Module 7

Spatial Cluster Analysis

For this task, we will use data on crime statistics in the southern U.S. states for the 1960s–1990s available in the "south" archive on our Google Drive (https://drive.google.com/file/d/1j5NVYi6UuBXEEm3jhXxyRAUelypn6aK9/view?usp=drive_link) and website (https://mgimo.ru/upload/2023/05/spatial-cluster-analysis.zip). Original source: https://geodacenter.github.io/data-and-lab//south/. Open the shapefile in GeoDa.

Principal component analysis (PCA)

Principal component analysis (PCA) is the definition of new features, which are linear combinations of the variables being studied and which essentially "absorb" most of their variability and thus convey most of the information about them.

We'll calculate the principal components for the following indicators for year 1990 from the database: homicides per 100,000 people (HR90), population (PO90), unemployment rate (UE90), divorce rate (DV90), median age (MA90), number of African Americans (BLK90) and the Gini index (GI89).

1) To do this, select Clusters → PCA.

2) Choose the variables on the left and start the calculation.

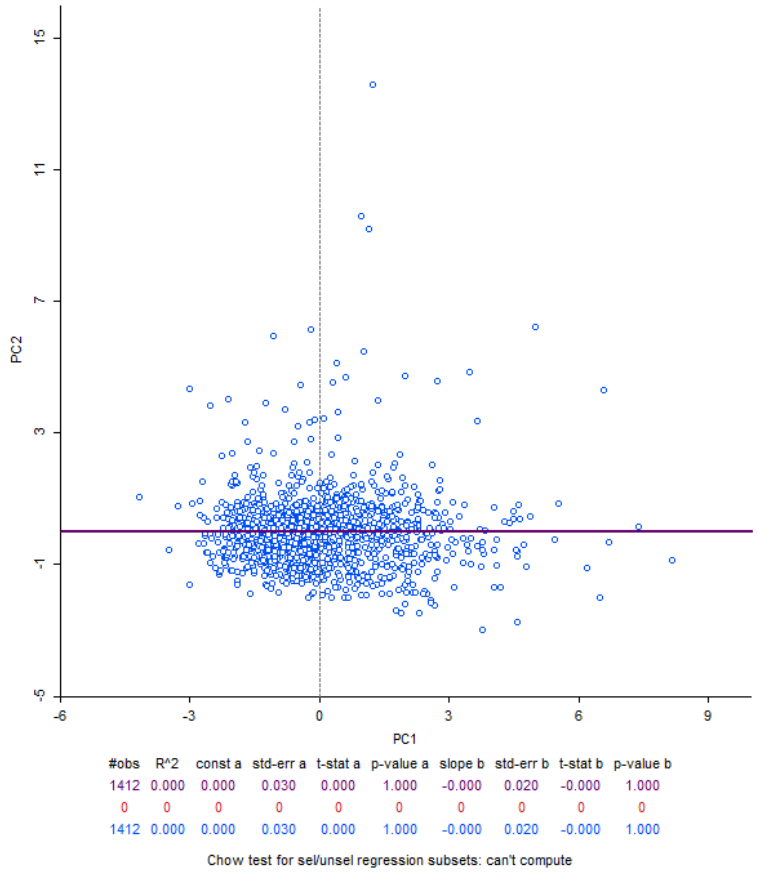
The results will be displayed in the field on the right. Save the indicators as separate columns in the attribute table.

The principal components are those whose eigenvalues exceed 1 (the Kaiser criterion), since the sum of the values of the components will be equal to the sum of the values themselves, in our case, seven.

