cdot)$ is a triangular distribution with minimum = 0.5, maximum = 2, and mode = 1.

In particular, social vulnerability refers to a background quantitative variable on the local community and demographic information, and their susceptibility to environmental risks. Community Resilience is defined as "the ability of a community to prepare for anticipated natural hazards, adapt to changing conditions, and withstand and recover rapidly from disruptions" by FEMA. This is quantified by the University of South Carolina's Hazards and Vulnerability Research Institute (HVRI) Baseline Resilience Indicators for Communities (HVRI BRIC), and is recommended by FEMA's Social Vulnerability and Community Resilience Working Group.

For each county of North Carolina, we collected the risk factor scores of coastal flooding, riverine flooding, hurricanes, and wildfires, as well as the community resilience metric.

2.4 The North Carolina Office of State Budget and Management

The North Carolina Office of State Budget and Management (OSBM) is a government-officiated resource management group, who works closely with other North Carolina departments. Data regarding transportation was found on the OSBM Log In to North Carolina (LINC) database, and was provided

by the North Carolina Department of Transportation (NCDOT). From this, we collected the number of registered automobiles and the mileage of both "primary" and "secondary" highways, defined by the NCDOT as providing the highest level of service at the greatest speed for the longest uninterrupted distance, with some degree of access control and a less highly developed level of service at a lower speed for shorter distances by collecting traffic from local roads and connecting them with respectively.

3 Statistical Methods

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a fundamental technique for *dimensionality reduction* used to transform a large set of correlated variables into a smaller set of linearly uncorrelated variables called **Principal Components (PCs)**. The goal is to retain the maximum variability (information) from the original data in the fewest possible components.

Given a data matrix $\mathbf{X}$ with n observations and p variables, PCA finds a set of weights (or loadings) $\phi_j = (\phi_{1j}, \ldots, \phi_{pj})^T$ such that the j-th principal component, Z_j, is a linear combination of the variables:

$$Z_j = \phi_{1j}X_1 + \phi_{2j}X_2 + \cdots + \phi_{pj}X_p = \mathbf{X}\phi_j$$

The first principal component (Z_1) maximizes the variance of the projected data by solving the following optimization problem:

$$\max_{\phi_1} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} X_{ij} \right)^2 \right) \quad \text{Subject to:} \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

After Z_1 has been determined, the second principal component (Z_2) is found as the linear combination of $X_1, \ldots, X_p$ that has maximal variance out of all linear combinations that are uncorrelated with Z_1.

The components Z_j are derived sequentially such that:

- Z_1 captures the maximum variance in $\mathbf{X}$.
- Z_j captures the maximum remaining variance, subject to being uncorrelated to all preceding components $Z_k, k < j$. For Z_2, this means the loading vectors are orthogonal: $\phi_2^T \phi_1 = 0$.

The importance of each component is measured by its Proportion of Variance Explained (PVE):

$$\text{PVE}_m = \frac{\text{Variance of } Z_m}{\text{Total Variance in Data}} = \frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2}$$

In our analysis, PCA is used to summarize and simplify the complex relationships among the **p** potential predictor variables (e.g., socioeconomic, demographic, or environmental factors) before their use in downstream spatial regression models, ensuring the selected components represent independent sources of variation. Since no formal response variable is available in this stage, PCA is essential for reducing feature redundancy and enhancing the robustness and interpretability of subsequent unsupervised methods.

3.2 Hierarchical Clustering

Hierarchical Clustering is an *unsupervised learning technique* used to group similar observations (or spatial units) into clusters. Unlike partitioning methods (like K-means), hierarchical methods build a

cluster tree, or dendrogram, which shows the relationships between clusters across different levels of similarity.

We employ an *agglomerative* (bottom-up) approach, starting with each observation as its own cluster and successively merging the closest pairs of clusters until only one cluster remains. The clustering process relies on two components: a distance metric and a linkage method. We utilize the **h** linkage method (e.g., Ward's method or complete linkage) and a **d** distance metric (e.g., Euclidean distance):

$$\text{Distance}(\mathcal{C}_i, \mathcal{C}_j) = f(\text{Distance}(\mathbf{x}_a, \mathbf{x}_b) \mid \mathbf{x}_a \in \mathcal{C}_i, \mathbf{x}_b \in \mathcal{C}_j)$$

Spatially Constrained Clustering: To ensure the resulting clusters form geographically contiguous regions, we implemented a spatially constrained hierarchical clustering approach. This method restricts the linkage process such that a pair of observations or clusters $(\mathcal{C}_i, \mathcal{C}_j)$ can only be merged if they are geographically adjacent (i.e., sharing a boundary or defined as neighbors in a spatial weights matrix). This constraint ensures that the identified clusters represent meaningful, continuous geographical regions rather than disparate units scattered across the study area.

The resulting clusters are used to identify homogeneous regions within the study area based on the PCA-derived component scores. These identified regions can then be incorporated into the spatial regression models as fixed effects.

3.3 Spatial Regression Models

When analyzing data indexed by geographic location (e.g., county-level data), the assumption of independent errors, $E(e_i) = 0$, in standard linear regression (Equation 1) is often violated. Spatial dependence where observations closer in space are more similar must be explicitly modeled. We consider two common spatial lag models that account for this non-independence using a predefined **W spatial weights matrix** (e.g., k-nearest neighbors).

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{i,k} + \epsilon_i \tag{1}$$

Spatial Autoregressive (SAR) Model

The Spatial Autoregressive (SAR) model (or Spatial Lag Model) accounts for dependence in the response variable (**Y**). It posits that the outcome at location i is directly influenced by the outcomes in neighboring locations, as defined by the spatial weights matrix **W**.

The model is defined as:

$$\mathbf{Y} = \rho\mathbf{W}\mathbf{Y} + \mathbf{X}\beta + \varepsilon$$

where:

- **Y** is the vector of the dependent variable.
- **X** is the matrix of independent variables.
- β is the vector of fixed effects coefficients.
- ρ is the *spatial autoregressive coefficient*, quantifying the strength of the spatial dependence in the response.
- **WY** is the spatially lagged response variable, representing the weighted average of neighbors' outcomes.
- ε is the vector of i.i.d. error terms.

Conditional Autoregressive (CAR) Model

The Conditional Autoregressive (CAR) model accounts for spatial dependence through the *error term (ε)*, implying that the spatial pattern is due to unobserved factors. The model assumes that the error at location i is conditionally dependent on the errors of its neighbors.

The model is typically expressed through the structure of the error term:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where the error term ε follows a conditional distribution:

$$\varepsilon_i \mid \varepsilon_{-i} \sim N\left(\lambda \sum_{j=1}^{n} w_{ij}\varepsilon_j, \tau_i^2\right)$$

The CAR model is often employed in Bayesian spatial analysis and is useful when dependence is driven by inherent spatial heterogeneity or measurement errors. λ is the spatial dependency parameter, and $\mathbf{W} = \{w_{ij}\}$ is the spatial weights matrix.

