# Clustering Experimentation with 2011 Housing Market Topology Data for the City of Baltimore

Robert Brasso

***Abstract*— As established in the accompanying paper, home vacancies remain a major problem that many medium-large cities deal need to address. They use up valuable land, are often uninhabitable, cause a tax burden on the city, and bring down property values for the neighborhoods they are in. In the previous paper, we identified that there is a clear relationship between the metrics related to vacancies and their designation of being distressed or not being distressed. Additionally, we created a supervised learning model that could accurately predict the market classification of a neighborhood. In this paper, we will continue exploratory analysis on this data by running three unsupervised clustering algorithms to see if they will be able to also accurately predict market typology for a neighborhood. If successful, we hope to be apply the unsupervised model in a way that can give us real-time analytics on neighborhoods and attempt to identify patterns that lead to homes to fall into distressed market categories.**

## I. Introduction

In the accompanying paper, we established the problem of home vacancies in Baltimore, MD resulting from the dramatic decrease in population from the 1950's to the present. A housing typology report from 2011 was used to attempt to identify factors that lead a neighborhood to receiving a distressed or middle market distressed designation, which associates with a higher rate of home vacancies. Through exploratory data analysis we identified that certain attributes, such as the percentage of vacant lots, the percentage of owner occupied properties, and the sales price coefficient correlate with the designation of market typology for a neighborhood.

The data was then used as inputs for three classification algorithms, K-Nearest Neighbors, Random Forest, and Gaussian Naïve Bayes. I attempted to establish whether or not supervised learning could be used to create a model that accurately predicts the housing typology designation for a census block in Baltimore. The results of this experimentation showed that K-Nearest Neighbors was the best algorithm of the three testing and could be optimized to provide us 96% accuracy, as well as an f-measure of .96. These results validated that supervised learning could provide us a useable model for predicting housing market typology for neighborhoods.

While using K-Nearest Neighbors gave us very strong results, supervised learning algorithms require us to have a previous dataset by which to train the algorithm to make future predicts. In contrast, unsupervised learning algorithms do not require any prior information about the dataset, which makes them more efficient in real world use. This paper intends to ask the question of whether unsupervised clustering algorithms could also provide us with highly accurate and replicable results.

## II. Dataset Description

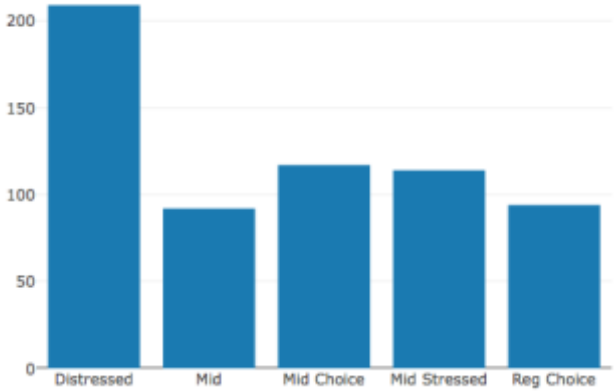
The dataset used in this paper was taken from *Open Baltimore***,** which is a website operated by the city of Baltimore that provides free datasets relating to the city. The dataset is called "2011 Housing Market Typology" and was created in order to inform neighborhood planning efforts, also informing residents of the local housing market conditions in their communities. The entire dataset has 626 records, contains 12 fields, and can be found by going to the following link: https://data.baltimorecity.gov/Housing-Development/2011-Housing-Market-Typology/782b-zpd7. Below each of the fields will be described in detail.

- **blockGroup** – Census block group. A map of all Baltimore census block groups can be found at http://www.mdp.state.md.us/msdc/census/cen2010/maps/blkgrp10/Baci_blkgrp10Roads.pdf
- **marketCategory** – Each block group is listed as one of the following groups…
    - **Regional Choice** – Competitive housing markets with high owner-occupancy and high property values
    - **Middle Market Choice –** Housing prices above the city average with strong ownership rates, low vacancies, but slightly increased foreclosure rates
    - **Middle Market –** Median sales of $91,000 as well as high ownership rates. Higher foreclosure rates, with slight population loss.
    - **Middle Market Stressed –** Slightly lower home sales than city average and have not shown significant sales price appreciation. Vacancy and foreclosure rates are high and the rate of population loss has increased.
    - **Distressed –** Experienced deterioration of housing stock. Contains high vacancy and the lowest homeownership rates. Most substantial population loss.
- **sales20092010 –** Total number of residential sales from 2009 to 2010.

