Homework 5: K-Means Clustering

CPLN 671/MUSA 501

Remember the data we used for assignments 1 and 2? The data set contained several variables for 1720 block groups in Philadelphia. These variables included:

1. **MEDHVAL**: median house value
2. **MEDHHINC**: median household income
3. **PCTBACHMOR**: percent of individuals with at least a bachelor's degree
4. **PCTSINGLES**: percent of single/detached housing units
5. **PCTVACANT**: percent of vacant housing units

For this assignment, you will be asked to use the R code provided in the slides to run a k-means cluster analysis with the 5 variables above.

Specifically, you will be asked to:

- Identify the optimal number of clusters based on the scree plot and the 26+ diagnostics available in the NbClust package in R.
- Run the k-means cluster analysis using the optimal number of clusters (i.e., the number of clusters that's identified as the best by the largest number of diagnostics in NbClust).
- As shown in the slides, use the aggregate command in R to examine the mean values of the variables in each of the resulting clusters.
- Specify whether the cluster solution makes sense. If it does, attempt to come up with descriptive names for each of the resulting clusters (see the slide called "**Do Variables Have To Be Lat & Lon?"** for an example of creative and descriptive names for clusters).
- Note: be sure you use the scale command to standardize the variables before running the cluster analysis, as is done in the code shown in class.

Outline

Introduction

1. Describe the data set and indicate what the purpose of this assignment is. That is, how can k-means clustering help you look at the data, and what kinds of questions can you answer?

Methods

1. How does the K-means algorithm work? Describe the steps in your own words.
2. What are some of the limitations of the algorithm?

3. What are some other clustering algorithms, and might they be more appropriate here?

Results

1. Present and describe the results from the NbClust command and the scree plot.
    a. What's the optimal number of clusters that you should choose based on the output?
2. Present and describe the table produced by the aggregate command showing the mean values of the **MEDHVAL**, **MEDHHINC**, **PCTBACHMOR**, **PCTSINGLES**, and **PCTVACANT** in each cluster.
    a. Here, state whether the cluster solution makes sense and if so, come up with descriptive names for each of the resulting clusters.
3. Look up the syntax for the write.csv command and export the table containing the cluster ID of each observation into a .csv file (i.e., the .csv file should contain the variable that indicates which K-means cluster each observation falls into).
4. In ArcGIS, join the .csv file that you exported from R to the RegressionData.shp by the field **POLY_ID**, and create a map showing the spatial distribution of the clusters.
    a. State whether observations falling into the same K-means clusters also tend to cluster in space. That is, does the K-means cluster membership variable seem to be spatially autocorrelated?
        i. You should not use Moran's I here, because the cluster membership variable is categorical.
    b. Does looking at the map yield any additional insight into the patterns you observe with the K-means analysis? Does it have any impact on how you might name the clusters?

Discussion

1. Briefly describe any patterns that you observe. Any surprising findings?