Spatial Dependency Diagnostics and Model Selection

Moran's I Statistic To formally test for the presence of spatial autocorrelation in the observed response variable or the residuals of a non-spatial model, we calculate Moran's I statistic (I). This statistic assesses the degree of linear association between the observations and the spatially lagged values ($\mathbf{WY}$). Moran's I is defined as:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

A statistically significant Moran's I value indicates the need for a spatial regression model. We use this diagnostic on the residuals of the initial Ordinary Least Squares (OLS) model to justify the adoption of the SAR or CAR structure.

Model Comparison The choice between the non-spatial (OLS) model and the spatial models (SAR, CAR), or between the SAR and CAR models themselves, is determined by comparing their goodness-of-fit and parsimony. We utilize the following metrics:

- **Log-Likelihood (ℓ)**: Measures how well the model fits the data (higher is better).
- **Akaike Information Criterion (AIC)**: Penalizes the log-likelihood for the number of parameters, rewarding better fit while penalizing model complexity.
- **Bayesian Information Criterion (BIC)**: Similar to AIC but imposes a stronger penalty on models with more parameters, especially useful with large sample sizes.

The model with the highest log-likelihood and the lowest AIC/BIC value is selected as the best fit for the data, as it most effectively accounts for the spatial dependency observed.

3.4 Point Processes Mapping

Point process analysis focuses on patterns of discrete events (points) distributed randomly or non-randomly in space. We utilize two primary tools to visualize and statistically test the spatial distribution of the events under study (e.g., college locations).

Kernel Density Estimation

We begin with Kernel Density Estimation (KDE) to visualize the intensity of the point pattern across the study region. KDE estimates the density function $\lambda(\mathbf{s})$ at any spatial location $\mathbf{s}$ by placing a smooth kernel function, $k(\cdot)$, over each observed event $\mathbf{s}_i$.

The estimated density function $\hat{\lambda}(\mathbf{s})$ is calculated as:

$$\hat{\lambda}(\mathbf{s}) = \frac{1}{nb^2} \sum_{i=1}^{n} k\left(\frac{\mathbf{s} - \mathbf{s}_i}{b}\right)$$

where:

- n is the total number of points.
- $k(\cdot)$ is the kernel function (e.g., Gaussian, Quantic).
- b is the bandwidth (smoothing parameter), which controls the degree of spatial smoothing.

KDE is used here to identify and visually represent areas of high spatial intensity (hotspots) for the events under investigation.

K-function Analysis

The K-function, $K(r)$, is a statistical tool used to test for Complete Spatial Randomness (CSR) and to quantify the type and scale of non-randomness (clustering or regularity/inhibition). It is defined as the expected number of neighboring events within a distance r of an arbitrary event, normalized by the overall intensity (λ). The empirical K-function, $\hat{K}(r)$, is estimated as:

$$\hat{K}(r) = \frac{A}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} I(\|\mathbf{s}_i - \mathbf{s}_j\| \leq r) w_{ij}^{-1}$$

where A is the area of the study region, and w_{ij} is a boundary correction factor (to account for edge effects).

The $\hat{K}(r)$ is typically compared against the theoretical $K(r)$ for CSR, which is πr^2.

- If $\hat{K}(r) > \pi r^2$, the points are **clustered** at distance r.
- If $\hat{K}(r) < \pi r^2$, the points exhibit **inhibition** (regularity) at distance r.

The analysis uses **Monte Carlo simulation** to construct an envelope around the CSR value, providing a formal test of significance for the observed clustering or regularity.

4 Capability Framework

To add structure to the principal components, our framework draws on social determinants of health (SDOH). SDOH are "the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks" [13]. While SDOH can be grouped into may domains, we group them into the following five domains:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

We provide a list of nine capabilities that represent the intersection of the results of our principal component analysis and the social determinants of health. The list of capabilities and their respective definitions is included in Table 1. Table 2 associates each capability to a component of the PCA.

Capability	**Definition**
Community Affiliation	Being able to see oneself represented and feel a sense of belonging within one's community.
Education	Being able to attain a high-quality education and engage in learning.
Healthcare	Being able to access timely, high-quality healthcare.
Economic Stability	Being able to maintain economic stability and have spending power as a consumer.
Transportation	Being able to travel to and from locations efficiently and affordably.
Sustenance	Being able to provide oneself and dependents with healthy food and be informed about nutrition.
Technology	Being able to access, understand, and use technology effectively and safely.
Built Environment	Being able to live in a location that is safe and supports healthy living.
Resilience	Being able to recover from natural disasters or other unprecedented events within a reasonable time.

Table 1: Framework components and their definitions.

Capability	**Community Factors**
Community Affiliation	Percent Asian, Percent White, Percent Black, Veteran Population, Median Age, Population Under Age 18, Population Aged 65+, Population Change Since 2014, Population
Education	Total Funding for K-12, Educational Attainment
Healthcare	Traditional Medicaid Enrollment, Medicaid Expansion Enrollment
Economic Stability	Percent Children in Poverty, Average Weekly Wage, Per Capita Income, Building Value
Transportation	Auto/Truck Registration, Primary Highway Mileage
Sustenance	Food Insecurity
Technology	Broadband Internet Access, Computing Device Access
Built Environment	Taxable Property Valuation per Capita, Mountains, Coastal, Number of Farms, Area (sq mi)
Resilience	Hurricane Risk Index Score, Coastal Flooding Risk Index Score, Community Resilience

Table 2: Community factors derived from PCA associated with capabilities.

Figure 1 provides a visual overview schematic of the development of our capability framework. By evaluating the average capabilities of the various COI, we are able to group them in ways that allow for fair representation.

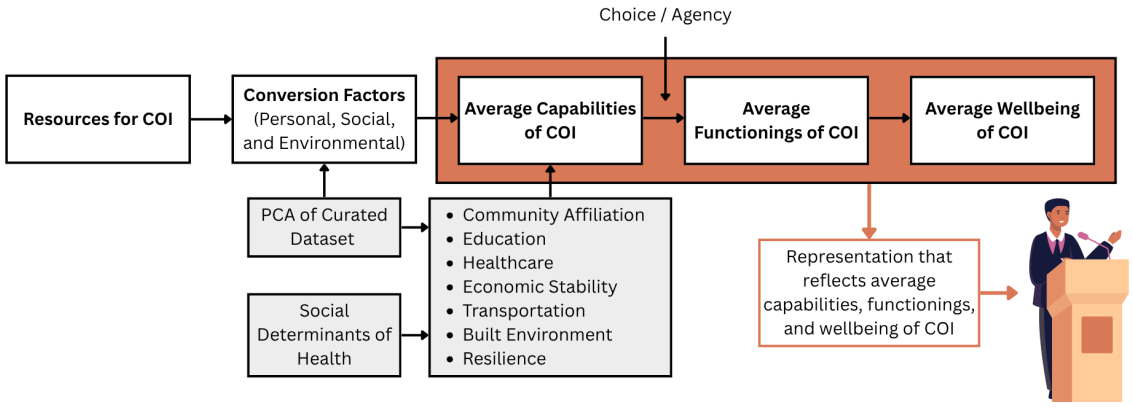


Figure 1: Diagram of the capability approach adapted for COI

5 Methodology

In this section, we introduce a way to quantitatively assess the capabilities of each COI and group them into districts accordingly. We then provide methods to evaluate our redrawn map.

