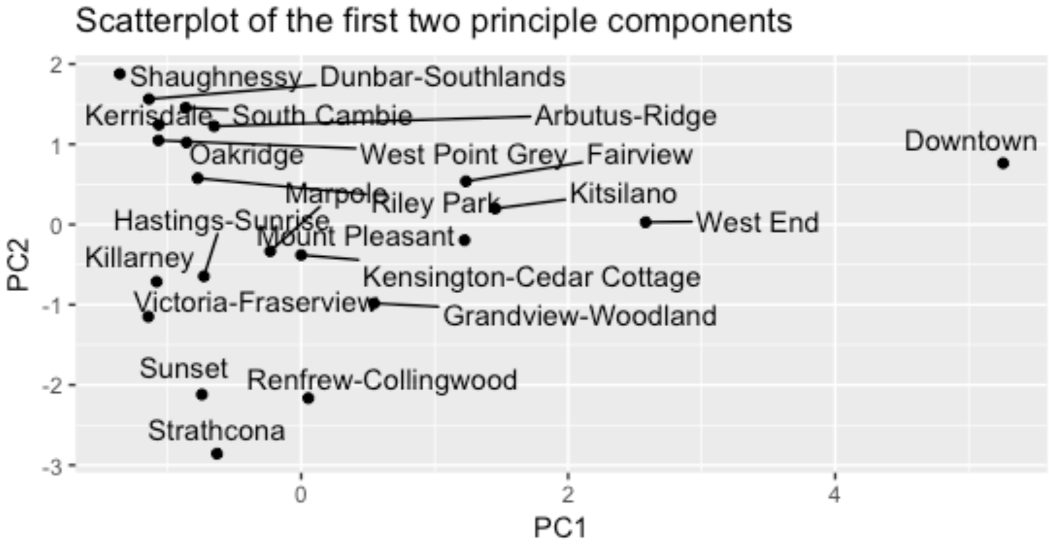
Cluster Analysis

Huanye Liu

The main purpose of this report is to illustrate and compare spatial clustering pattern using different approaches for several variables of interest in the city of Vancouver, Canada. The city of Vancouver is divided into 22 communities with different numbers of neighboring communities. Here we want to examine the relationship between neighbors in multivariate space of interest and the neighbors in geography.

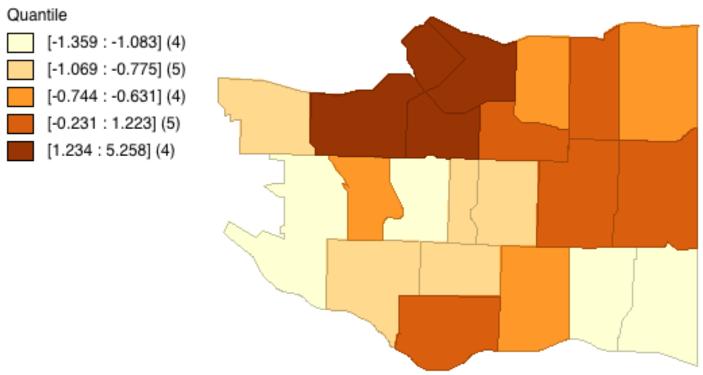
1. Principle components analysis

The five variables of interest are number of dwellings, average rent, unemployment rate, number of crimes and mobility level for each community in Vancouver. We want to first reduce the five variables to two principle components by extracting and combing common features among these variables. After applying principle component analysis using R, we find from the loading matrix that the first principle component mainly represents the combination of mobility level, number of dwellings and number of crimes, which can be seen as a measure of public security. The second principle component mainly represents the combination of average rent and unemployment rate, which can be referred to as an economic index. The scatterplot of the first two principle components are shown below.

