Project 3: Clustering Algorithms and Development of a Kmeans/Jenks Algorithm

Abstract

This project seeks to unearth patterns in the tract population density data through the use of R's clustering algorithms. In addition to the use of in-built functionality, this project will detail the development of a "home-made" kmeans algorithm.

1. Initializing the Sf Objects and Setting the Stage for Analysis.

As with any project, the first step in this process was bringing in the data. The sf package made this extremely easy to do. The primary focus for this project was that of the population density. Calculating the block density required two steps. The first was to generate the area using the sf's st_area function and the units package set_units function to calculate the area for each polygon in square kilometers. Finally, density could be calculated by dividing the population, which was identified in the tracts data as column DP0010001, by the previously calculated area. One point of interest was that the units from the shape area did carry over from the calculation and can be seen in the density column as 1/km^2.

2. Tmap Clustering.

Once the population density was calculated, mapping by different clustering algorithms was extremely easy using tmap. The difficulty for mapping proved to be the formatting. Most of the formatting required to develop the maps shared in this report were done in the tm_layout function. The tm_layout function provides a great deal of control over the final product with everything from margin settings, positioning of elements, to setting titles or labels. In order to set the different clustering methods, tmap conveniently provides a style argument. The methods used in the following figures are: Kmeans, Jenks, Hierarchical Clustering, Bayesian Clustering, Quantiles, and Fisher. The kmeans and Jenks methods in Figure 1 were almost identical, which makes intuitive sense as they are very comparable algorithms. The Quantiles method in Figure 2 seemed to show the greatest amount of color on the map.

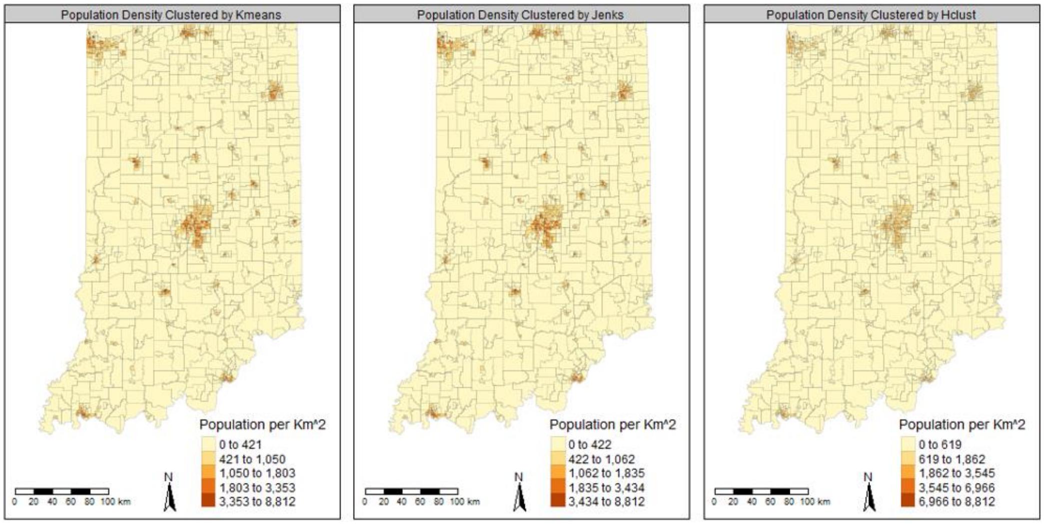

Figure 1: Kmeans, Jenks, and Hierarchical clustering methods shown for population density in each Census tract in Indiana.

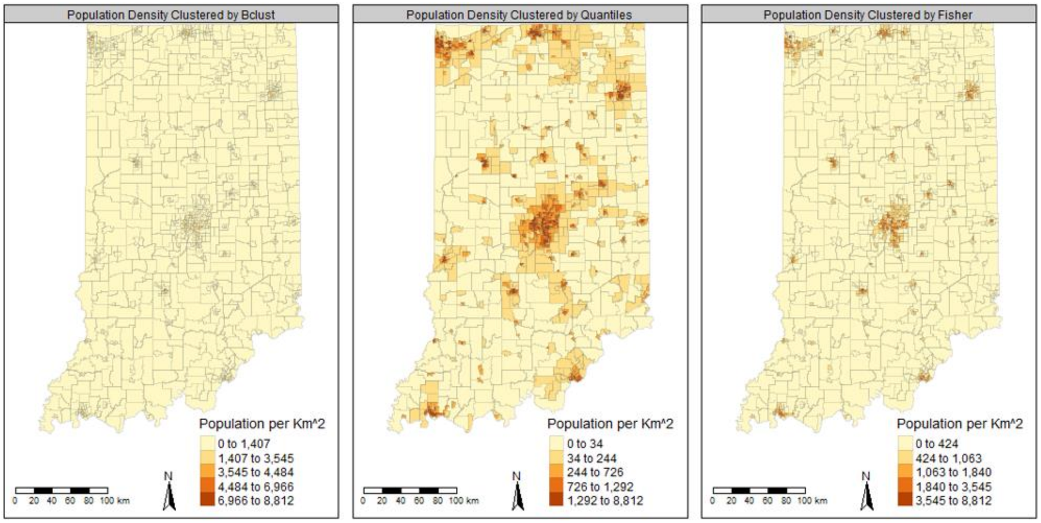

Figure 2: Bayesian, Quantiles, and Fisher clustering methods shown for population density in each Census tract in Indiana.

