Cpln HW 4

Ajjit Narayanan

Bill Cohen

Introduction

With both spatial and non-spatial data, it is useful to group subsets of a population by shared characteristics. When those shared characteristics are known and defined, many useful statistical regression modeling techniques can be used to analyze the data. Spatial proximity can be one such grouping, but data can also be better understood by grouping non-spatial variables such as home value, income, education, and property type. However class membership is often unknown in advance.

Cluster analysis can be used to identify groups of data with similar traits within a sample population when those class distinctions do not exist a priori. For this analysis, we have a dataset of all the block groups in Philadelphia. For each block group, we have the following variables: median house value (MEDHVAL), median household income (MEDHHINC), percent of individuals with at least a bachelor's degree (PCTBACHMOR), percent of single/detached housing units (PCTSINGLES), percent of vacant housing units (PCTVACANT). The goal of this paper is to find clusters of block groups with similar characteristics. The method that we will be using is k-means clustering, which takes a user specified number of clusters and groups each observation into a cluster. It can help us understand the various kinds of block groups that exist in Philadelphia, see how the different variables are distributed across all block groups, and look at relationships between variables within the cluster.

Methods

As stated above, we will use the k-means clustering algorithm to find out if there are discrete clusters of block groups in Philly. The k-means algorithm groups interval (numerical) data into a user-specified number of clusters, such that each observation belongs to exactly one cluster. To do this, the K-means algorithm uses a 6 step iterative process.

1) Randomly selects k points as cluster centers within the n dimensional space of our data, where n is the number of variables we are clustering the data on.
2) Calculate the (Euclidean) distance between each data point and each of the randomly assigned cluster centers
3) Assign each data point to the cluster center that it is closest to

4) Using the newly calculated clusters of data points, recalculate the cluster centers
5) Update the distance between each data point and the newly calculated cluster center
6) If no observation changes membership, the process concludes. If not, repeat from step 3

The underlying objective is to locate the cluster centroids such that they minimize the overall distance to cluster member observations. This is calculated as the sum of squared errors (SSE), or the sum of squared distances between each observation in a cluster and the cluster centroid: $SSE = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$

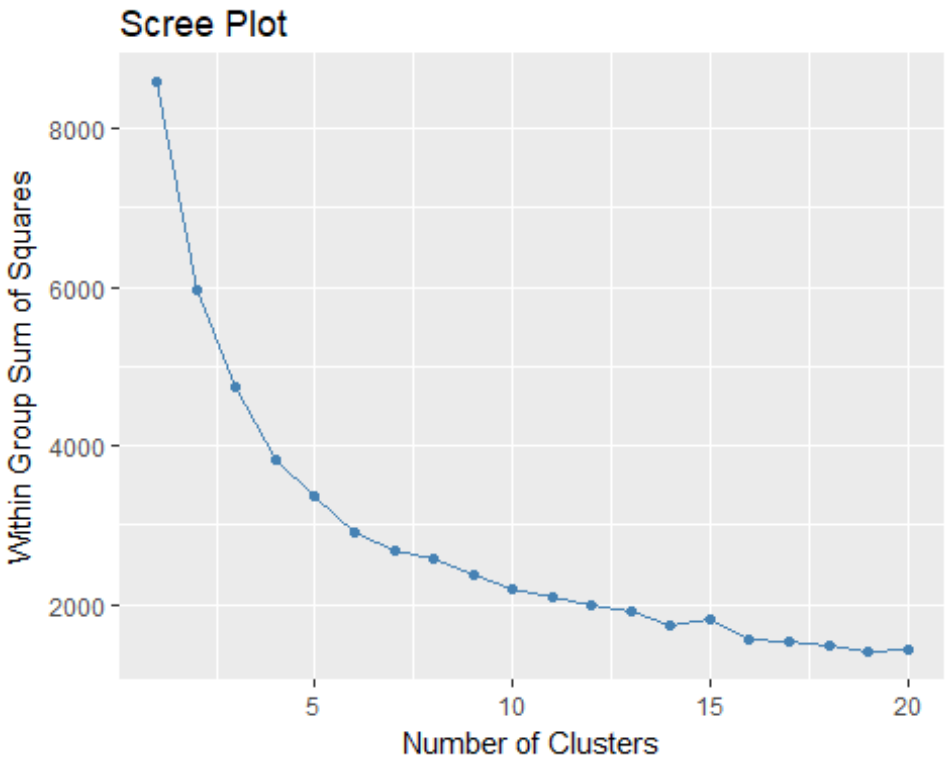
where x_i and y_i are the coordinates of observation i, and x_c and y_c are the coordinates of the nearest centroid.