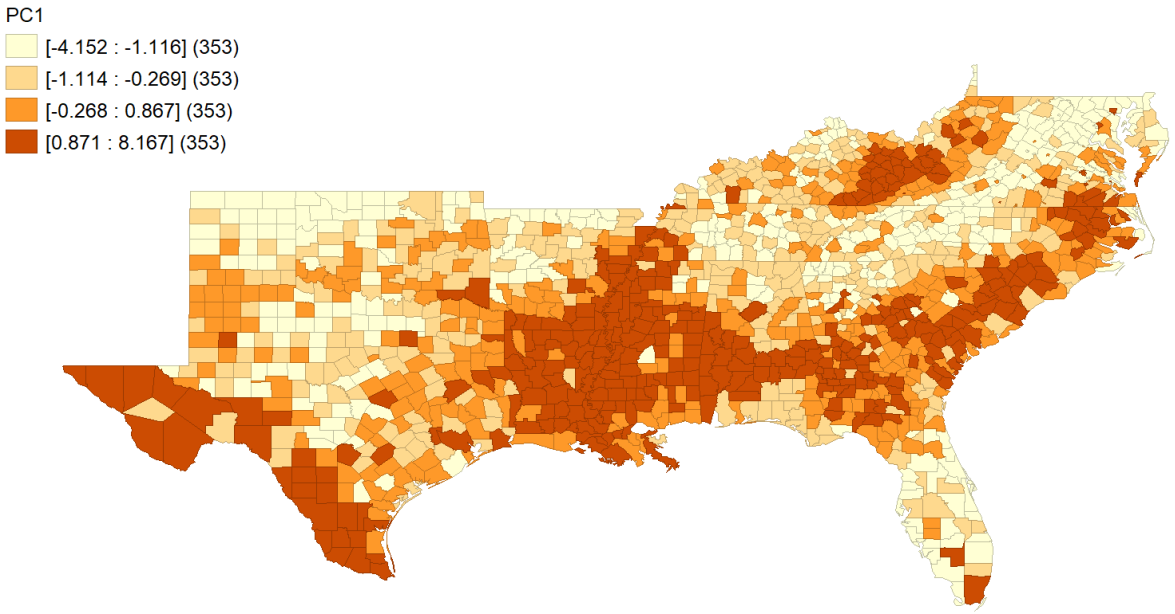
In our case, the first three components will be the principal ones, and the explanatory power of the first component (2.32) will be significantly higher than the power of the second and third components (1.29 and 1.14, respectively).

The proportion of variance gives us an idea of the percentage of change each of our principal components accumulates – 33, 18 and 16, respectively, or 67% in total (approximately two thirds).

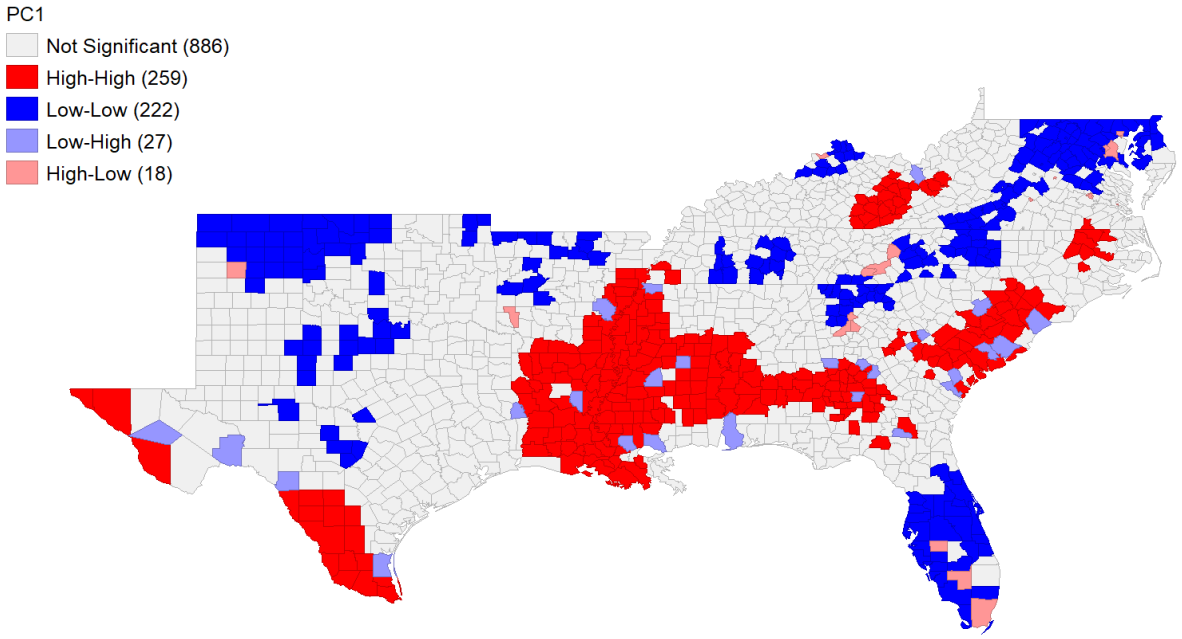
Our task now is to try and understand exactly what our components mean. To do this, we need to look at the columns under the graph, which describe the correlation coefficient of each component with the original variables. As we can see, our second variable is closest to the "population" indicator, the third is closest to "divorce rate," and the first accumulates the remaining indicators. We'll plot the first two components on a scatter plot to make sure there is no correlation between them.



