HW5. K-Means Clustering

Yuhao Jia, Zhonghua Yang, Zile Wu

1. Introduction

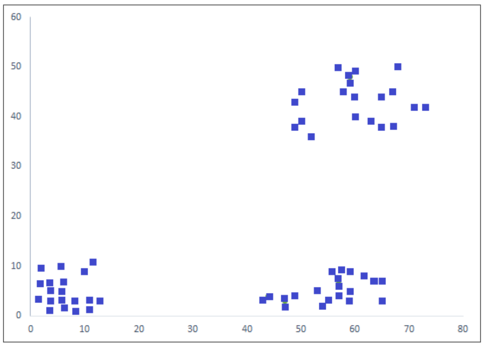
Cluster analysis is a technique based on the partitioning of data, which aims to divide data into different clusters, and data in the same clusters often have a strong degree of similarity. Cluster analysis is often suitable for dealing with large amounts of data, and dividing the data into relatively small groups is more conducive to handling and studying the data.

For this assignment we will regress the data from Assignment 1 and Assignment 2 and use the K-means algorithm to divide the data set with a large volume of data into non-overlapping clusters that contain all the data, with each individual object in a particular set. Our dataset has five variables, they are PCTBACHMOR, MEDHVAL, MEDHHINC, PCTVACANT, PCTSINGLES. we need to divide these five types of variables into numerous clusters, we need not only to find the optimal number of this cluster, but also to study the autocorrelation situation between variables in the same cluster.

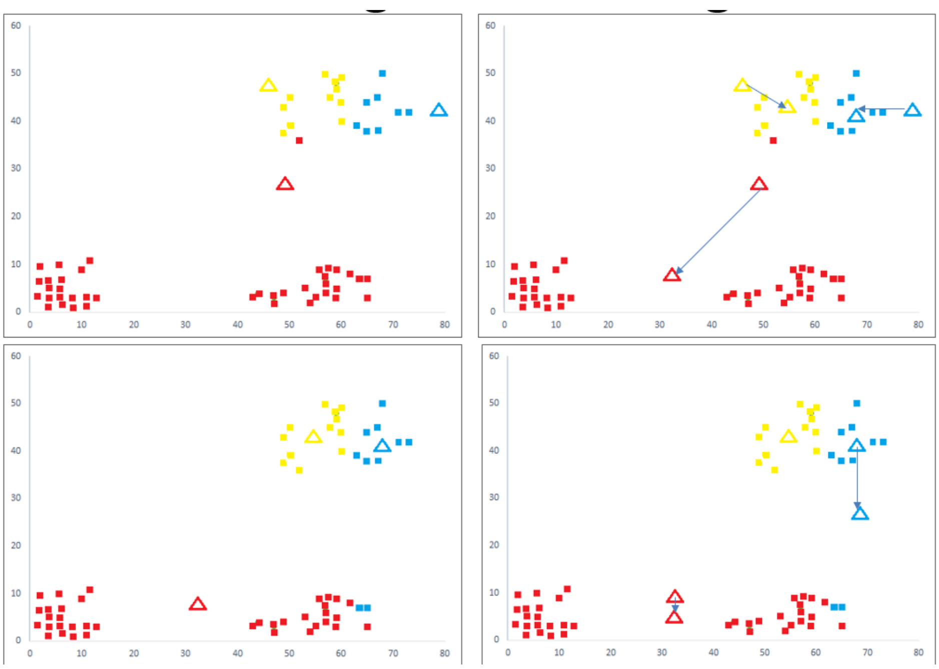
Using K-means cluster analysis we can answer questions such as: how do we classify houses in different situations in the neighborhood, how many clusters are more appropriate for this dataset, and what are the characteristics between houses in the dataset? In addition, adding spatial characteristics to K-means cluster analysis allows us to bring in data from specific spatial regions and explore whether the variables in these data are also spatially clustered? What real-world problems do these spatial correlations reflect?