As with other cluster analysis methods, K-means clustering seeks to separate objects such that the resulting groups are easily interpreted and meaningfully actionable. The biggest factor in the success of the k-means clustering method is the number of clusters chosen. When classes are unknown, it is often difficult to know how many clusters to use. Because this is user-defined, there is great room for influencing the results and this is in fact one of the limitations of the k-means algorithm. There are a number of tests and indices that calculate an optimal number of clusters for K-means, such as the Hubert index and the D index. Some other limitations with k-means clustering include issues dealing with noise and outliers, groups of different sizes and densities, and those with non-globular shapes.

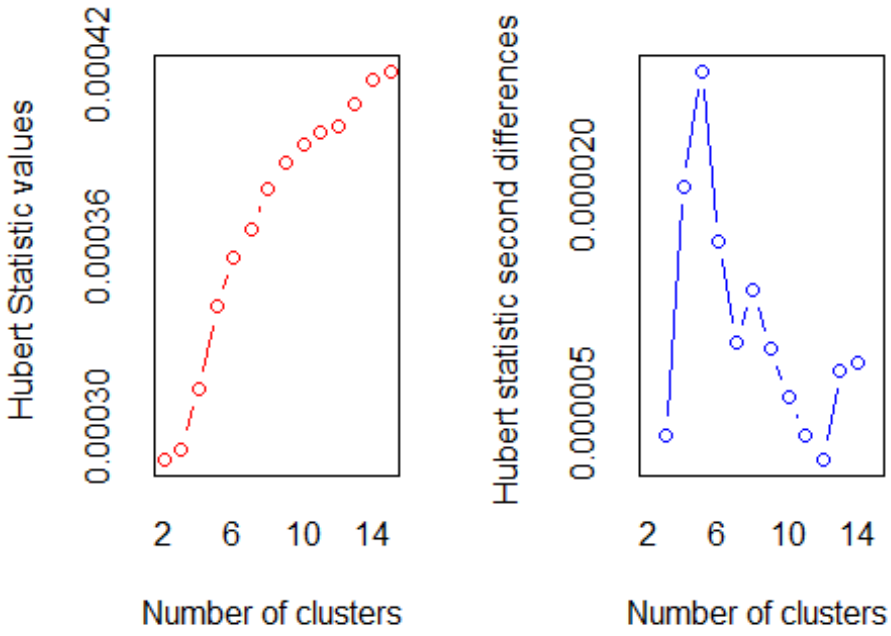
Hierarchical clustering can offer an alternative for smaller datasets, working from the bottom up to group data into cluster hierarchies. Density-based clustering (DBSCAN) groups points by establishing a meaningful neighborhood size and minimum number of neighbors.