- **salesPriceCoefficientVariance –** Sales price standard deviation from 2009-2010 divided the mean sales price from 2009-2010.
- **commericalResidentialLandRatio –** Commerical and institutional land area divided by residential land area.
- **unitsPerSquareMile –** Amount of housing units per square mile.
- **residentialPermits –** Residential permits greater or equal to $50,000 divided by residential lots plus vacant lots.
- **vacantLots –** Amount of vacant lots divided by residential plus vacant lots.
- **vacantHouseNotices –** Vacant housing notices divided by residential plus vacant lots.
- **foreclosureFilings –** Foreclosure filings from 2009-2010 as a percentage of privately owned residential lots.
- **medianSalesPrice20092010 –** The median sales price for homes in census block from 2009-2010.
- **ownerOccupied –** Estimation of all occupied residential units that are owner occupied.

## III. Data Preprocessing

During the preprocessing phase, the data was reviewed to ensure there is not any null or missing values. After verifying the data, I wanted to make the dataset more conducive to clustering. Since distance measures - such as Euclidian distance - can be skewed by values that have larger ranges, I decided to perform a min max normalization on all of the fields except the Block Group identifier and the market category, which were removed for testing the algorithms. Lastly I created a bar chart to show the distribution of records for each market category, so I could compare the results from the clustering algorithms to see how accurately they grouped the clusters in relation to market category.



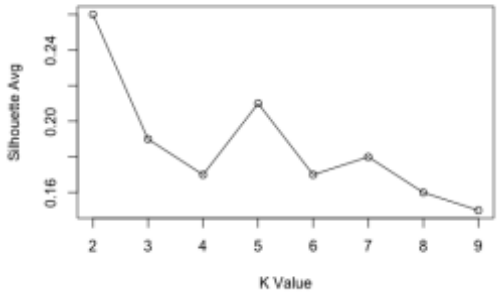

The totals for each category were distressed = 209, mid stressed = 114, mid market = 92, mid choice = 118, and reg choice =94.

## IV. Clustering Model Results

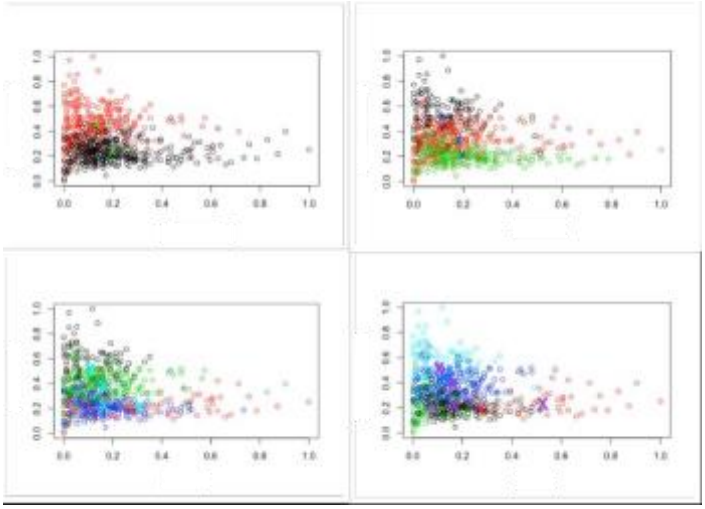
As stated at the onset of this paper, our purpose in testing these algorithms is to create an unsupervised learning model that will accurately predict market category based on various other attributes. We have established that supervised learning models work well at accomplishing this task, but to give us options in our future application of this data, we also want to test clustering models to see how accurately they can predict market category. The three clustering models we will use are K-Means Clustering, DBSCAN, and PAM (Partitioning Around Medoids).

### *A. K-Means Clustering Results*

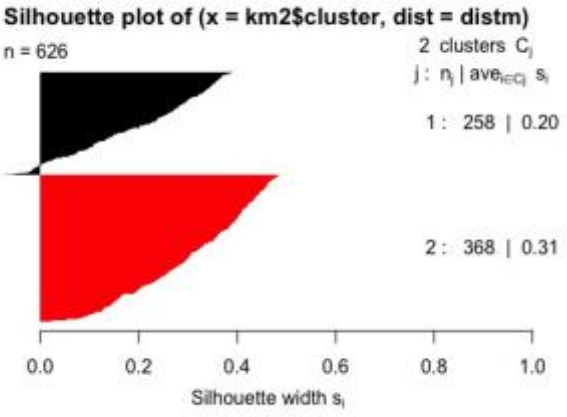
For the K-Means clustering portion of this experiment, I decided to test various K values to determine which would provide us the best results. Below is a line graph of the the silhouette averages based on changing K values.

