Rashmi Mariyappa
Statistical Models + Methods Class

Chicago Community Area: Exploratory Analysis

Introduction

Background

The City of Chicago has one of the largest public data portals in the United States. For my project, I wanted to take advantage of this availability and do an exploratory analysis of the city we live in using this real world data.

Chicago is divided into 77 official community areas. When most people think about the city they group these community areas by larger location: the North, West, and South sides of the city. Another way Chicago is commonly talked about is in terms of racial segregation with community areas grouped by racial composition. I was interested in ignoring location and racial community identities to see unconventional ways in which these 77 community areas could be grouped, and what commonalities may not be immediately apparent.

Additionally, it seems everywhere you look in Chicago there is a construction crane for new buildings. I also wanted to use the data collected to see what community features are driving building permits in the city.

Business Application

From a business perspective this information can be very useful for the Chicago real-estate market. For example, maybe you are a real-estate agent with a buyer interested in the Logan Square neighborhood, but it is too expensive for them. What neighborhood is most like it that is more affordable? What will be the next hot neighborhood in the city based on important features?

Data Collection, Cleaning, Summary

Data Collection & Cleaning

As mentioned in the introduction, a part of this project was constructing my own data set. This aspect of the project took a significant amount of time. Therefore, I wanted to discuss what was involved in the process.

First, I researched where I could collect data and aggregate statistics about the community areas. One large repository of data was the City of Chicago Data Portal. I collected a variety of individual data sets from this site including: community names, community numbers, 311 requests for rat complaints, garbage cart, graffiti removal, vacant property complaints, electricity use, city owned lots, blue and red line rail stations, and building permits. Each of these were an individual data set most of which had to be cleaned and transformed in some way. For example, the building permit data set contained permits for many types of requests. I had to research the code words for new construction and renovation to be able to pull this specific type of data I was looking for. Most of the files had data from 2011 onwards and I choose to only keep the records from 2016 onwards to get a more recent “look” at the community areas. Additionally, many of the files such as the one for 311 requests had individual observations for each request. Therefore, I had to aggregate this information to community areas.

The DePaul Institute for Housing Studies and the Chicago Metropolitan Agency for Planning (CMAP) “Community Data Snapshots” were a good resource for information. Both sources had useful data already compiled by community area. I obtained a variety of variables from these sources including: population total, median age, income, residential sales, residential stock, mortgage activity, and walk score.

Finally, the most time consuming part of the data collection process was data that had to be manually calculated. For the distance variable I wanted a measure in miles of the distance from the community area to the center of the city. I was not able to find this information pre-existing from a list. I devised a way to calculate this variable in Google maps using the measure feature. However, this was a manual process that involved searching each community area in Google maps determining the center and then clicking to the center of the city. I used 1 S State Street for the center as it is the

0,0 mark in the Chicago address grid. I found average home prices on the website Trulia. However, they were listed by neighborhood and not community area, so I had to map these back to community area through research.

The most challenging part of the data collection was that the building permit file and red and blue line data were not listed by community area. The building permit data listed the address of the request and latitude longitude coordinates. The building permit data set contained 13,931 variables so it would not be possible to map these manually. Additionally, if I was not able to map these to community area I would not be able to complete part of the project that was proposed. A solution to this problem was found by using the open source program QGIS. Using QGIS I mapped the boundaries of the community areas, and then plotted the coordinates of each permit onto that map. In order to translate the data seen on the map into a readable data file I used the "add polygon attributes to points" process within QGIS. This added the community area number to each permit. The CTA dataset contained the latitude and longitude of each rail station in a single field. I split that into separate fields in order to plot them on the map and used the same process in to obtain the dummy variables for whether a red or blue line station was present in the community area.

Data Summary

I had two final data sets. One file contained the Chicago community areas and various information describing the community. This file had 77 observations, one for each community area. The second file contained individual building permits for renovation and new construction over $45K from 2016 to present (March 17, 2018). This file has 13,931 observations: 11,162 for permits for renovation and 2,769 for new construction.

The brief variable description for each of the two data sets is listed below. I have included a data citation file with the data sets that have been submitted which includes where the data was collected and comments about details and preprocessing needed.