3. Development of Kmeans with Jenks Criterion.

The point of Kmeans with the Jenks criterion is to maximize the value for F, as shown in the Equation 1. This equation was sourced form the lecture notes provided by Prof. Shan.

Equation 1

$$F = \frac{Between\ Class\ Variance}{Within\ Class\ Variance}$$

The between class variance was simply defined as the variance of the center points for the k classes and the within class variance was the sum of the variance for all values in each class separately, summed up. Using this method, I was able to produce a very comparable classification system to that of the built in Kmeans/Jenks methods but not exactly. The maximum F-value selected by the algorithm developed here can be seen Figure 3 along with the tmap.

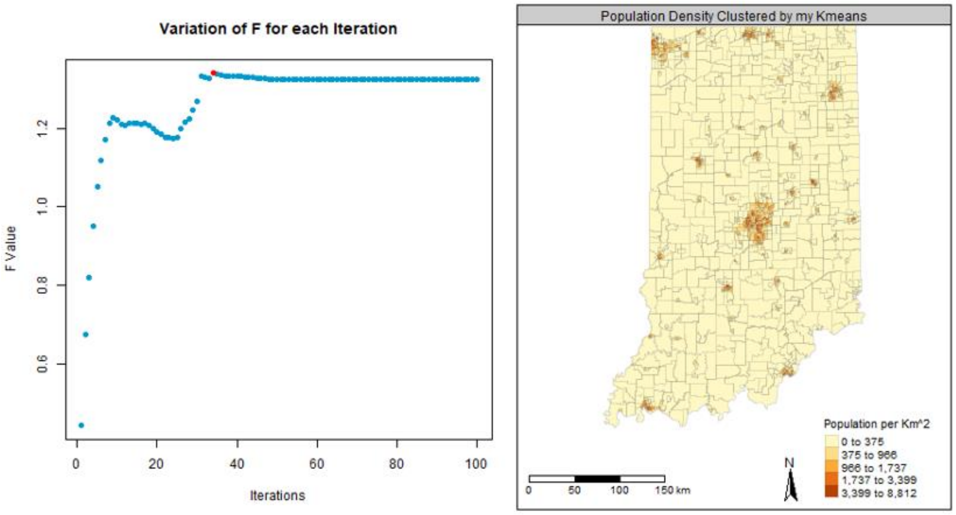

Figure 3: F-values for Kmeans/Jenks calculations along with the breaks mapped to Indiana tracts.

When compared with ArcGIS, the values were again very close but not the same as what I had calculated or what R calculated. The method by which R, myself, and ArcGIS terminates calculation of the F value must be different. The histogram and map can be seen in Figure 4.

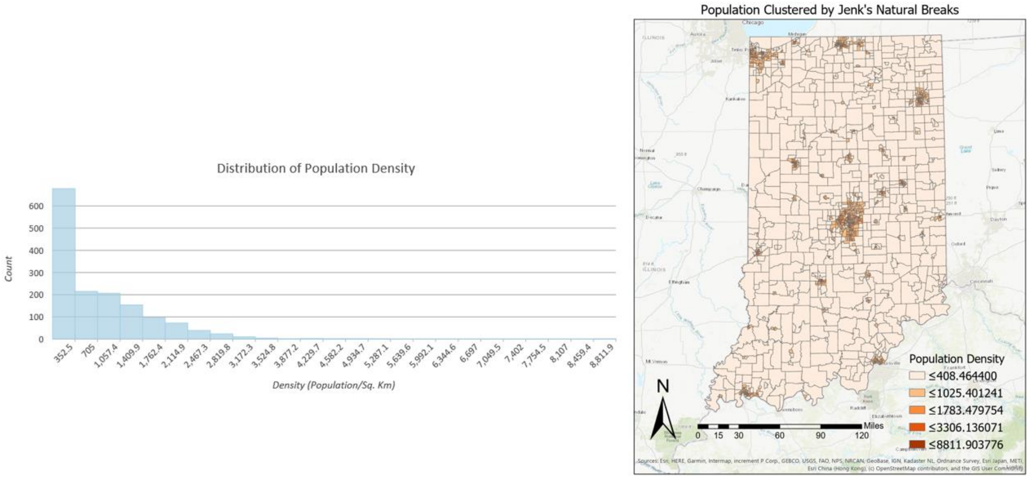

Figure 4: Population density histogram and Jenks method with natural breaks created in ArcGIS.

4. Summary/Conclusion/Concluding Remarks

This project proved to be challenging from a computational standpoint but was quite rewarding to develop a well known and quite powerful clustering algorithm. It should prove to be useful to have an intuitive sense of how Kmeans/Jenks works going forward.