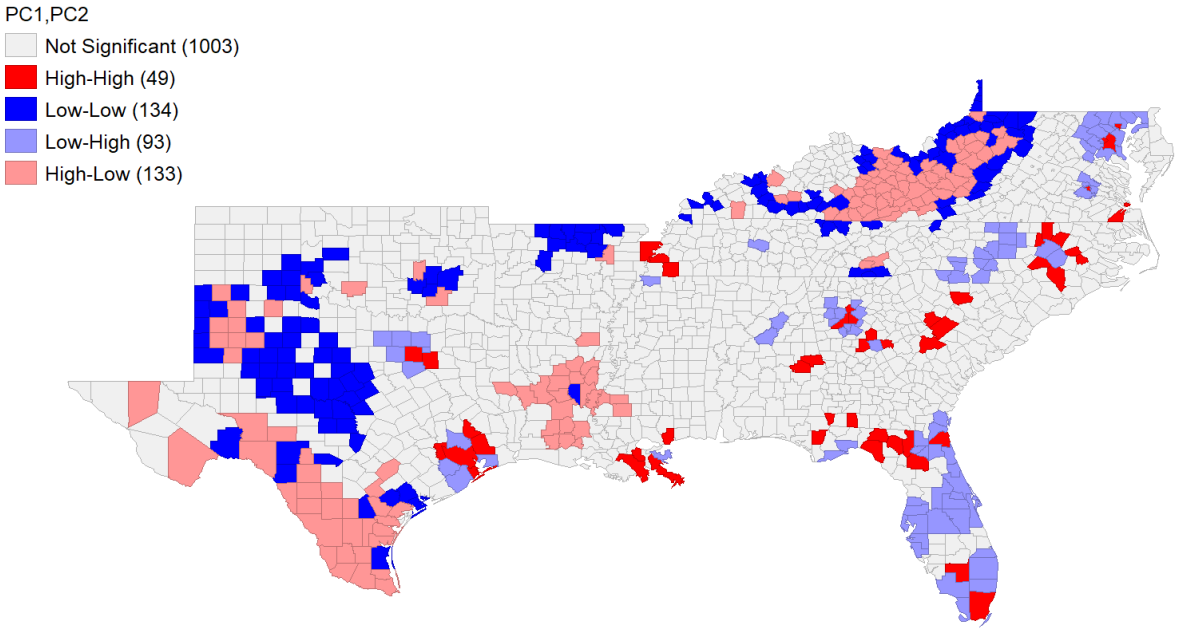
Now let's create a cartogram with four classes for the first principal component to visually analyse the territorial pattern of its distribution.



Our next step after creating a spatial weights matrix (say, using the k-nearest neighbours method, with k = 6) is to calculate and map LISA for the first principal component in order to identify those spatial clusters where it has a high explanatory power and those where it does not.



To finish up, we will analyse bivariate local indicators of spatial autocorrelation of the first and second principal components in order to identify spatial clusters where they cumulatively have explanatory power.