5.1 Defining COI

To quantify the capability of each COI, we define **Grand COI Score**. The dependent variable for all subsequent spatial regression analysis, the Grand COI Score, is a synthetic, composite measure constructed from the raw COI metrics. This score serves as a surrogate for the latent construct of overall community well-being.

The construction of the score involved two main steps. First, the raw features were grouped into nine conceptual determinants (e.g., Economic Stability, Resilience, Technology) based on domain expertise and polarity (positive or negative association with well-being). All raw variables were first standardized to Z-scores (Z_{ij}) to ensure equal weighting and comparability. The composite score for each determinant D in county i was calculated as the sum of the standardized positive indicators minus the sum of the standardized negative indicators:

$$\mathrm{Score}_{D_i} = \sum_{j \in D_{\mathrm{Pos}}} \mathrm{Z}_{ij} - \sum_{k \in D_{\mathrm{Neg}}} \mathrm{Z}_{ik}$$

The second step involved aggregating these nine determinant scores into a single outcome variable. The Grand COI Score ($\mathrm{COI}_i^{\mathrm{Grand}}$) for each county i was calculated by taking the arithmetic mean of the nine determinant scores:

$$\mathrm{COI}_i^{\mathrm{Grand}} = \frac{1}{9} \sum_{D=1}^{9} \mathrm{Score}_{D_i}$$

This final score, ranging from low to high well-being, was used as the continuous response variable in the OLS, SAR, and CAR models. The median values of these scores, profiled across the 20 COI groups, reveal the distinct well-being signature of each contiguous region.

5.2 Evaluating Recent District Maps

To evaluate if recent election maps preserved our defined communities of interest, we evaluated COI evaluated with PCA at $k = 40$ (Fig. 2) against the 2020 NC District Map (Fig. 3), the 2023 NC District Map (Fig. 4), and the 2025 NC District Map (Fig. 5).

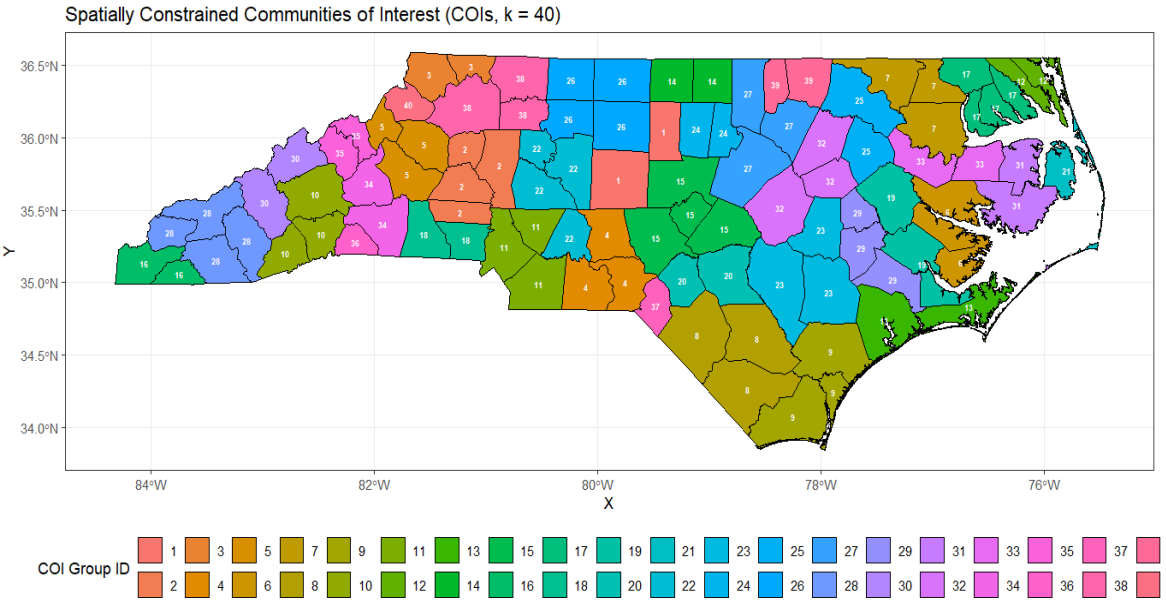


Figure 2: 40 COIs determined by PCA and HSC

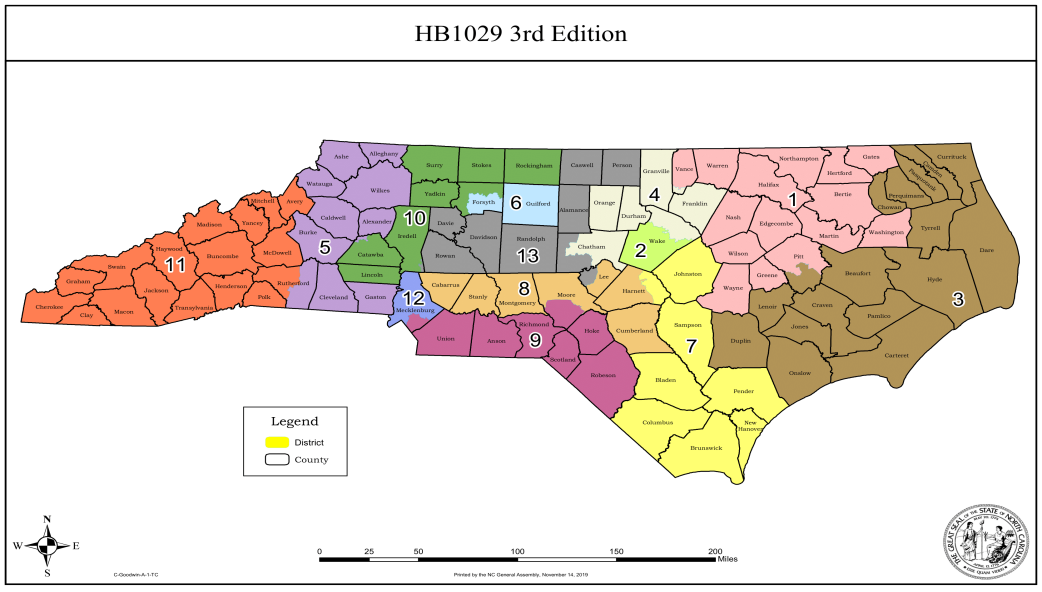


Figure 3: 2020 District Plans for North Carolina

The 2020 District Map is considered to be one of the fairest maps for North Carolina in decades, splitting its then 13 seats into 8 Republican and 5 Democratic seats in the House elections. Since then, the lines have been redrawn, the current 2023 District Plans have split its 14 seats into 10 Republican

and 4 Democratic seats. The recent 2025 District Plans for North Carolina redraw the lines between District 1 and District 3.

We chose $k = 40$ for this assessment, as a k-value that is too low makes assessment too rigid, and a k-value that is too high makes the results too arbitrary. Some assumptions have to be made, as the actual district cross county lines and we do not. As a rule of thumb, the entirety of Mecklenburg and Wake County are always their own district, and for counties that are split on district lines, we group by the district that has more of that county in it by area.