Results

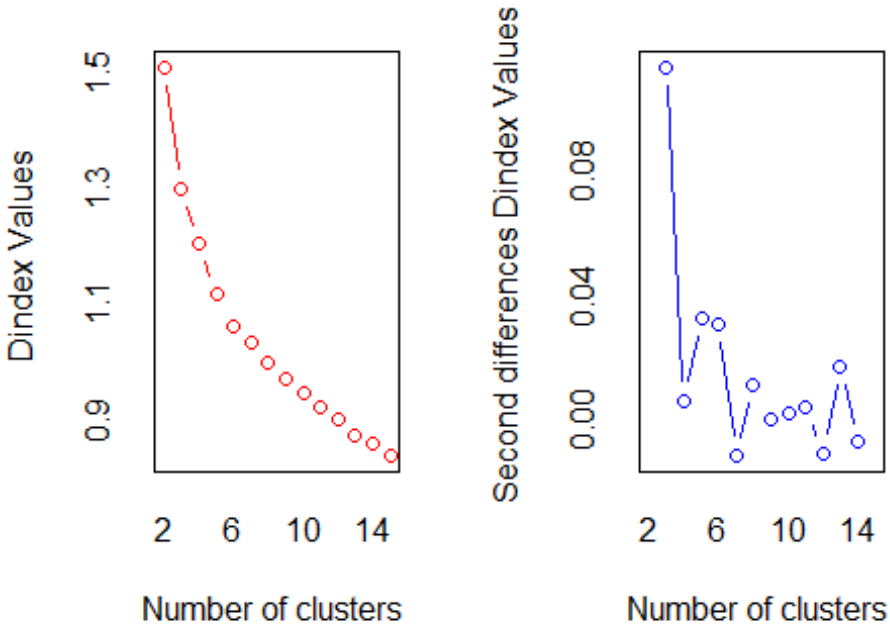
Below is the the results from the Scree plot, which is a graph of the number of clusters vs the within group sum of squares. An appropriate cluster solution could be defined as the solution at which the reduction in SSE slows dramatically. This produces an “elbow” in the Scree plot.