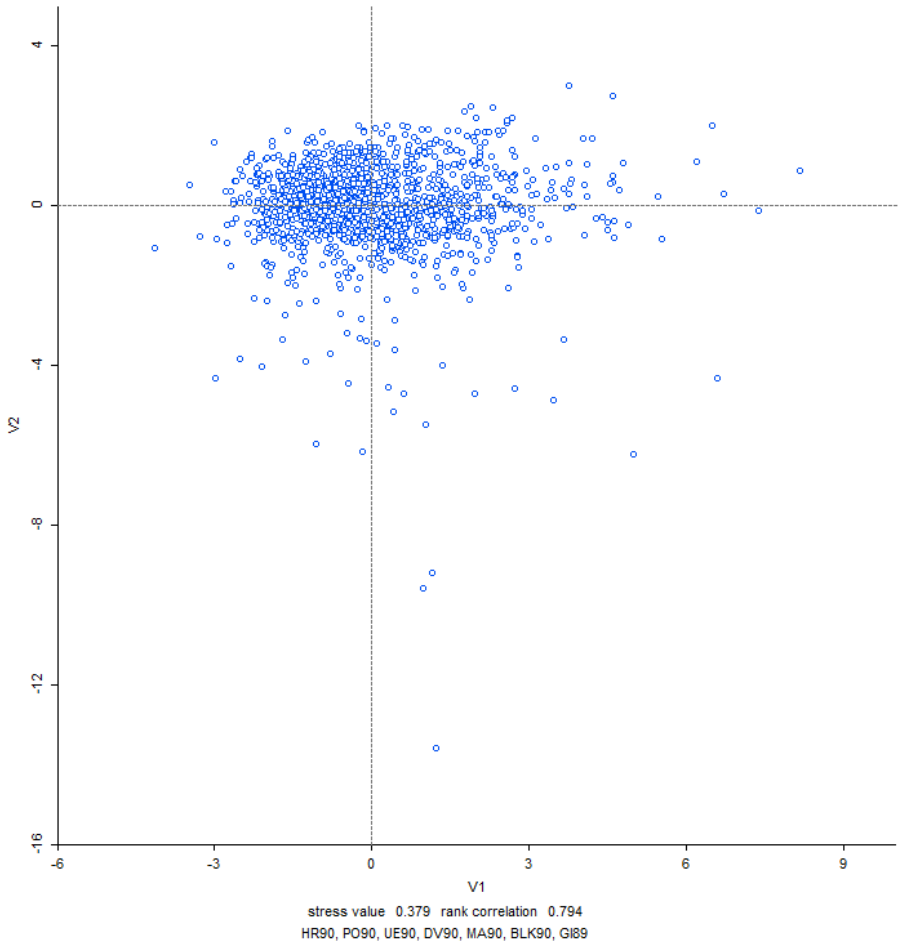


Multidimensional scaling

Another possible dimensionality reduction technique is multidimensional scaling (MDS). This involves showing the differences between objects in the form of a two-dimensional graph. In other words, the closer two points are on a graph, the more similar they are.

1) To perform multidimensional scaling in GeoDa, select Clusters & MDS, choose the same variables as in the PCA, and save the two resulting variables.

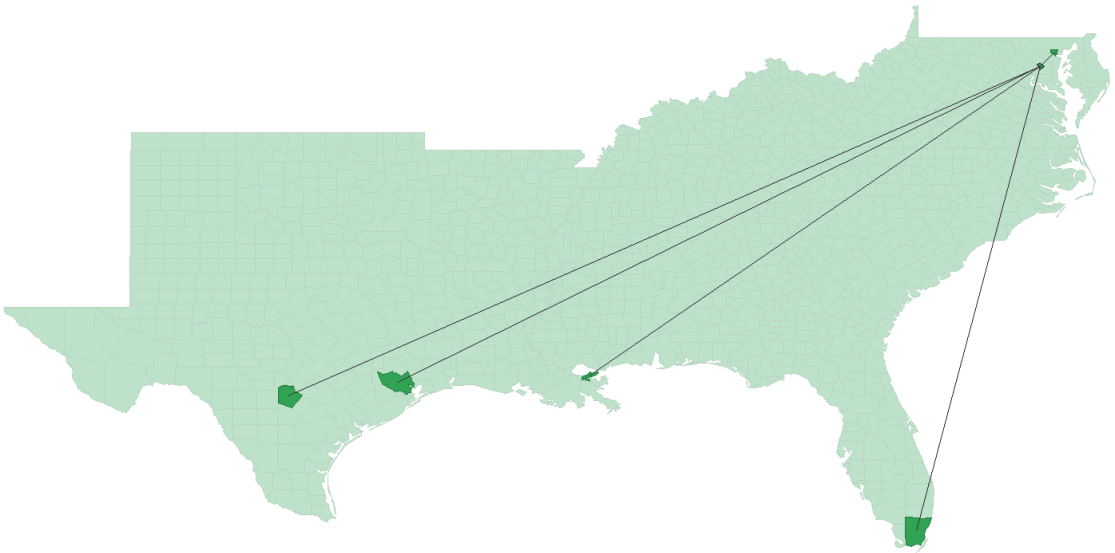
2) Create a scatter plot based on these two parameters.