We look at how the district maps interact with the COIs on three metrics:

- **Breaks:** Breaks are defined as when a district line intersects and partitions a COI. If a COI is split into n districts, it is counted as *n-1* breaks. This serves as our measurement of cracking. Lower values preferred.
- **Number of districts that contain only one COI:** Excluding Mecklenburg and Wake County, the number of districts that contain only one COI. This serves as our measurement for packing. We chose one COI as the cutoff since the expected number of COIs per district is around 3, and we allow for an interval of [2, 4] as acceptable leeway for population consideration. Lower values preferred.
- **Number of distinct COI per district:** Simply counting up the number of distinct COI per district. This measures how heterogeneous the district is. Lower values preferred, as we want to avoid clustering a large amount of different COIs together.

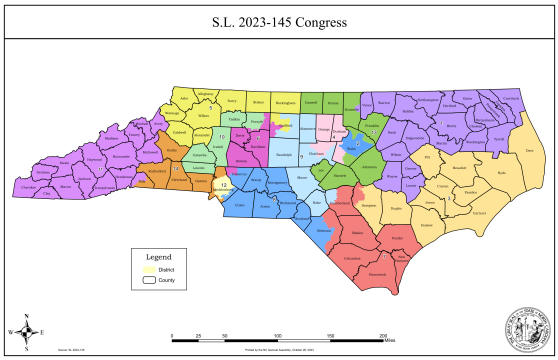


Figure 4: 2023 District Plans for North Carolina

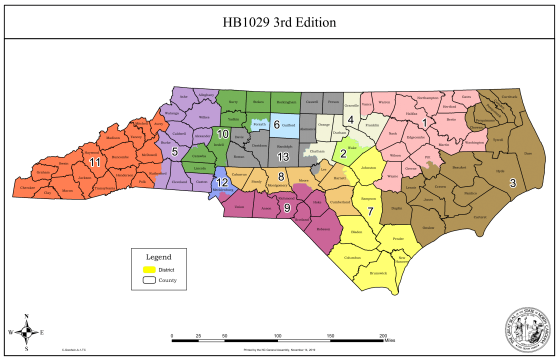


Figure 5: 2025 District Plans for North Carolina

6 Results

Our analysis proceeds in the sequence outlined in Section 3, beginning with data reduction and culminating in the spatial modeling of the *Grand COI Score*, displayed in Table 6, a surrogate measure derived from initial component analysis. We first used PCA to condense the large set of socioeconomic variables into a few uncorrelated factors. The scores from these factors, alongside the results of the Spatially Constrained Hierarchical Clustering, were then used as predictors in the regression models. Crucially, the initial Ordinary Least Squares (OLS) model confirmed the presence of significant spatial autocorrelation, necessitating the comparison and selection of the appropriate spatial model (SAR or CAR) using AIC and BIC. Finally, we present the findings from the point process analysis, detailing the spatial distribution and clustering patterns of the institutions across the study area.

6.1 Dimensionality Reduction and Unsupervised Findings

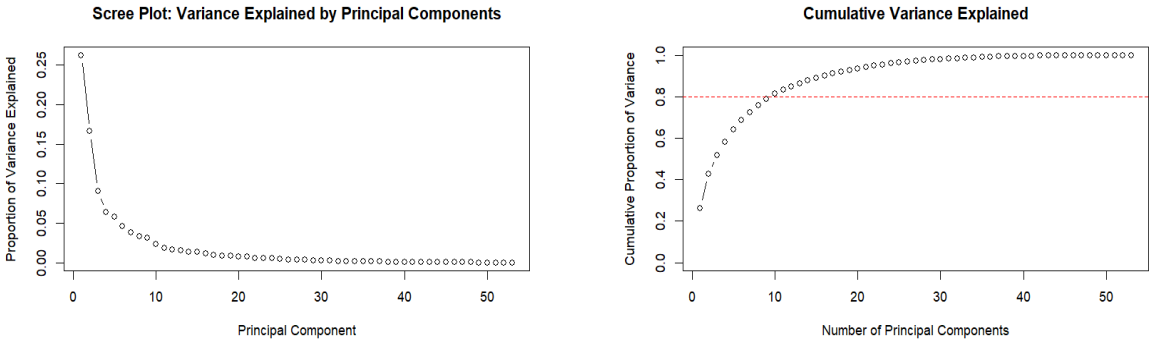
We first applied Principal Component Analysis (PCA) to the initial set of p socioeconomic, demographic, and environmental predictor variables. The goal was to reduce dimensionality and establish a set of

uncorrelated features for subsequent spatial modeling.

Instead of relying on a high cumulative variance threshold (e.g., 80%), which risked creating a component space that too closely resembled the full dataset, we adopted a criterion focused on maximal explanatory power with minimal dimensionality. We selected the minimum number of components K such that the total variance explained by the union of the selected features slightly exceeded 50%. Specifically, we sought the minimum K satisfying:

$$\sum_{k=1}^{K} \text{PVE}_k > 0.50$$

where PVE_k is the Proportion of Variance Explained by PC_k. The first three principal components (PCs) met this criterion: PC1 explained approximately 25.1%, PC2 explained 14.8%, and PC3 explained 11.2%, resulting in a cumulative variance explained of approximately 51.1%. Thus, $\mathbf{K = 3}$ principal components were retained.



(a) Proportion Variance Explained

(b) Cumulative Variance

Figure 6: PCA Results from finalized results

The final set of features used in the regression models was derived via a union-based feature selection method. This method identified all original variables whose absolute loading on any of the top three PCs exceeded a threshold of 0.2. This process ensured that the final feature set of 30 variables represented the maximal explanatory power of the three most significant latent dimensions. These 30 features were subsequently standardized and used as the independent variables ($\mathbf{X}$) in all regression models.

Component	**PVE (%)**	**Primary Interpreted Dimension / High Loadings**
PC1	≈ 25.1	Population Density and Wealth *High loadings:* Population, Building Value, Taxable Property Valuation
PC2	≈ 14.8	Socioeconomic Disadvantage *High loadings:* Percent Children in Poverty, Food Insecurity, Medicaid Enrollment
PC3	≈ 11.2	Demographic Mix and Infrastructure *High loadings:* Percent White/Black, Highway Mileage
Cumulative Variance Explained by 3 PCs: 51.1%		

Table 3: Top Principal Components and Associated Interpretive Loadings

Spatially Constrained Hierarchical Clustering

To identify geographically cohesive regions with similar socioeconomic profiles, we applied Spatially Constrained Hierarchical Clustering. The clustering was performed on the standardized 30-variable feature set, utilizing Ward's linkage method and a distance metric penalized by a spatial contiguity matrix ($\lambda = 50$) to guarantee all resulting clusters were contiguous.

Cluster Selection and Profiling: Based on the dendrogram structure, we selected $\mathbf{k = 20}$ (COI) groups. This selection provided an optimal balance between regional homogeneity and overall group size. As visualized in the map of North Carolina counties , the resulting 20 COI groups are contiguous and reveal distinct geographic patterns, which confirmed the successful segmentation of the state into regions based on the underlying drivers. These 20 COI groups were subsequently introduced into the spatial regression models as categorical fixed effects to capture local heterogeneity.