2. Methods

The principle of the K-means clustering algorithm is not difficult to explain. Suppose we need to group classes on a dataset, the number of classes being artificially specified. For example, in the dataset shown in the figure, 60 points in the coordinates are to be grouped based on two variables (y=latitude and x=longitude), and a total of k=3 groups are specified.



Our method is to first assume the location of the centroids of the three class clusters subsequently, and then find the nearest cluster center for each data point based on the Euclidean distance, after the centroid of mass of each cluster is calculated, and the data points are assigned again by using the centroid of mass as the new cluster center. This calculation is repeated until the allocation of data doesn't change anymore, and the optimal answer is successfully found.



We also have a metric to evaluate the clustering of the data called within-cluster sum of squared errors (SSE). SSE is specifically the calculation of the squared distance between each observation and the centroid of its cluster and the sum of these squared distances. The SSE for overall K-means is calculated by summing these SSEs across clusters.

Of course, in the practical application of K-means clustering algorithm, the variables do not necessarily have to be spatially meaningful latitude and longitude, for example, education level, income level, etc. can be used as the object of calculation. But these variables should be continuous.

Limitations of the K-means clustering algorithm：

1. The K-means algorithm requires the priori number of clusters, which is not easy to determine in advance, although some methods can be found, such as the SSE scree plot.

2. K-means clustering is suitable for continuous numeric variables, although some people also bring binary data sets into the calculation.

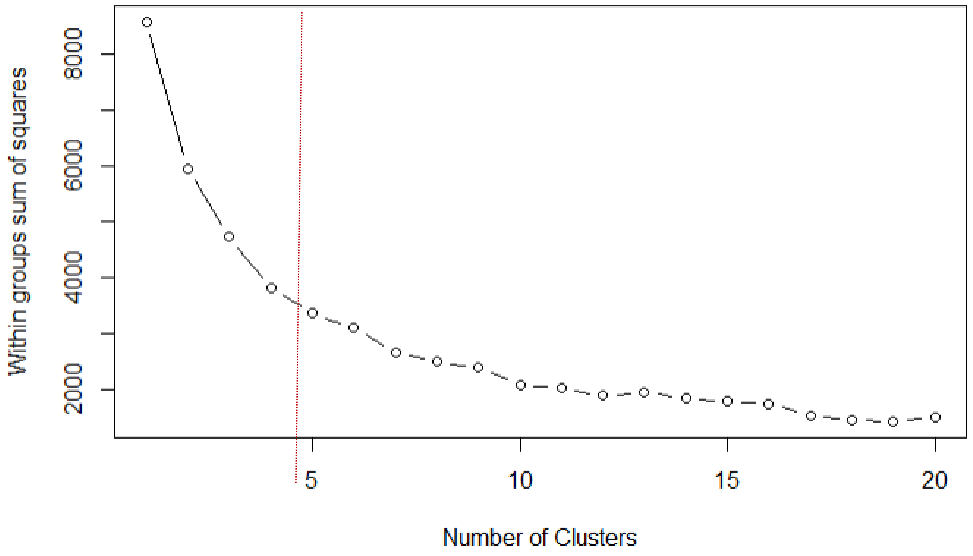
3. Some problems may arise when the size of the clusters is not consistent. In the K-means clustering algorithm it is assumed that the clusters are all of the same size, and similarly, it is assumed that the density of the data in the clusters is the same, so that the clusters are globular. However, in practice, the calculation can be biased, resulting in clusters of non-standard size, density and shape.

4. K-means clustering is difficult to handle noisy and anomalous data because each data must be classified into a certain cluster in this algorithm.