-COMMUNITY AREAS DATA- 19 VARIABLES

"ComArea" (77 – names of community areas of Chicago)
"ComNo" (77 – number id of community areas of Chicago)
"Dist" (distance from downtown miles)
"PopTot" (total population)
"MedianAge" (population median age)
"Income" (median income)
"HomePrice" (median sales price)
"ResSales" (sales per 100 residential parcels)
"ResStock" (housing stock % single family homes)
"MortgAct" (mortgage activity per 100 residential parcels)
"ElecUse" (average electrical use per sq ft in kwh)
"CityLots" (# of city owned lots)
"BlueSta (has blue line train station)
"RedSta (has red line train station
"WalkScore" (walkability score)

COMMUNITY AREAS DATA - CONTINUED

"R311Rat (# 311 requests regarding rat baiting)
"R311Graf (# 311 requests regarding graffiti removal)
"R311Cart (# 311 requests regarding garbage carts)
"R311Vac (# 311 requests regarding vacant buildings

-BUILDING PERMIT DATA- 9 VARIABLES

"ID" (ID number)
"Permit_Type" (2 - kind of permit)
"Issue_Date" (date permit was issued)
"Issue_Year" (year permit was issued)
"Estimated_Cost" (estimated cost of request)
"Work_Description" (description of permit)
"Latitude" (latitude of location)
"Longitude" (longitude of location)
"ComNo" (77 – number id of community areas of Chicago)

Grouping Chicago Community Areas: Cluster Analysis

In project report 1 one of the hypothesis that was proposed was:

Hypothesis 1: *The best cluster groups of the 77 communities will not divide them simply into two binary groups (rich vs. poor, North vs. South, etc.) but rather will provide a diverse set of communities in each with more than 2 clusters present.*

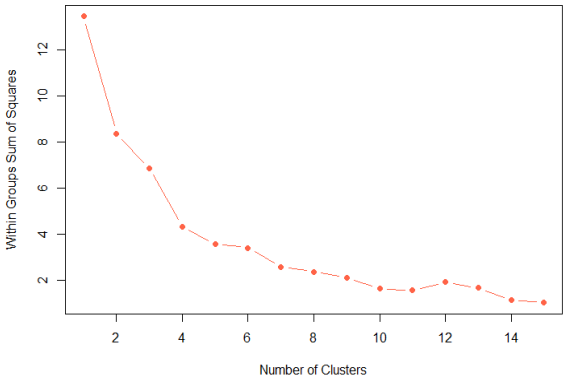
The methods proposed to test this hypothesis were using K-means and hierarchical agglomerative clustering. I ran both these methods on two different subsets of variables for the communities.

311 Request Clustering

The first subset of variables that I chose were the 311 requests by community for rat complaints ("R311Rat"), graffiti removal ("R311Graf"), garbage carts ("R311Cart"), and vacant building complaints ("R311Vac"). I chose this subset of variables because I was curious to see how the communities grouped by the types of requests they were making to the city. Additionally, these variables did not have any obvious relationship to race or community location which I was looking to avoid.

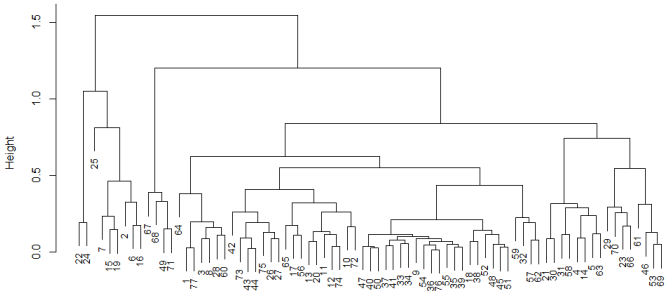
First, the k-means cluster algorithm was used. The data was normalized to a 0 – 1 scale. I used a scree plot with the total within cluster sum of squares to determine the number of clusters. The graph is pictured to the right. There is not a clear "elbow" in the curve, so we tested the K=4 to K=8 results. I chose K=4 as our final model for k-means as the cluster means/centroid differentiation was maximized. The total within cluster sum of squares for this model was 4.32 and the between cluster sum of squares to the total sum of squares was 67.8%.