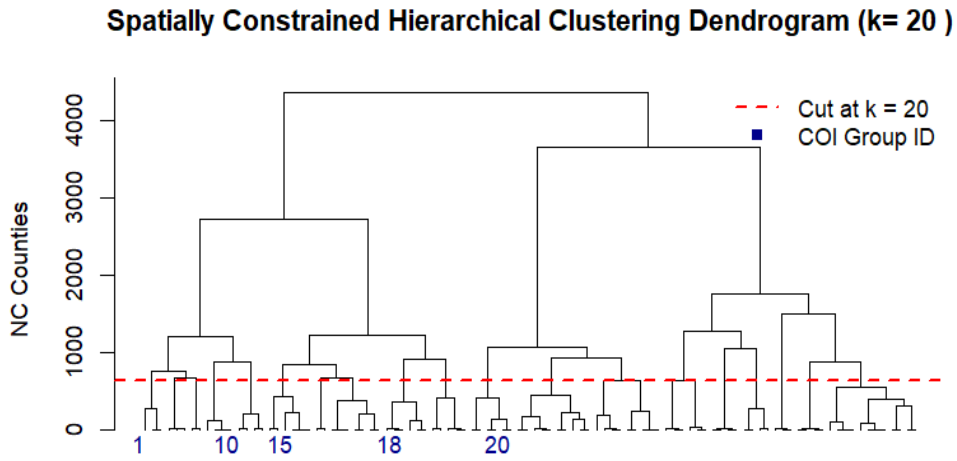


Figure 7: Dendrogram representing clustering of NC counties.

To identify geographically cohesive regions with similar socioeconomic profiles, we applied *Spatially Constrained Hierarchical Clustering*. The clustering was performed on the standardized 30-variable feature set using Ward's linkage and a distance metric penalized by a spatial contiguity matrix ($\lambda = 50$), which guaranteed that all resulting clusters formed contiguous geographical regions. Based on the structure of the dendrogram, we selected $\mathbf{k = 20}$ Communities of Interest (COI) groups. The geographical locations of these 20 COI groups are presented in Figure 8.

The result of this unsupervised learning step was twofold: first, the resulting Group ID was attached to the spatial data, allowing for the direct visualization of the contiguous regions and confirming the successful segmentation of the study area; and second, the grouping variable enabled the calculation of robust median profiles for all 30 features across the 20 clusters. These 20 COI groups were ultimately introduced into the spatial regression models as categorical fixed effects to account for unobserved regional heterogeneity, as seen in Figure 8.

6.2 Spatial Regression and Model Selection

We began by estimating the Grand COI score across North Carolina counties using an ordinary least squares (OLS) regression. The model included six predictors selected to capture overarching community characteristics representative of the state's population: *% Children in Poverty*, *Educational Attainment*, *Broadband Internet Access*, *Hurricane Risk Index Score*, *Population Change*, and *Median Age*. The OLS estimates, shown in Table 4, serve as a baseline for assessing the magnitude and direction of associations between the predictors and the Grand COI score.

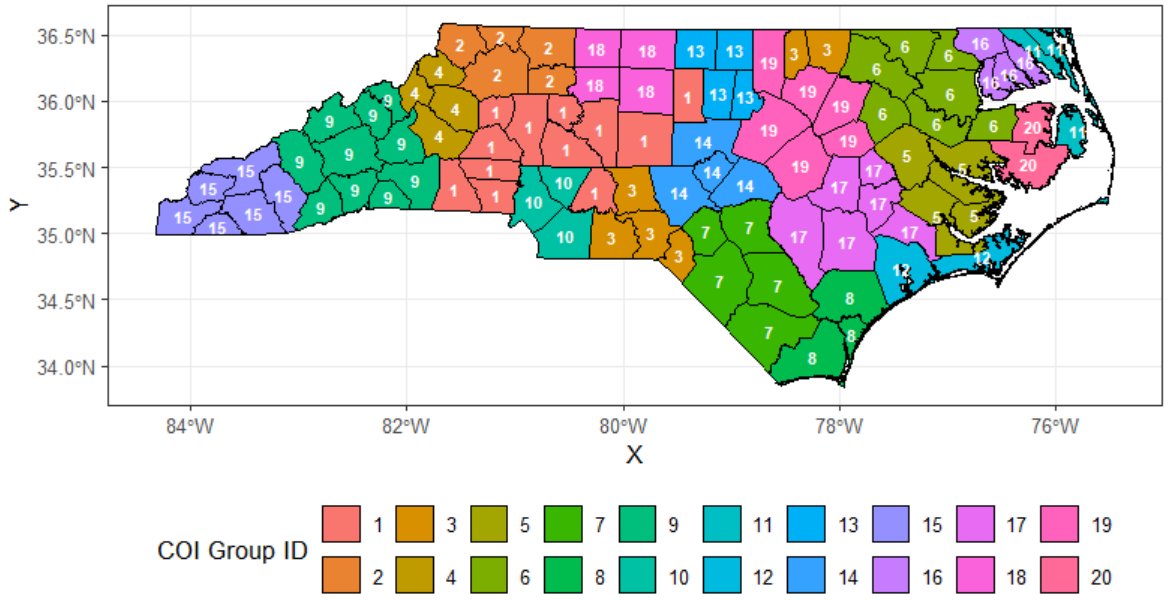


Figure 8: COI derived from PCA and HSC with $k = 20$.

Predictor	Estimate	Std. Error	t value	p-value
Percent Children in Poverty	-4.502	0.730	-6.172	1.75×10^{-8}
Educational Attainment	3.748	0.424	8.833	6.00×10^{-14}
Broadband Internet Access 2022	4.339	0.683	6.351	7.79×10^{-9}
Hurricane Risk Index Score	-0.00511	0.00199	-2.570	0.0118
Population Change Since 2014	0.00001789	0.000001141	15.674	$< 2 \times 10^{-16}$
Median Age	-0.01661	0.006362	-2.611	0.0105

Table 4: OLS Regression Results for Grand COI Score (Intercept excluded). RSE = 0.2741 on 93 DF; Adjusted $R^2 = 0.9579$; F-statistic = 376 on 6 and 93 DF, $p < 2.2 \times 10^{-16}$.

6.2.1 Assessment of Spatial Dependence

To evaluate the presence of spatial autocorrelation in the residuals of the OLS model, we computed Moran's I statistic. Significant spatial autocorrelation indicates that geographically neighboring counties tend to have more similar COI scores than expected under independence.

The map of standardized residuals (Figure 9) highlights areas where the OLS model systematically over- or under-predicts the Grand COI score, suggesting potential spatial dependence.

To assess whether the OLS residuals exhibit spatial autocorrelation, we computed the Global Moran's I statistic. Let e_i denote the residual for county i, and $W = [w_{ij}]$ the spatial weights matrix based on adjacency. The null and alternative hypotheses are:

H_0 : Residuals are spatially independent, i.e., $e \sim N(0, \sigma^2 I)$

H_A : Residuals are spatially autocorrelated, i.e., neighboring residuals are more similar than expected.