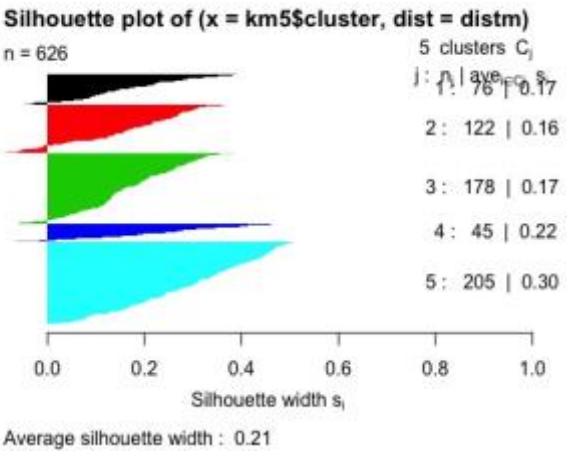



As you can see, the maximal silhouette averages occurred when K was equal to 2 and when K was equal to 5. Since the values beyond 5 become smaller as the k value rises, I decided to only plot the K values from 2 to 5. Below is a collection of scatter plots for K values 2 to 5. The centers have been marked with x's.



As stated above the two best K values were 2 and 5. Therefore, I plotted the silhouette graphs for each of these values and included them below.

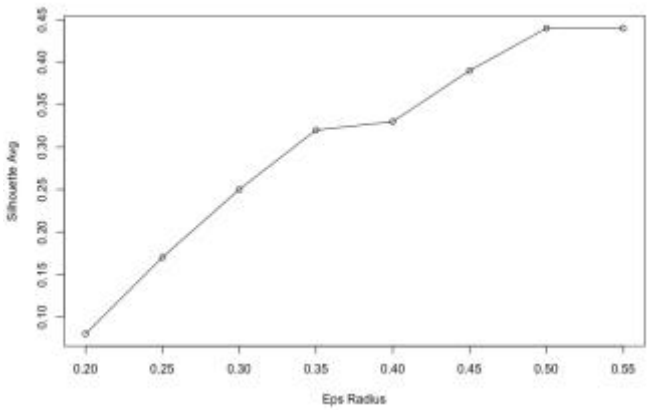


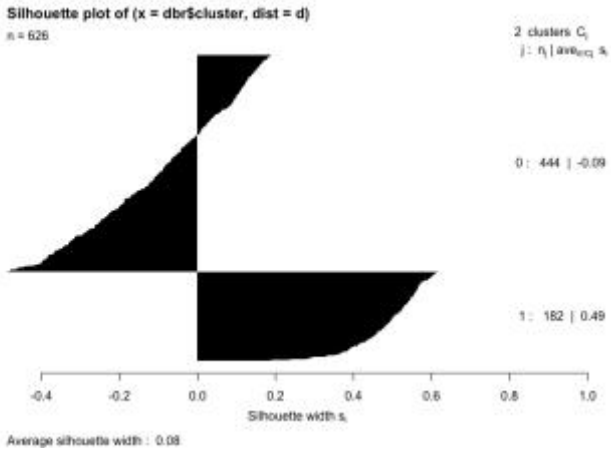

The silhouette averages for both of these are not particularly strong, but it does appear that the fifth cluster where k=5 might be fairly representative of the distressed market category as the total counts are very similar and the silhouette coefficient for that category is higher than the rest. The two clusters where k=2 seem to potentially group distressed and mid market stressed into one cluster and the rest into another cluster, however the totals are a bit off. In both of the results, it appears that the algorithm is potentially running into problems clustering the middle market categories likely due to similarities in the composition of metrics.

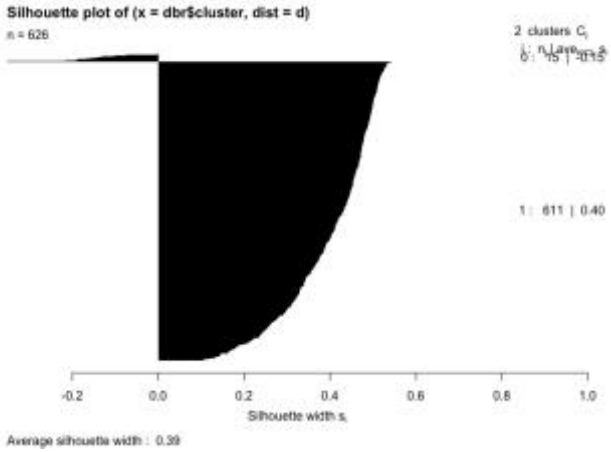
*B. DBSCAN Results*

For the DBSCAN testing, I set the minimum number of points to 25 and then tested epsilon values of .2 to .55. That range was used since the data was normalized to a min-max of 0 to 1. Again, like the K-Means testing I charted the different silhouette averages and then compared a few of the silhouette graphs. The figure below contains the charted values:




As seen in the chart, as the epsilon radius values become larger, the silhouette average becomes larger. While these results seem like an improvement over K-Means, they are actually inflated and looking at a few silhouette graphs will emphasize a major problem. Below are the silhouette graphs when epsilon was set to .2 and .45:

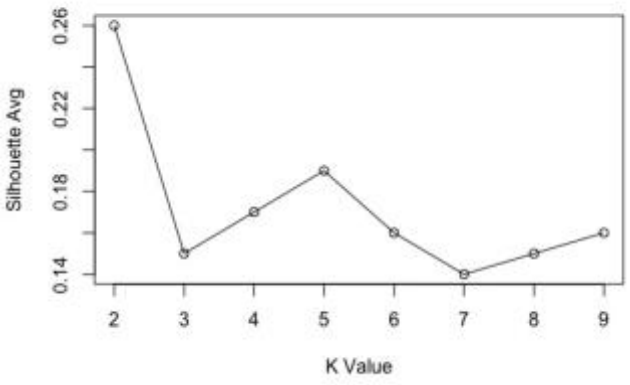


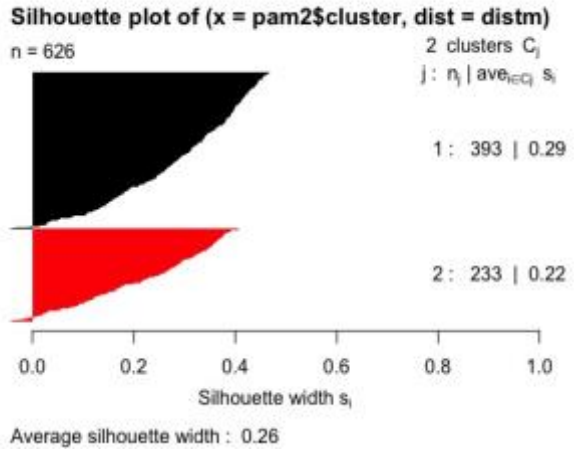

The major problem seen above is that DBSCAN is attempting to cluster the data into only two categories and as epsilon increases, the data is further stratified to one cluster. We then get one cluster with a very poor silhouette coefficient and one with a very strong coefficient, giving a dramatically inflated average. This trend continues as epsilon values grow and DBSCAN further attempts to create on large cluster.

### C. *PAM Results*

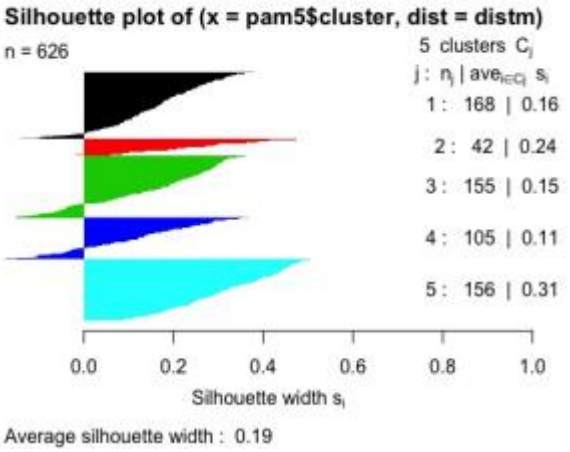
Partitioning Around Medoids is a variation of the K-Means algorithm and also uses k values to determine the amount of clusters the algorithm finds. We used the same K values as the K-Means algorithm and charted the differences between the K values below:




This chart is very similar to the K-Means algorithm results, however the silhouette coefficients are all slightly lower except for k=2 which is the same. Looking at K=2, we see similar results to the K-Means algorithm and it follows that it also suffers from the same problem as classifying the middle market data.




The second cluster seems likely to contain most of the distressed records and perhaps some of the mid market stressed records, but it is difficult to determine. When k=5, we hope to see a distribution similar to the true distribution of the market category records.




In reality, what we get doesn't seem to accurately resemble the original distribution at all. It appears that this algorithm does not do well with identifying the market typology categories.

## V. Conclusion

From these clustering experiments, it is clear that a supervised learning approach to classifying market typology is much more effective than an unsupervised approach. DBSCAN seemed to be a poor choice for this data because it wanted to collapse the data into two clusters and there was a large amount of variance in the clustering. PAM also seems like a poor choice since it was not able to come close to resembling the original distribution from the dataset. The K-Means algorithm offered some promise when K=5 in that it seemed to accurately create a cluster for the distressed market category. Our goal with this research is primarily to identify distressed and mid market stressed neighborhoods as soon as possible, so to this extent K-Means may be a viable solution, however K-Nearest Neighbors from the previous paper on supervised learning algorithms was far more accurate.

It appears the major problem with using unsupervised learning in this dataset is that middle market categories are too similar and thus cause problems when the algorithms try to cluster the data. These algorithms could potentially help identify the extremes of distressed and regional choice, but the middle market data makes them difficult to use for our purposes.

## References

https://www.theatlantic.com/business/archive/2014/10/can-homeless-people-move-into-baltimores-abandoned-houses/381647/