Scree Plot – K-Means



I also wanted to test hierarchical agglomerative clustering before settling on our final model for the 311 clustering request. I built hierarchical agglomerative models using complete, average, and single linkage methods. Unfortunately, all three models were producing clusters with one large cluster and several very small clusters. This can clearly be seen by the dendrogram for the hierarchical agglomerative clustering pictured to the right. For example, the tree was cut to produce 5 clusters and the number of observations for each cluster is in the chart to the right. One can see that cluster 1 has 48 (62%) observations while clusters 4 and 5 only have 2 (3%) and 4 (5%) observations respectively. Therefore, for the final model for analysis I chose the K-means clustering with K=4.

Hierarchical Agglomerative Clustering - 311 Requests
Complete Linkage - Dendrogram



Insights

I analyzed the K-means K=4 clustering further to obtain some insights into the groupings. First, I aggregated the mean of the variables for each cluster on the remaining variables not used to cluster by. I plotted the cluster means/centroid plot. These are pictured below. From these two charts I was able to draw some insights about the clusters. The variables that are highlighted in the chart below correspond to the way I identified the clusters. Below and on the next page is a listing of identifications.

Hierarchical Agglomerative Clustering - 311 Requests
5 Cluster – Number of Observations per Cluster

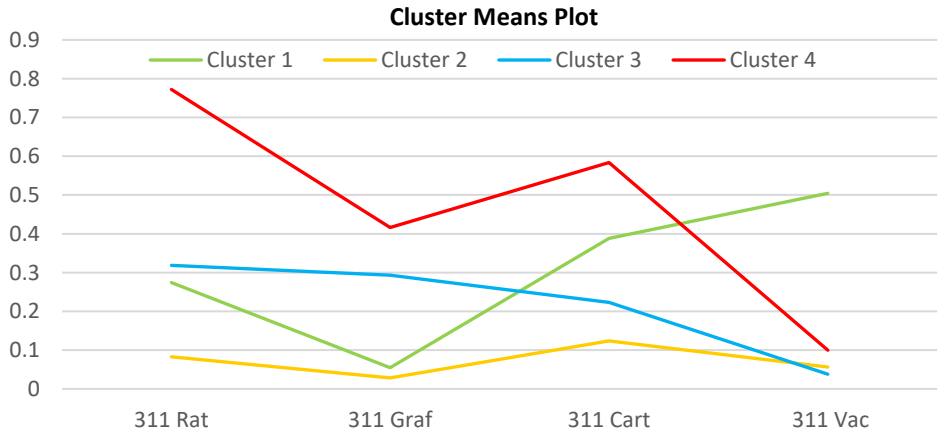
Cluster	1	2	3	4	5
# of Obs.	48	7	16	2	4

311 Requests K-Means K=4 Clustering - Aggregate Mean Variable Values by Cluster Assignment

Cluster	#Obs	Dist	PopTot	Median Age	Income	HomePrice	ResSales	ResStock	MortgAct	ElecUse	CityLots	BlueSta	RedSta	WalkScore
1	13	9.01	36,679	35.28	30,111	117,538	4.46	0.37	4.56	14,508	564.23	0	0.31	2.44
2	32	8.74	17,338	36.57	47,573	209,524	4.71	0.42	7.87	72,666	129.53	0.19	0.13	13.56
3	23	6.4	43,540	33.62	54,097	252,964	5.44	0.32	10.17	33,667	23.52	0.09	0.17	1.79
4	9	6.29	75,970	33.33	52,660	302,843	6.34	0.21	12.27	18,956	107.22	0.44	0.11	1.53

Cluster 1: Vacant Lots & No Graffiti
-Farthest to Downtown
-Highest City Owned Lots
-Lowest Home Price
-Lowest Home Sales

Cluster Means Plot



Cluster 2: No Problems Here!
-Lowest Population Total
-Medium Income
-Medium Home Price
-Highest % Single Family Home

Cluster 3: Average Problems Here!
-Higher Population Total
-Highest Income
-Higher Home Price
-Higher Home Sales