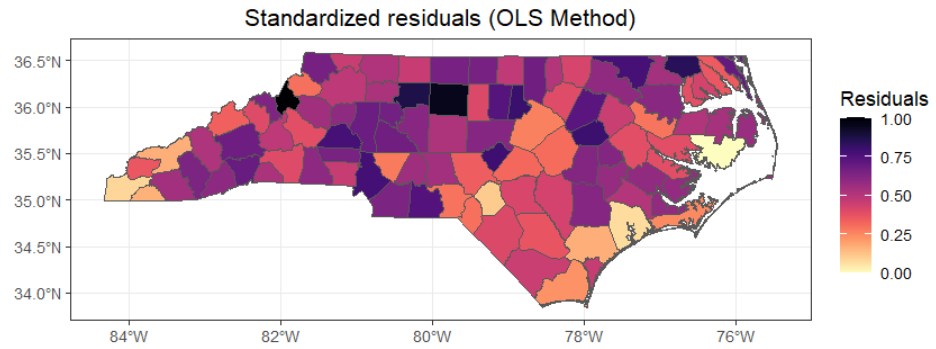


Figure 9: Standardized residuals from the OLS model, illustrating spatial clustering.

The Moran's I test yields $I = 0.153$, with a standard deviate of 2.843 and $p = 0.0045$. Since $p < 0.01$, we reject H_0, indicating that the OLS residuals are not independent across space and suggesting that a spatial regression model (SAR or CAR) may be more appropriate.

6.2.2 Spatial Regression Models: SAR and CAR

Given evidence of spatial autocorrelation, we fitted two spatial regression models: a spatial autoregressive (SAR) model and a conditional autoregressive (CAR) model. Both models incorporate a spatial weights matrix to account for dependencies between neighboring counties.

Model	df	AIC	BIC	Log Likelihood
CAR	2	67.168	72.378	-31.584
SAR	2	63.499	68.709	-29.750

Table 5: Comparison of Spatial Regression Models (SAR vs. CAR)

These spatial regression models adjust the coefficient estimates to account for spatial structure, resulting in different parameter magnitudes and standard errors compared to the OLS model. As shown in Table 5, the SAR model exhibits lower AIC and BIC values than the CAR model, as well as a higher log-likelihood, suggesting that it provides a slightly better fit to the data. Therefore, the SAR model may be preferred for future modeling of Grand COI scores across North Carolina counties.

6.2.3 Simulation of Parameter Distributions

To further examine the variability and discrepancies of the parameter estimates across models, we simulated 1000 draws from the estimated distributions of each coefficient for the OLS, SAR, and CAR models. Boxplots of these simulations (Figure 10) illustrate both the central tendency and variability of each predictor's effect under the three modeling approaches.

Notably, the magnitude and dispersion of coefficients differ across models, with spatial models (SAR and CAR) generally exhibiting shrinkage and reduced variance for certain predictors compared to OLS. These differences highlight the importance of accounting for spatial dependence when interpreting associations at the county level.

Overall, the results demonstrate that while OLS provides a straightforward estimate of associations between community characteristics and the Grand COI score, spatial dependence is present and materially affects parameter estimates. SAR and CAR models offer adjusted estimates that account for spatial autocorrelation, and simulations confirm that ignoring spatial structure may lead to over- or under-estimation of key effects.

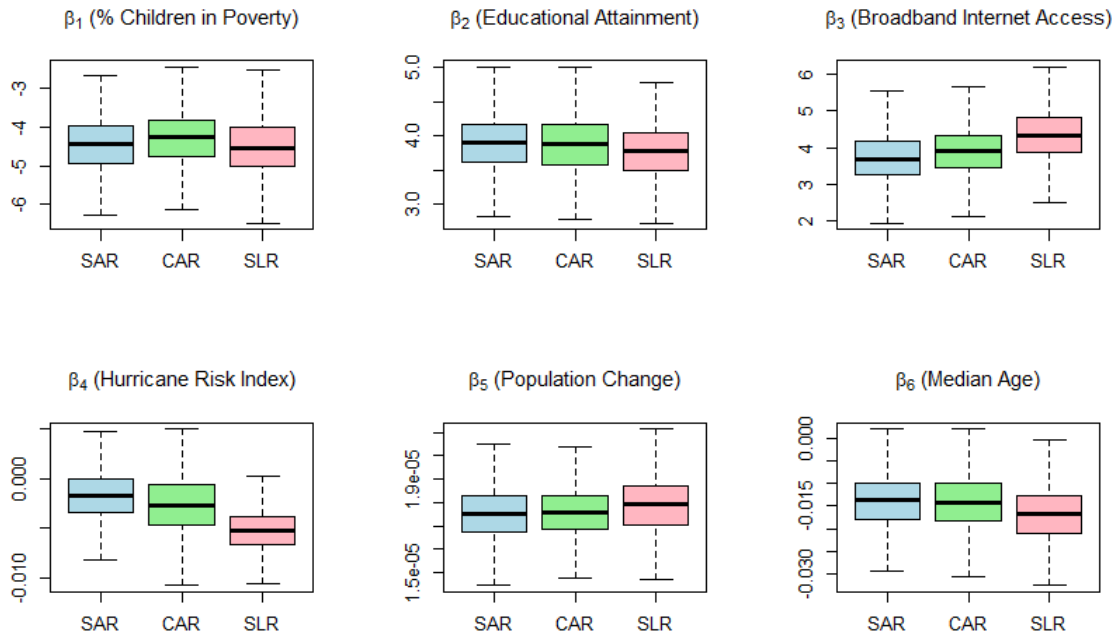


Figure 10: Simulated distributions of coefficients for each predictor under OLS, SAR, and CAR models. Each boxplot summarizes 1000 draws for each paramter estimate.

6.3 Point Process Analysis

To investigate the spatial patterns of higher education institutions in North Carolina, we obtained data from *NC OneMap*. Institutions were categorized based on whether their enrollment was above or below the median across all schools.

6.3.1 Point Pattern Mapping

We visualized the locations of institutions on a North Carolina polygon map. Figure 11 shows the distribution of institutions with enrollment above the median (left panel) and below the median (right panel). These maps provide an initial visual assessment of potential clustering patterns.

6.3.2 Kernel Density Estimation

This final section analyzes the spatial patterns of higher education institutions, segmented by their enrollment status (greater than or less than the state median), using point pattern methods. We first applied a *2-dimensional kernel density estimator* to quantify the spatial intensity, $\hat{\lambda}(\mathbf{s})$, for each institution group.

The resulting intensity maps (Figure 12) highlight distinct spatial regimes: institutions with enrollment **greater than the median** exhibit clear and pronounced clustering, with high concentration "hotspots" forming around major urban and economic centers, such as the Research Triangle and the Piedmont Triad regions. Conversely, institutions with enrollment **less than the median** show a much lower overall intensity and a more dispersed pattern, suggesting a distribution primarily serving surrounding smaller metropolitan and rural areas.