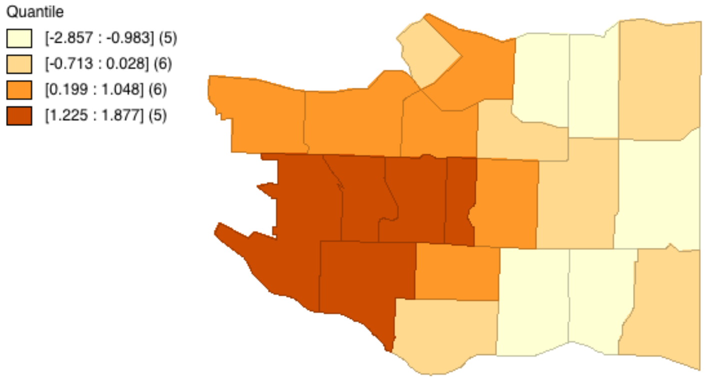


From the scatterplot we can find several loose clusters in multivariate space of interest, so we then use GeoDa to examine the spatial features in terms of the first and second principle components respectively, and two maps are shown as below.

The map using the first principle component:



The map using the second principle component:

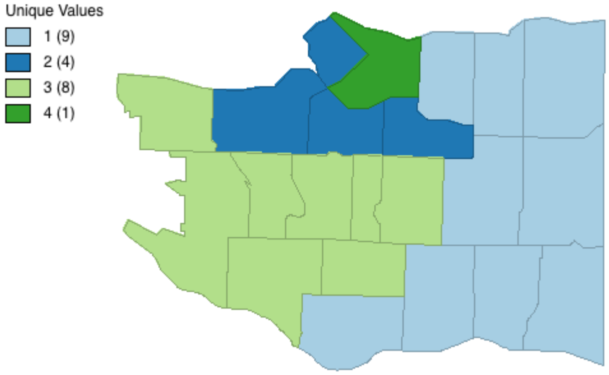


As we can see from the two maps above, the geographical neighboring patterns can be found in both maps, especially for communities taking lower values in these two principle components.

2. Classic Clustering Methods

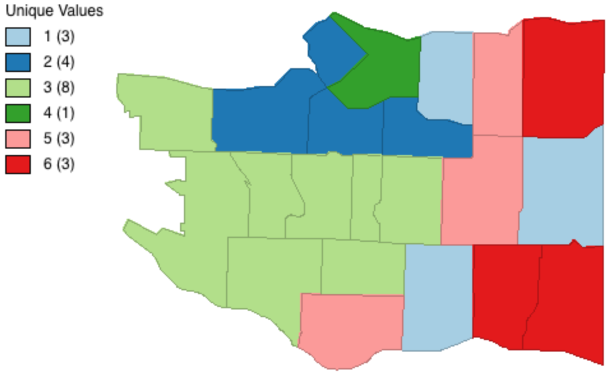
2.1 hierarchical clustering

Given the first two principle components, we first apply the hierarchical clustering to them using complete linkage. By choosing the number of clusters to be 4 and merging the group information with shape files using GeoDa, we examine the spatial features of the 4 clusters using the first two principle components, and the map is shown as below:



From the map above we can see the clustering based on the first two principle components also indicates the spatial clustering pattern, which illustrates the high local spatial correlation for the five variables of interest.

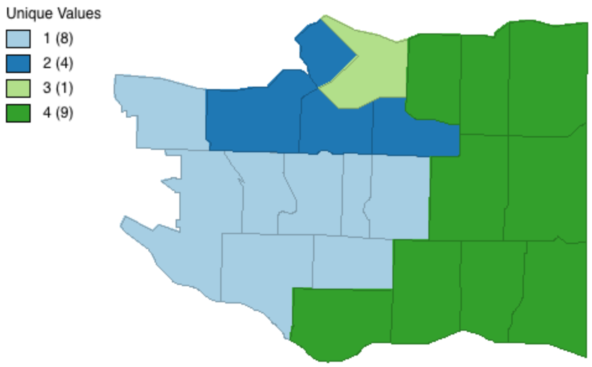
Then we choose the number of clusters to be 6. After following the same steps as k=4, we plot the map as shown below:



Compare the map above with the one using 4 clusters, we can see the spatial clustering pattern is not as good. There emerges some absences of spatial contiguity for 3 clusters.