MDS allows you to create a spatial weights matrix based not on physical proximity, but on analytic proximity – that is, it presents the data in terms of multidimensional scaling.

To do so, right click on the scatter plot we have just created, select Create spatial weights matrix, then set the variable name and adjacency type.

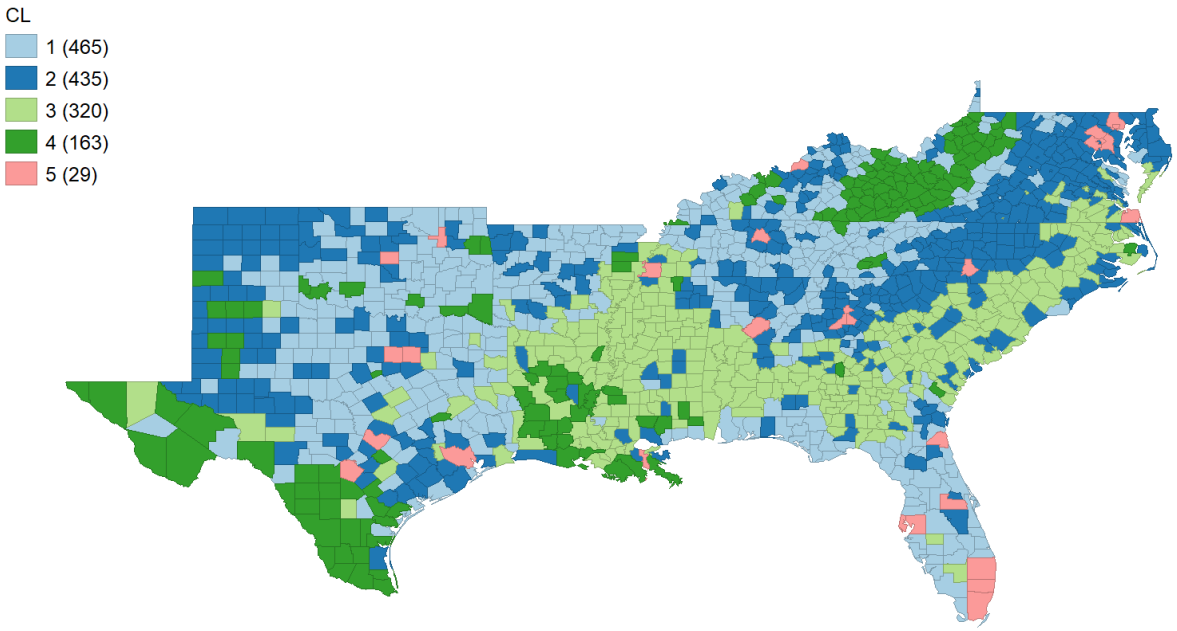
On the adjacency graph for this spatial weights matrix (queen neighbourhood) you can see that the counties with similar crime rates are distant from Washington geographically.