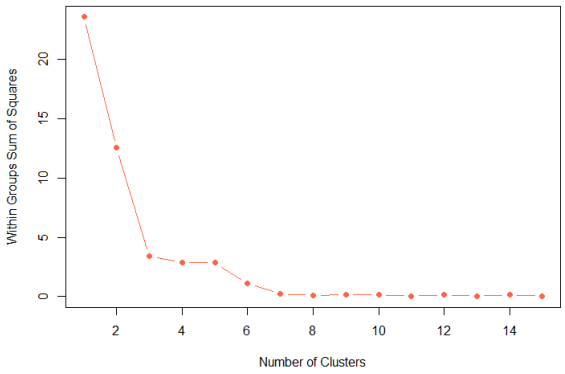
Cluster 4: Rats & Garbage Carts
-Closest to Downtown
-Highest Population
-Highest Home Price
-Highest Home Sales

"Green" Clustering

The second subset of variables that I chose to cluster with were ones that people might consider environmentally friendly or "green." These variables were electricity use ("ElecUse"), living close to the blue or red line trains ("BlueSta", "RedSta"), and an open space score of open acres divided by total population ("WalkScore").

I tested clustering with both K-means and hierarchical agglomerative clustering using complete, average, and single linkage. Hierarchical agglomerative methods were once again providing clusters that had very few observations 1-2 in many of the clusters. Therefore, I used K-means with a scree plot to determine the best number of clusters. From the scree plot we see two "elbows" at K=3 and K=6. I checked the cluster centroids and the K=3 clusters were simply dividing the clusters into 3 groups of community areas containing blue line, community areas containing a red line, and all other communities. Therefore, we picked K-means with K=6 as our final model. The total within cluster sum of squares for this model was 0.259 and the between cluster sum of squares to the total sum of squares was 98.9%.

Scree Plot – K-Means



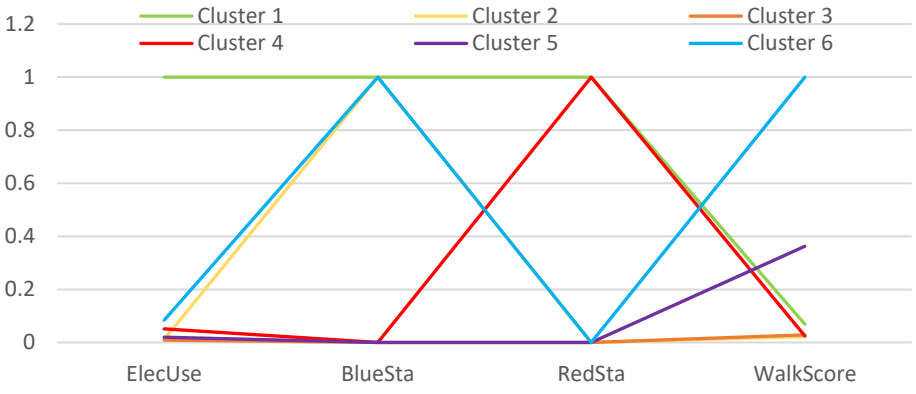
Insights

While whether a community area had a blue or red line station was a major factor in this clustering it was able to divide out a few more groupings. Additionally, it did pick out the community areas of the Loop and O'Hare as outliers and put them in their own clusters. In order to garner insights, I again aggregated the mean of the variables and plotted the cluster means/centroid plot, pictured below. From these two charts I drew some insights about the clusters. The variables that are highlighted in the charts correspond to the way I identified the clusters. Below and on the next page is a listing of identifications.

"Green" K-Means K=6 Clustering - Aggregate Mean Variable Values by Cluster Assignment

Cluster	#Obs	Dist	PopTot	Median Age	Income	HomePrice	ResSales	ResStock	MortgAct	CityLots	R311Rat	R311Graf	R311Cart	R311Vac	BlueSta	RedSta	WalkScore	ElecUse
1	1	0.31	33,442	32.31	93,254	290,250	6.00	0.00	9.20	10	478	4,211	9	1	1	1	7.47	1,154,623
2	10	6.05	51,769	33.84	51,235	294,508	6.31	0.23	12.27	182	2,387	6,198	2,391	80	1	0	2.53	23,713
3	48	8.16	32,464	35.28	45,384	202,725	4.71	0.41	7.98	150	1,192	2,882	1,419	107	0	0	3.23	20,360
4	12	5.99	44,316	36.23	48,117	238,323	5.83	0.18	8.65	234	1,502	2,267	1,124	161	0	1	2.80	68,298
5	5	12.11	12,384	33.24	44,777	163,675	4.08	0.64	5.96	230	282	236	568	47	0	0	38.62	32,749
6	1	15.87	13,695	37.14	46,638	132,144	5.20	0.12	6.70	0	21	209	147	1	1	0	106.21	106,502