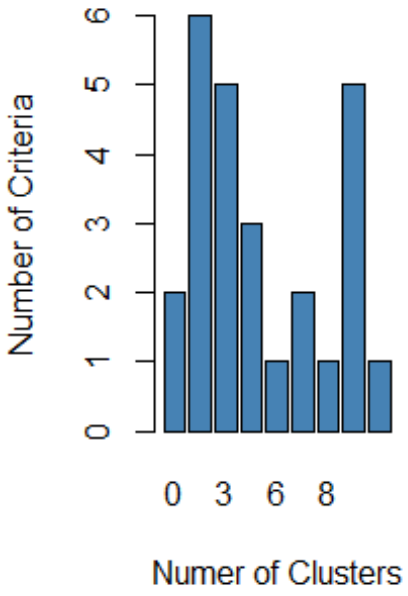
The figure is fairly inconclusive, without a distinctive elbow. It appears however that the greatest decline in slope between clusters 3 and 5, suggesting 4 clusters should be used. Next we use the NbClust package in R, which has 30 different methods to determine the optimal number of clusters. Below is a barplot of the number of clusters and the number of criteria that choose the cluster size as optimal.




```
## *** : The Hubert index is a graphical method of determining the number of
clusters.
##                 In the plot of Hubert index, we seek a significant knee
that corresponds to a
##                 significant increase of the value of the measure i.e the
significant peak in Hubert
##                 index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of
clusters.
##                 In the plot of D index, we seek a significant knee (the
significant peak in Dindex
##                 second differences plot) that corresponds to a significant
increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 6 proposed 2 as the best number of clusters
## * 5 proposed 3 as the best number of clusters
## * 3 proposed 4 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
## * 1 proposed 8 as the best number of clusters
## * 5 proposed 13 as the best number of clusters
## * 1 proposed 15 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```



We can see from the output of the NbClust plot that approximately 6 of the methods choose an optimal cluster size of 2. This is the cluster size with the highest number of criteria, but clusters of size 3 and size 13 are also noticeably large. For now, we proceed with using a cluster size of 2, as that had the most number of criteria to back it up. Below are the results of running a 2-mean clustering algorithm on our data. We present the sizes of the clusters along with the standardized and unstandardized (ie actual) means of each cluster

cluster	size
1	1446
2	274

Standardized Cluster Means

cluster	MEDHVAL	PCTBACHMOR	MEDHHINC	PCTVACANT	PCTSINGLES
1	-0.27	-0.33	-0.24	0.13	-0.19
2	1.44	1.74	1.25	-0.67	0.99

Actual Cluster Means

cluster	MEDHVAL	PCTBACHMOR	MEDHHINC	PCTVACANT	PCTSINGLES

1	49952.4	10.2	27668.5	12.5	6.8
2	152495.3	46.9	51982.6	4.8	22.3

So it seems as if the model grouped the majority of the data points into cluster 1, and a small amount into cluster 2. Broadly, we can define cluster 1 as "working class" and cluster 2 as "hardly working class". The working class cluster, which is most of the block groups in Philly, have lower Median Household values of around $50,000, have only 10% of the population with a bachelors degree or higher, have household incomes of around $27,668, have 12% of lots that are vacant, and only have 6.8% of houses with singles. In contrast, the hardly working cluster has Median Household values of around $152,000, have 47% of the population with a bachelors degree or higher, have household incomes of around $51,928, have around 4.8% of lots that are vacant, and have 22.3% of houses with singles. Although basic, these clusters do make sense in the context of Philadelphia, which is one of the poorest large cities in the United States. For completeness sake, we also run k-means again with cluster sizes of 3 and 13, they were the runner up cluster sizes based on the criteria. We present the cluster sizes and actual cluster means for 3 and 13 clusters respectively.