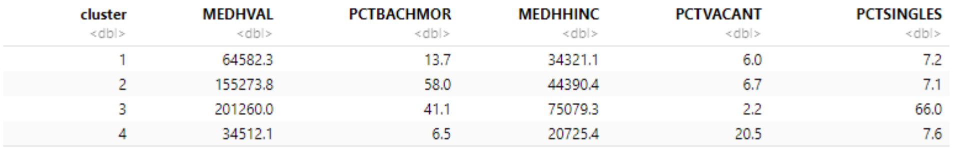
5. K-means clustering looks for local minima rather than global minima, which can lead to incorrect results.

Besides K-means clustering algorithm, there are other clustering algorithms, including hierarchical clustering and density-based clustering algorithm (DBSCAN). But hierarchical clustering is more suitable for exploring small data sets, while DBSCAN, although it can identify different cluster shapes, it cannot include all data in the clusters because observations that are too far from the nearest domain will be identified as outliers. Therefore, in comparison, the most suitable method is still K-means clustering, because our dataset is large and not suitable for hierarchical aggregation, and also our dataset does not discard data, so DBSCAN is not suitable either.

3. **Results**



Using the NbClust in R, we can generate a scree plot to determine the optimal number of clusters. The x-axis of the scree plot is the number of clusters. The y-axis is the Sum of Squares Error (SSE) for each number of clusters. In the scree plot, the drop in SSE between 1 and 2 clusters is vast, the drop between 2 and 3 clusters is pretty substantial, the drop between 3 and 4 clusters is somewhat noticeable. After 4, there is relative little drop in SSE between two adjacent number of clusters. So we stick with the 4 cluster solution based on the scree plot.



cluster	MEDHVAL	PCTBACHMOR	MEDHHINC	PCTVACANT	PCTSINGLES
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	64582.3	13.7	34321.1	6.0	7.2
2	155273.8	58.0	44390.4	6.7	7.1
3	201260.0	41.1	75079.3	2.2	66.0
4	34512.1	6.5	20725.4	20.5	7.6

After determining the number of clusters as 4, we can calculate the mean values of the MEDHVAL, MEDHHINC, PCTBACHMOR, PCTSINGLES, and PCTVACANT in each cluster. The result seems reasonable. For example, the cluster 4 describes which census tracts with lower household incomes and lower housing values have both lower bachelor's degree populations and higher housing vacancy rates. These descriptions are consistent with our perception of census tracts that are in decline, which similar to some neighborhoods in North Philadelphia. The cluster 3 has the highest home values, household income, percent of single/detached housing units and lowest home vacancy rates. This is also consistent with our impression of Northwest Philadelphia, which is a developed area with a high-income population. In summary, based on the data in the table, we can name these clusters as follows:

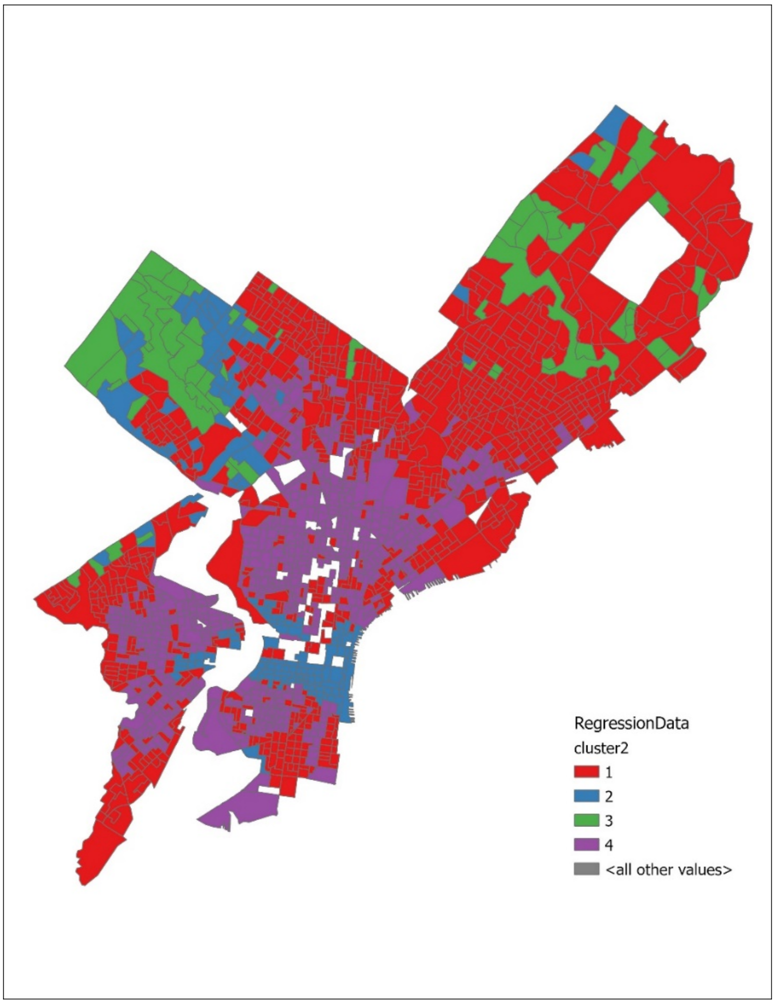
Cluster 1: Low to middle income cluster.