Cluster Means Plot



Cluster 1: The Loop
-Energy Hog
-Highest Income
-Multiple Rail Lines

Cluster 2: Pricey Blue Line Only
-Highest Population
-Highest 311 Requests
-Highest Mortgage Activity

Cluster 3: No Trains & Average
-Medium Values Across
-Lowest Electricity Use

Cluster 4: Red Line Only
-Higher Electricity Use
-Higher Population
-Highest City Owned Lots

Cluster 5: No Trains & Far
-Low Population
-Highest % Single Family Homes
-Higher Walk Score

Cluster 6: O'Hare
-Farthest to Downtown
-Lowest Home Price
-Low 311 Requests

Did We Prove the Hypothesis?

Finally, I wanted to check the original hypothesis and prove it to be true or false. To do this I looked at the communities assigned to each cluster. For the 311 request clustering, in clusters 2, 3, and 4 I found a variety of communities not usually grouped together. For example, cluster 2 included the communities: East Garfield Park, Loop, and Beverly. All three are located on different "sides" of the city and have different racial makeups. Another example was cluster 3 which included communities such as Lincoln Park, South Lawndale, and Archer Heights. Again, much like cluster 2, these three communities have different ethnic compositions and are located on differing sides of the city. However, cluster 1 seemed to consist of communities mainly on the south side of the city. We can see this same thing in the clustering for "green" features. Cluster 2 contains community areas Norwood Park and West Garfield Park which are located on different sides of the city and have different ethnic compositions. This can be seen in cluster 4 as well that was mainly grouping community areas with red line stations. This cluster has community areas Lakeview and Roseland which are also located on different sides of the city and have different wealth and ethnic make-ups. Overall, I would say that our hypothesis that the best clustering of the 77 communities will not divide them simply into two binary groups (rich vs. poor, North vs. South, etc.) but rather will provide a diverse set of communities in each with more than 2 clusters present proved to be true.

Community Area Features Driving New Construction Over Renovation: Decision Tree

In project report 1 one of the hypothesis that was proposed was:

Hypothesis 2: *Two important features driving new construction permits are having a blue line stop nearby and being close to downtown.*

The method proposed to test this hypothesis was using a type of decision tree classification model, not for prediction but rather to determine top features. Therefore, the reason the tree model was used was for interpretability over prediction. First, in order to do this, I needed to create a response variable. I decided to create this variable based on the kind of permit, 1 = new construction and 0 = renovation. With this response variable the community area features driving new construction building permits over renovation building permits was checked.

Variable Importance

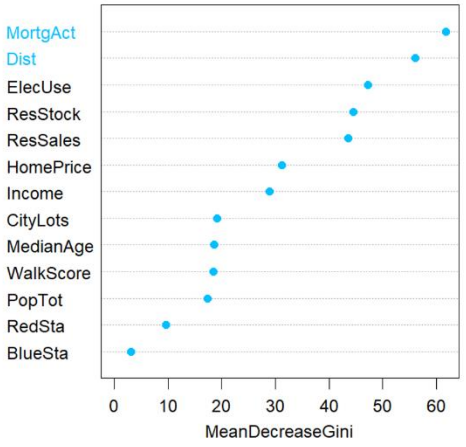
Next, I ran random forest and bagged decision tree models to determine variable importance. The data was split with a 70% - 30% train to test ratio with observations randomly selected. There were 9,751 observations in the training set and 4,180 observations in the test set with 13,931 observations total. I built 11 random forest tree models using all the community area features (17 total not including id variables) with the m parameter ranging from 1 to 16. Additionally, 7 random forest models were built excluding the 4, 311 request features (13 total not including id variables) with the m parameter ranging from 1 to 12. Finally, I built 2 bagged models with and without the 311 request variables (m=17 and m=13, respectively). For all these models the number of trees was set to 1,000. I checked test error to pick the final models even though the models were solely being used for variable importance interpretability. For all these models a test error of 18.35% and a training error of 19.70% was obtained. This is unusual and will be discussed in the section on the next page titled, "Did We Prove the Hypothesis?" Since error could not be used as a means to select a model, I used the standard $m=\sqrt{p}$ to select the final models. I rounded the decimal values for m up so a model of m=5 and m=4 was used for the final random forest with and without the 311 request variables, respectively.