To formally assess spatial clustering relative to the null hypothesis of **Complete Spatial Randomness (CSR)**, we computed *Ripley's K-function,* $K(h)$, for both enrollment groups (Figure 12). The results statistically confirm distinct spatial patterns: institutions with enrollment above the median exhibit significant clustering, forming tightly-concentrated "hotspots" often located near urban centers, whereas

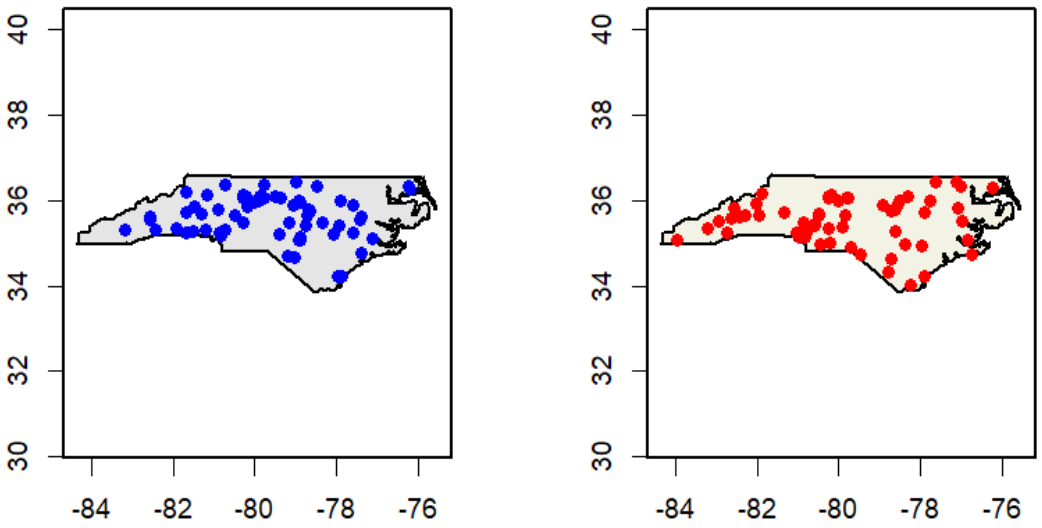


Figure 11: Point processing patterns of higher education (i.e., college, university, trade schools) institutions in North Carolina. Left (BLUE): enrollment above median enrollment count. Right (RED): enrollment below median enrollment count.

institutions below the median are more *evenly* distributed across the state. A continuation of this analysis could examine the difference between the two underlying point processes rather than just their K-function summaries. For example, fitting spatial point process models to each enrollment group and use Markov Chain Monte Carlo (MCMC) methods to estimate the posterior difference in their spatial intensities. This would yield a spatially explicit "difference surface", highlighting regions where low-enrollment institutions are systematically underrepresented and may require greater educational investment. This spatial heterogeneity highlights areas of concentrated higher education activity and has important implications for resource planning, student accessibility, and the regional economic impact of higher education in North Carolina.

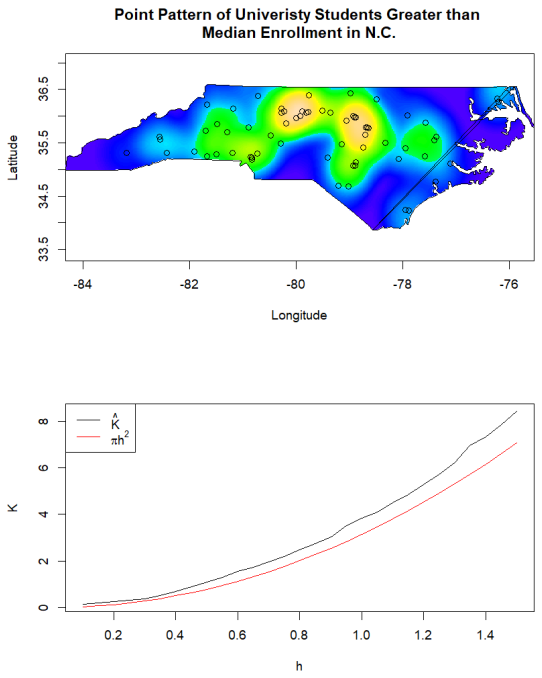
7 COI Protection of Recent District Maps

Using the metrics outlined in the methodology, we evaluate the effectiveness of recent North Carolina district maps in protecting COI.

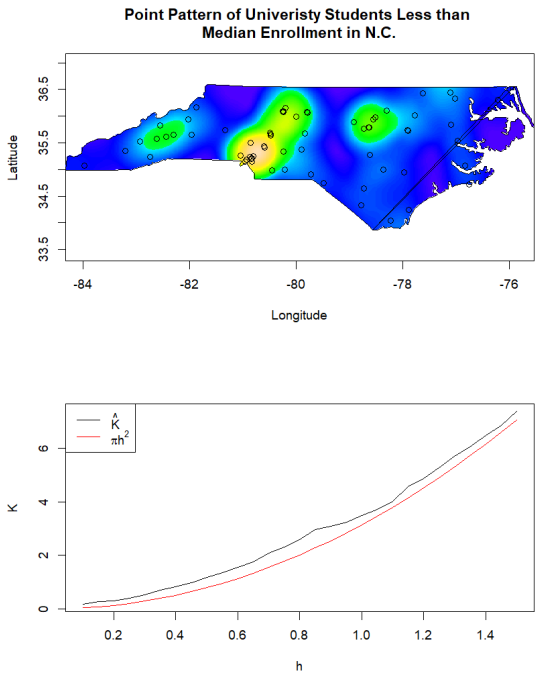
Breaks: We find that each of the three different district maps yielded 17 breaks, indicating that none of the maps are better at retaining COI than the other maps.

Number of districts that contain only one COI: We find that the 2020 map only contains part of COI group 26 (representing Forsyth and Guilford County) into what at the time was District 6. However, both the 2023 and 2025 district map group together part of COI 22 (representing Davie, Davidson, and Rowan County) into the current and future District 6 and all of COI 24 (representing Orange and Durham County) into the current and future District 4.

Number of distinct COI per district: Fig 13 shows the number of distinct COI per district. We find that the 2020 map has the lowest maximum value at 9, and that the 2025 map has the highest maximum value, as it pulls together 12 COI into District 1. This indicates that the heterogeneity of the districts in the more recent maps is higher than the 2020 district map. From the differences in the last two metrics, we see that the 2023 and 2025 district maps are less protective of COIs than the 2020 district map, and the 2025 district map is less protective of COIs than the 2023 district map, implying that the district maps have become more unfair over time.



(a) Greater than Median Enrollment - Intensity surface and K-function plots



(b) Less than Median Enrollment - Intensity surface and K-function plots

Figure 12: Stacked visualizations of higher education spatial patterns by enrollment group.

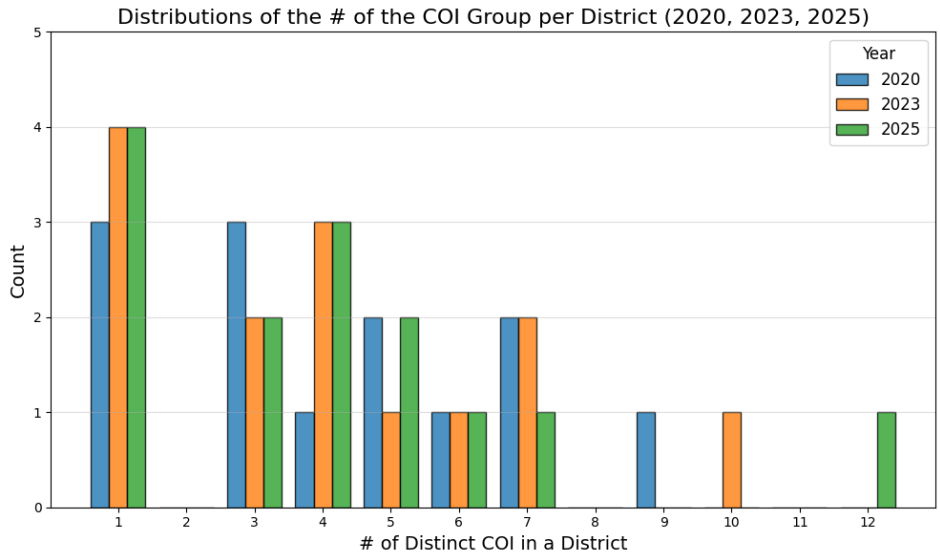


Figure 13: Distributions of the Number of COI Group per District in 2020, 2023, and 2025. We see that (including Mecklenburg and Wake County) a total of 3 districts had only 1 COI group represented in 2020 while in 2023 and 2025, this number jumps up to 4. The maximum number of distinct COI also increases over the years.