Cluster 2: Middle to high income cluster.

Cluster 3: Wealthy and developed cluster.

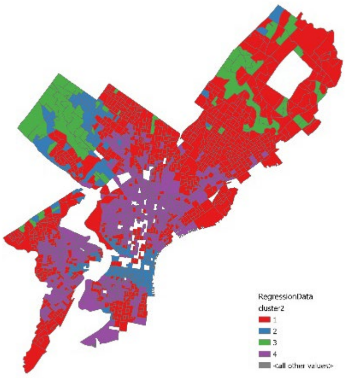
Cluster 4: Poor and less developed cluster.

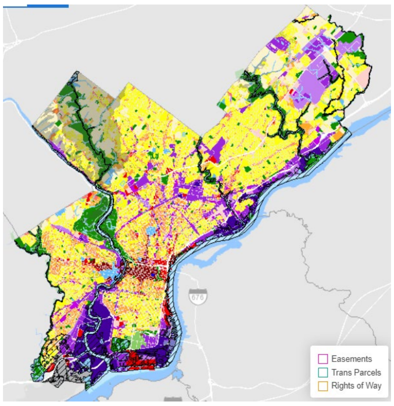
Below is a map showing the spatial distribution of 4 clusters. There are clear patterns showing spatial autocorrelation of 4 clusters. Cluster 2 tends to spatially cluster in center city, university city and north-west Philadelphia. Cluster 4 tends to spatially cluster in west, south and north Philadelphia. Most of tracts in cluster 1 tend to spatially cluster in north and north-east Philadelphia, while there are some tracts clustering in the margin of south and west Philadelphia. Cluster 3 tends to spatially cluster in the margin of north-west Philadelphia and middle of north-east Philadelphia.



The spatial pattern of 4 clusters is somewhat similar to the pattern of concentric rings of land use in “Bid rent theory” which describes the change of land price and demand as the distance from the central business district (CBD) increases. We can add some key words to those clusters’ name. Though it’s not entirely appropriate to add “CBD” to cluster 2 as it has some tracts in north-west Philadelphia. It might be reasonable to add “high density residential (the ring near CBD)” to cluster 4 and “low density residential (the ring far from CBD)” to cluster 1.

We also compare the cluster map with the zoning of Philadelphia as follows. We found that most tracts in cluster 3 surrounds green lands or water. So we can add words like “nature around” to its name.





4. Discussion

The spatial pattern of 4 clusters is similar to the pattern of concentric rings of land use in “Bid rent theory” which is not so surprising because we include home value and household income as clustering features. However, it’s hard to use this theory to explain why some tracts in north-east Philadelphia are in the same cluster with CBD area. It’s reasonable that tracts in cluster 3 surround green land and water because nature landscape may increase the house value, but we didn’t expect the high percent of single in this area.