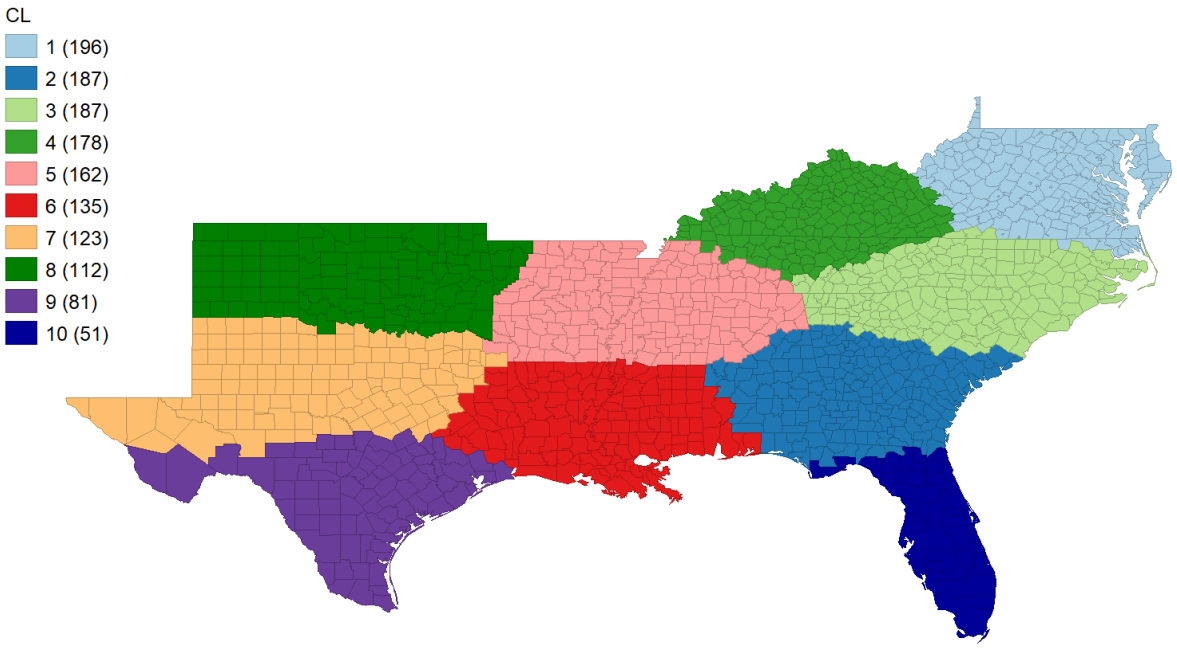
k-Means Clustering

K-means clustering divides the entire dataset of objects into a number of statistical clusters so that the average values differ as much as possible for each of the variables among the clusters.

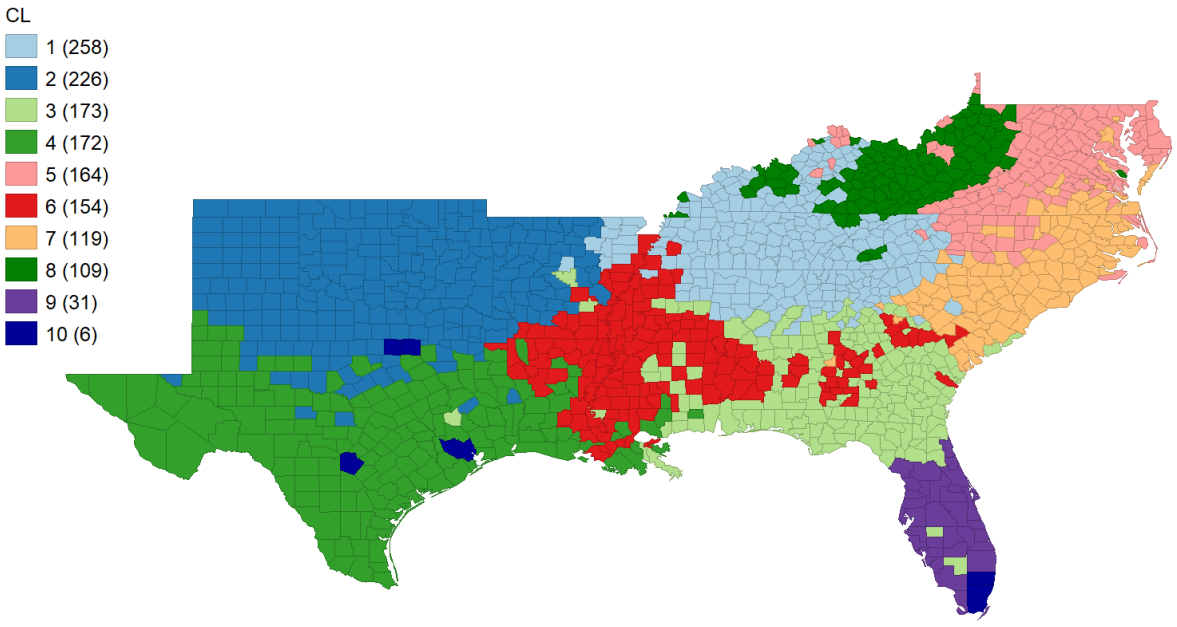
To perform cluster analysis in GeoDa, select Clusters & K-Means, set the same variables, and set the number of clusters to 5. The cluster cartogram will be created.