8 Discussion

8.1 Limitations and Future Considerations

Some key considerations of the North Carolina government when redistricting counties lie in the concepts of contiguity and county-splitting. As mentioned in the introduction, the North Carolina Constitution Article 2 Section 5, 2 states that "Each representative district shall at all times consist of contiguous territory", meaning that any congressional district should consist of areas that never be split into more than one continuous area. The North Carolina Constitution Article 2 Section 5, 3 states that "No county shall be divided in the formation of a representative district," with a very direct implication of no county should be divided in order to form a congressional district. Our methods follow both of these legislatures, however North Carolina congressional district maps have historically (and notoriously) divided up counties; We defined a county to be the smallest unit we could work with.

By constraining the granularity of our data to the county level, we lose the ability to properly account for population or geographical features like rivers that are considered when developing district maps. Future work should look at more granular measures such as precincts or zip codes. Due to the availability of the data we wished to consider, our compiled dataset consists of data spanning from 2021-2024. Future work should use data complied from a smaller time-span to yield more temporally accurate results.

Future work should look at defining COI with different k values and considering different variables, such as precincts, in COI analysis. Additionally, future work should explore how compact these COI regions can be.

8.2 Conclusion

The *Grand COI Score* provides a quantitative measure of multidimensional well-being for each North Carolina county. Since the final set of 20 contiguous COI groups represents regions of internal homogeneity in terms of this score, the subsequent step focuses on defining a simpler, more granular set of $k = 40$ COIs to enable a detailed assessment against recent congressional maps. This larger number of regions, though sharing the same foundational PCA and clustering methodology, provides a finer scale of segmentation necessary to detect instances where political boundaries may crack or pack these empirically derived communities.

Our core evaluation metric is the degree to which existing congressional districts (from 2020, 2023, and 2025) violate the integrity of these 40 COI boundaries. An ideal map would have district lines that closely follow the boundaries of the COI groups, thereby ensuring communities with shared socioeconomic interests are represented by a single official. This allows for a direct, objective comparison of how well the politically drawn maps align with the statistically determined communities, providing empirical evidence to assess the impact of recent redistricting efforts, particularly the controversial 2025 map, on the fundamental principle of fair representation.

References

[1] The Associated Press. 2024 election results. `https://apnews.com/projects/election-results-2024/`, 2024. Accessed November 16, 2025.

[2] North Carolina General Assembly. North carolina state constitution, article 2, 2025. Accessed November 16, 2025.

[3] Merriam-Webster, Inc. Gerrymandering, n.d. Accessed November 16, 2025.

[4] Madhukara Kekulandara and Edmund A Lamagna. A partial map mcmc algorithm for addressing racial gerrymandering challenges: A case study of alabama's congressional districts. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 77–88, 2025.

[5] Daniel D. Polsby and Robert D. Popper. The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale Law & Policy Review*, 9:301–353, 1991.

[6] Ernest C. Reock. A note: Measuring compactness as a requirement of legislative apportionment. *Midwest Journal of Political Science*, 5(1):70–74, 1961.

[7] Administration & Cost of Elections Project (ACE). Communities of interest – establishment of criteria for delimiting districts (bdb05c), 2025.

[8] Amartya Sen. Commodities and capabilities. *Oxford University Press*, 1999.

[9] Ingrid Robeyns. *Wellbeing, Freedom and Social Justice: The Capability Approach Re-examined*. Open Book Publishers, 2017.

[10] Frances Stewart. Groups and capabilities. *Journal of Human Development*, 6(2):185–204, 2005.

[11] North Carolina Association of County Commissioners. County data and information, n.d.

[12] North Carolina Department of Health and Human Services. 2022 ethnicity and racial data for north carolina by county, 2024.

[13] Office of Disease Prevention and Health Promotion. Social determinants of health, 2020.

9 Appendix

COI	AFFIL	EDUC	HCARE	ECON	TPORT	SUST	TECH	ENVR	RESIL
1	0.392	-0.249	0.588	1.365	0.499	0.318	0.859	0.413	1.126
2	-0.607	-0.736	0.139	-1.494	-0.906	-1.033	-0.437	0.202	0.731
3	-0.137	-1.332	-3.38	-2.978	-1.027	-0.65	-1.552	-0.118	-0.293
4	-0.325	-0.695	0.476	-0.993	-0.871	-0.09	-0.287	1.05	1.415
5	-0.651	-0.316	0.072	0.054	0.135	0.063	0.198	-1.181	-1.351
6	-1.052	-1.184	-2.614	-3.394	-0.428	-0.472	-3.901	-0.361	-2.0
7	0.196	-1.11	-2.464	-1.916	1.363	-0.778	-1.408	-1.74	-1.983
8	0.529	0.518	1.554	1.987	0.583	0.853	2.977	0.491	-1.419
9	-1.126	0.04	0.61	-0.78	-0.67	-0.064	-0.58	0.513	1.771
10	2.867	2.138	1.815	5.943	1.507	2.025	2.989	0.467	1.194
11	0.3	0.637	3.292	1.881	-1.386	1.719	3.139	0.998	-0.773
12	0.609	0.364	1.311	1.085	0.042	0.165	2.634	-1.865	-1.094
13	0.223	1.77	1.044	3.465	-0.55	0.955	1.712	0.504	1.434
14	-0.239	0.784	1.552	1.418	0.352	0.751	0.958	0.049	0.189
15	-1.64	-0.529	-0.449	-1.396	-1.404	-0.727	-0.961	1.889	1.385
16	-0.826	-1.08	0.249	-1.156	-1.733	0.293	-0.345	0.611	-2.161
17	0.03	-1.038	-1.235	-1.724	0.252	-0.523	-1.502	-0.717	-2.107
18	0.524	0.789	0.319	1.455	1.586	0.191	0.697	-0.484	0.678
19	0.482	-0.063	1.638	0.813	0.793	1.286	1.052	-0.831	0.511
20	-0.904	-1.715	0.038	-2.859	-1.722	-1.644	-3.004	-1.51	-3.52

Table 6: Capability scores by COI group. The highest and lowest values per capability are highlighted. The capabilities are abbreviated as follows: Community Affiliation - AFFL, Education - EDUC, Healthcare - HCARE, Economic Stability - ECON, Transportation - TPORT, Sustenance - SUST, Technology - TECH, Built Environment - ENVR, Resilience - RESIL

Code and Files: GitHub - Project Link