cluster	size
1	859
2	156
3	705

cluster	size
1	257
2	15
3	86
4	2
5	188
6	113
7	30
8	76
9	258
10	364
11	29
12	254
13	48

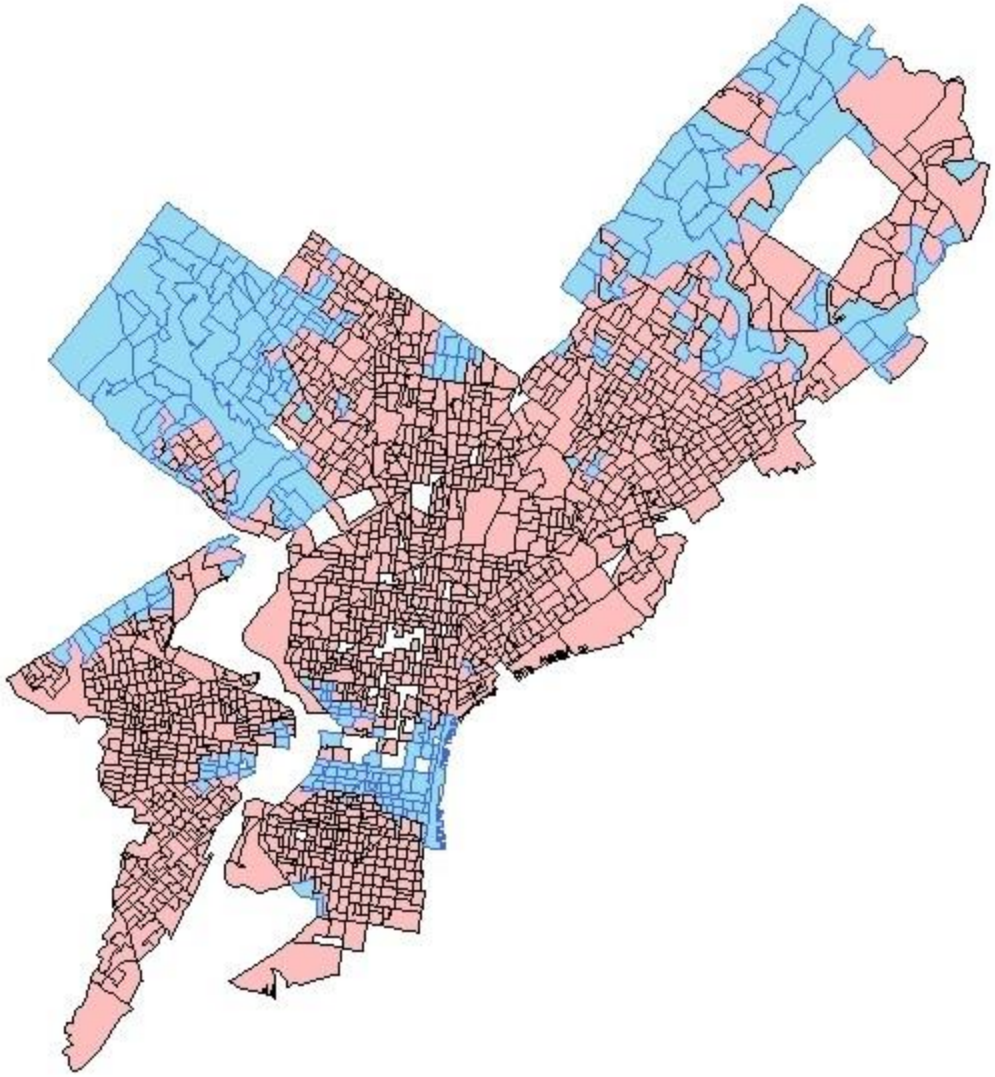
Actual Cluster Means

cluster	MEDHVAL	PCTBACHMOR	MEDHHINC	PCTVACANT	PCTSINGLES
1	69041.3	16.8	35371.8	5.8	7.0
2	191203.9	55.5	58111.0	4.7	29.5
3	35291.6	6.5	20995.9	19.4	7.5

cluster	MEDHVAL	PCTBACHMOR	MEDHHINC	PCTVACANT	PCTSINGLES
1	43149.8	9.4	28845.0	15.7	6.1
2	274033.3	73.2	106624.0	2.5	65.9
3	95988.4	23.7	45426.1	4.1	29.5
4	921900.5	60.1	200001.0	2.9	65.1
5	80452.7	18.3	46799.8	3.7	5.4
6	84346.9	36.3	33570.1	9.3	5.3
7	295753.3	71.1	51458.5	6.9	5.2
8	141075.0	62.9	42204.8	6.2	5.9
9	29445.7	4.9	18242.6	23.3	8.1
10	58976.6	10.1	32443.9	5.6	5.1
11	131289.7	27.6	57003.8	1.5	72.3
12	38362.2	5.8	17493.3	9.3	8.0
13	38533.3	7.0	20353.7	43.8	8.0

The 3 cluster analysis is very similar to the 2 cluster analysis, but now there is a low income group, a middle income group, and a high income group. The low income and high income groups are more extreme than the groups we had under 2 clusters. The 13 cluster output is a little harder to read, but it basically just shows even more gradation. i

Now we want to analyze the spatial distribution of clusters for our original 2 cluster analysis. To do this, we imported the data into ArcMap and then generate choropleth maps. The pink zones represent block groups in cluster 1 and the blue zones represents block groups in cluster 2.



There is clear spatial autocorrelation within clusters, with cluster 2 primarily located in the northeast and northwest, with an additional pocket in Center City/University City. Based on Philadelphia census data used in previous assignments, we know that these areas tend to have high median household income, high median home value, and higher levels of college educated residents. Another accurate name for these clusters could be "The 1%" (cluster 2) and the "99%" (cluster 1).

Discussion

In conclusion, we tried to find clusters of Philadelphia block groups based on several socioeconomic characteristics. Using a k-means algorithm, we first identified 2 as the optimal number of clusters to split our data into. We then saw that the 2 clusters in our data roughly corresponded to "The 1%" (cluster 2) and "The 99%" (cluster 1).

The 1% tended to have higher Median Household Values, higher Median Household Incomes, lower vacancy rates, higher education rates, and a higher amount of singles. This is mostly in line with our intuition as it makes sense that the richer parts of the city also have higher levels of education, lower rates of vacancy and higher amount of single households.