The resulting clusters do not particularly reflect the spatial concentration of the phenomenon, which can be avoided by checking the box Use geometric centroids. This is because k-means cluster analysis necessarily involves the calculation of centroids – an abstract entity representing a typical representative of the cluster – for input indicators. Adding geometric centroids with a weight of 1 to the calculation will mean that the cluster analysis will ignore all indicators that are not geometric centroids – we will be left with a cartogram with clusters that are not related to the indicators we are analysing.



If, however, we set the weight of geometric centroids to, say, 0.5, then we will get a mixed picture, with the resulting clusters having both a spatial and a statistical variable.



Hierarchical Clustering Method

The hierarchical clustering method allows you to explore the structure of the differences between objects and choose the optimal number of clusters.

1) To perform hierarchical clustering in GeoDa, select Clusters & Hierarchical. Spatial Cluster Analysis

The goal of conventional statistical cluster analysis is to divide a set of observations into statistical clusters based on their similarity. Spatial cluster analysis aims to divide a set of observations into statistical clusters based on their similarity while at the same time grouping them into spatially continuous regions. In this sense, spatial cluster analysis is a mathematical expression of the traditional geographical regionalization method.

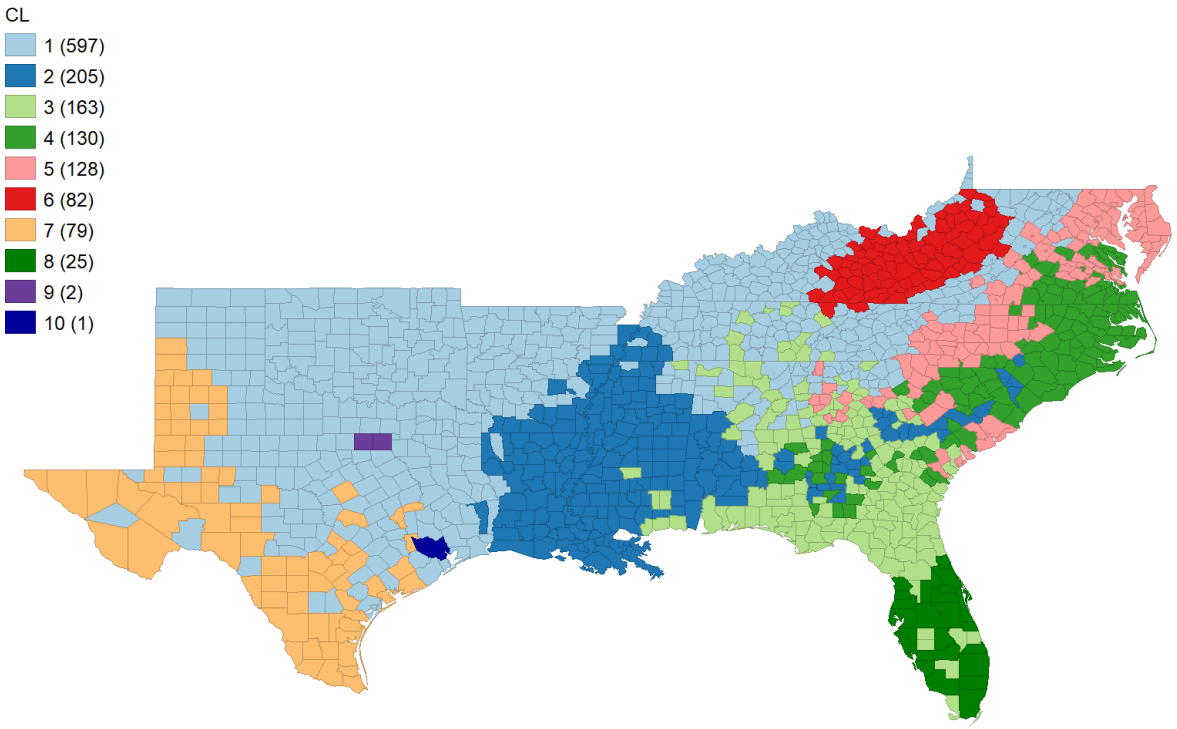
This means we are dealing with zoning – the combining of multidimensional data into statistical clusters while respecting spatial constraints. The resulting clusters (regions) should: 1) be as different from each other as possible; 2) contain elements that are as similar to each other as possible; and 3) include objects that are geographically located next to each other. Spatially bounded clustering can be soft (when geographic coordinates are introduced into the feature set or, as shown above, the weight of geographic centroids in the feature set changes), or hard (when it is impossible to create a cluster that combines observations of entities that are not neighbours). For the latter, the program needs to know which neighbourhood principle to use in the analysis, so you need to specify the spatial neighbourhood weights you want.