Insights

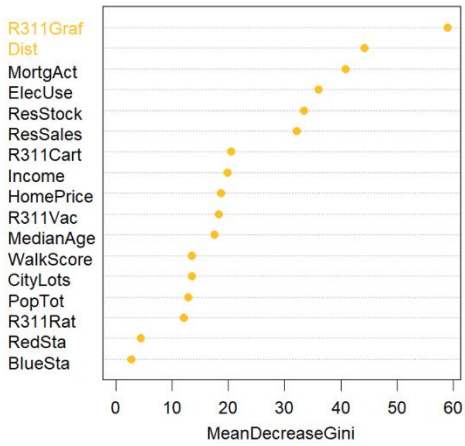
Next, I reviewed the variable importance graphs to see the top features driving new construction compared to renovation permits. Since the figures for accuracy were problematic (they did not change for any model), I reviewed the mean decrease in the Gini index to determine the top features. Below is the graph of the variable importance for the

models including and not including the 311 request features. The top 2 features not including the 311 request variables were mortgage activity (MortgAct) and distance to city center (Dist). The top 2 features including the 311 request variables were 311 graffiti request removal (R311Graf) and distance to city center (Dist).

Top Features without 311 Requests



Top Features with 311 Requests



Did We Prove the Hypothesis?

My original hypothesis predicted that being close to the blue line station (BlueSta) and being close to downtown (Dist) would be 2 of the top features. Our results showed the hypothesis to be half correct. Being close to the blue line was not important but being close to the city center was in the top 2 features with both sets of variables. However, should we trust these results given that the error was not changing for any of the models? I reviewed the confusion matrix (it was the same for both models) pictured to the right and saw that while the accuracy was 80.30% the sensitivity was very low at 6.69%. This model was mainly predicting most observations to be renovation permits as opposed to new construction permits. With this low sensitivity I evaluated this to be a poor model and we should not trust the findings. What can be done to correct this is discussed in the "Conclusion, Improvements, Next Steps" section below.

Confusion Matrix		Actual	
		0	1
Predicted	0	3360	744
	1	23	53

Conclusions, Improvements, Next Steps

In conclusion, I found this project to be valuable as I was able to develop my own research topic and questions. Additionally, it was interesting to build my own real-world data set to address the research question. Building the data set took much longer than I expected. Finding the data needed, cleaning it to the appropriate form, and creating various variables needed for the analysis was challenging. The cluster analysis yielded interesting insights about possible unconventional ways that community areas could be grouped and proved my hypothesis to be true. I mainly found two subsets of variables to cluster the community areas by that were not directly related to race and location, 311 requests and "green" factors. The next steps for the community clustering analysis would be to build out this data set further and find other features that can be clustered on which are not correlated too much to race or location. One example of this would be variables relating to the type of housing stock each area is comprised of single family homes, condos, apartments, etc. The feature importance analysis on what was driving new building compared to renovation permits turned out to be unreliable as the model was poor. Upon examining the data further, I realized that the data was unbalanced. There were only 19.88% of the observations that were new construction (response = 1). This unbalanced data could lead to a poor model. Additionally, upon reviewing the work description field in the building permit data set I believe these requests can be further filtered to possibly only residential projects. However, this would be difficult, as this is not hard coded into the data and is only a text field. Balancing the building permit data set and further cleaning the observations would be a way to possibly improve the results for the model created for variable importance. Finally, in thinking about the next steps proposed above, this project has made me realize the iterative nature of data analysis.