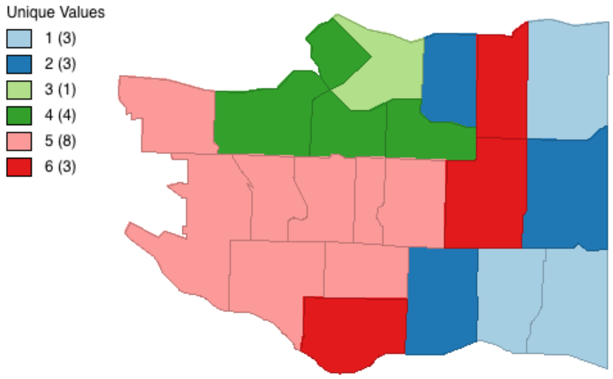
2.2 K-means clustering

Next we apply the K-means clustering to the first two principle components by choosing 4 clusters and 20 initial random assignments. Again, we merge the cluster information with the shape files using GeoDa and examine the spatial features of the 4 clusters using the first two principle components. The map is shown as below:



Comparing the map above with the one using hierarchical clustering, we find the two spatial clustering patterns are exactly the same, which means the effects using both clustering methods when k=4 are the same.

Then, again, we choose the number of clusters to be 6, and now the map is shown as below:

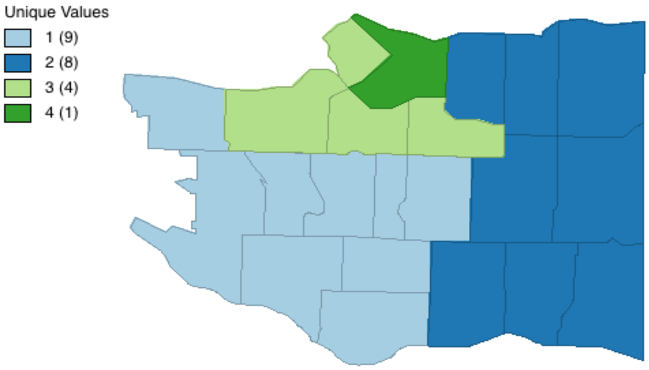


and again, we find the two spatial clustering patterns by applying two clustering methods to the first two principle components are exactly the same again, which means the effects using both clustering methods when k=6 are also the same.

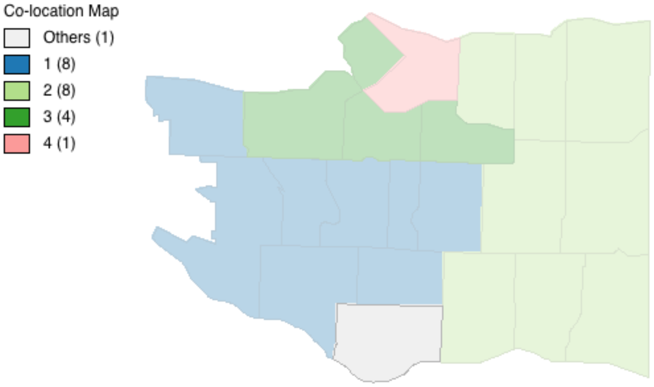
3. Spatially constrained clustering

3.1 Skater

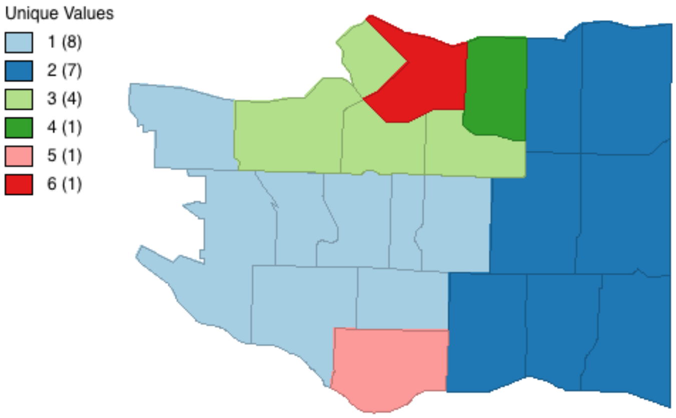
When considering contiguity constraints imposed on the multivariate clustering methods, we first use the skater algorithm. After selecting the first two principle components, choosing the number of clusters to be 4 and specifying the queen contiguity weight file, we run the algorithm and the resulting cluster map as shown below:

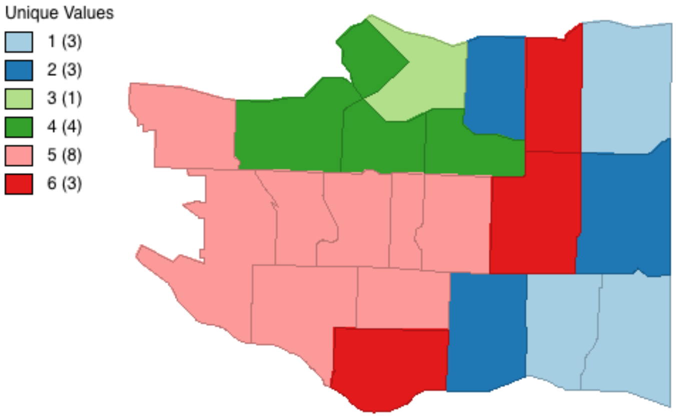


The ratio of between sum of squares to total sum of squares is 0.82601, and the clustering pattern on the map above and it is guaranteed to meet the spatial contiguous constraint, which is very similar to cluster maps using classical methods when k=4. The only difference is the community Marpole in the south of Vancouver, as the co-location map shows below:



Then we try to choose the number of clusters to be 6. After following the same steps as k=4, we get the cluster map as shown below:

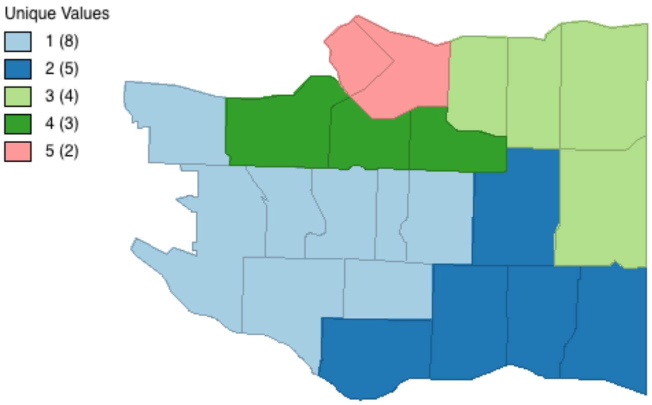




The ratio of between sum of squares to total sum of squares is 0.898. Compared with the lower cluster maps using classical methods when k=6, the upper cluster map using skater picks out the two communities disconnected to others from their respective clusters on the lower map to form two 1-community clusters, and combine all 7 communities on the right side which belongs to 3 different clusters on the lower map to form one big cluster, to meet the constraint of spatial contiguity.

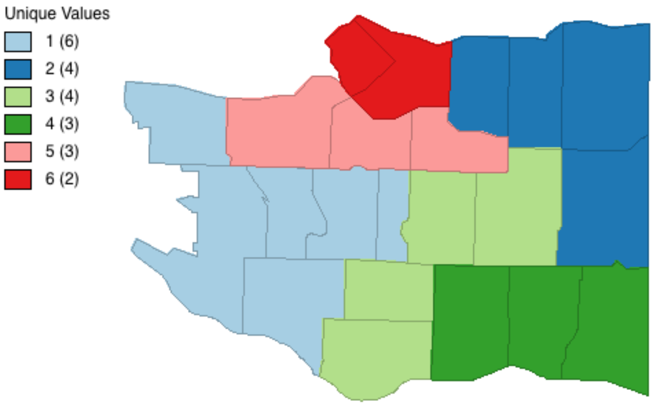
3.2 Max-p

Next we use the max-p algorithm for spatially constrained clustering. After selecting the first two principle components, specifying the queen contiguity weight file, setting the minimum size constraint for population as 10% by default, and setting the number of iterations to 100, we run the algorithm and the resulting cluster map is shown below:



We can see from the map above that the algorithm yields 5 clusters consisting of 2 to 8 communities. The ratio of between sum of squares to total sum of squares is `0.842488`.

Next we do sensitivity analysis by setting the number of iterations to 5000, and the resulting map is shown as below:



We can see now that the algorithm yields 6 clusters consisting of 2 to 6 communities The ratio of between sum of squares to total sum of squares is `0.852041`, which is greater than the one using 50 iterations, and the number of iterations bigger than 5000 will not improve the ratio. So it means by using the Max-p algorithm, we find the optimal number of clusters is 6 and the optimal ratio of between sum of squares to total sum of squares is `0.852041.` Compared with the skater algorithm which produces the ratio 0.898 when the number of clusters is also 6, the max-p algorithm performs less well.