GeoDa includes a number of spatial cluster analysis methods:

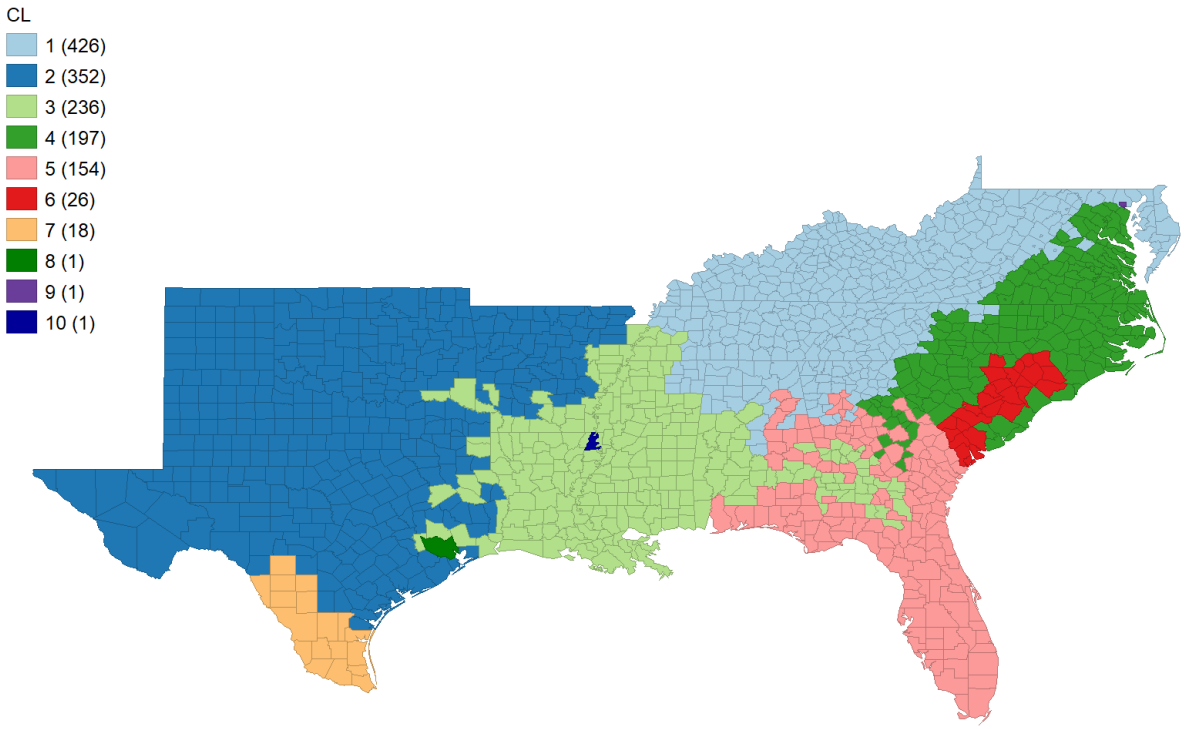
1. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
2. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)
3. Spatially Constrained Hierarchical Clustering (SCHC)
4. Spatial ‘K’luster Analysis by Tree Edge Removal (SKATER)

5. Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP)
6. Automatic Zoning Procedure (AZP)
7. Max-p Regions Model (max-p).

The first two are used exclusively to analyse point objects (or centroids of polygons) based on the density of their placement and the identification of noise points (outliers). Methods 3–5 are used to analyse all objects, and the algorithm assumes their hierarchical clustering with the spatial restraint that only neighbouring observations can fall into a single cluster. The REDCAP model is preferable, as it combines the procedures contained in the SCHS and SKATER models in its algorithm. Finally, methods 6 and 7 are not based on hierarchical clustering, so instead of looking for similarities between observations (with spatial constraints), they carry out zoning based on differences between regions. Note that the max-p model is the only one where you do not need to specify the desired number of clusters, only the maximum number of observations in a given cluster, which apparently makes it the most progressive for research in the social sciences.



REDCAP Analysis



AZP Analysis