Baltimore Housing Prices Disparity for Comparable Neighborhoods: A Case for Enabling Interactive, Visual Exploration of Neighborhoods

Akshay Peshave, Siraj Memon, Vedmurtty Chavan, Tim Oates

Computer Science & Electrical Engineering
University of Maryland, Baltimore County
{peshave1, siraj1, ved1, oates}@umbc.edu

Abstract

As government agencies increasingly make public data available online, it provides opportunities to leverage such data for descriptive, predictive and prescriptive analytics. One domain where these technological capabilities are applicable is real-estate development and housing market domain. This domain is of interest to home buyers, investors and policy makers. Diverse and varying preferences of residents of a geography are latent behavioral factors that affect residential property prices.

This paper describes a geographical area agnostic housing typology classifier for Baltimore City communities or neighborhoods. Further, it discussed correlation analysis and composite Vital Signs scores to characterize city population perceptions of different community development categories. These scores enable community clustering to investigate price disparity in comparable communities based on configurable categories and year-on-year trend analysis. Various visualization possibilities are discussed in conjunction with these approaches to make a case for interactive, visual exploration of geographical communities which may be extended to comparative analysis across geographies.

Keywords— Multi-label Classification, Correlation Analysis, Clustering, Data Visualization, Baltimore City, Housing and Community Development.

1 INTRODUCTION

As government agencies increasingly make public data available online, it provides opportunities to leverage such data for descriptive, predictive and prescriptive analytics. State-of-the-art data technologies and analytic methods can be used to drive innovation and development of platforms to help inform public policy effectively and efficiently for social good. This also makes direct information provision to and engagement of the public-at-large possible at scale. Interactive technology platforms which enable custom-tailored analyses and engagement processes will benefit both policy makers, policy enforcers and the general public. Data science technologies also allow federated open data consumption and and analytics to provide insights at varied abstractions which will help citizens and policymakers make informed choices.

One domain where these technological capabilities are useful is real-estate development and housing market. Consumers buy residential properties for personal use and as investment instruments. Consequently, the housing market is susceptible to volatility due to sensitivity to varying factors that affect public perception and pricing bubbles. A wide array of data needs to be acquired and analyzed to inform such high monetary value decisions. In addition to geographic and housing market macro-data, a more subtle and latent behavioral factor affects property pricing: diverse and varying preferences of residents of a geography.

The ability to conveniently access and analyze aggregated forms of data which affects home prices is important to analyze such latent factors. Interactive platforms consuming relevant data from government agencies and third-party oganisations and applying state-of-the-art data science can efficiently enable this. Such platforms can serve to inform public policy and developmental efforts as well as help consumer decision making. The analytics workflow may take many forms based on the use-case and requirements thereof.

This works explores one such analytics workflow for the residential property market in Baltimore City. The task is to enable visual exploration of Baltimore City neighborhoods which are similar in terms of some character-

istic features but have disparity in home prices. The datasets used in this work are sourced from OpenBaltimore (http://data.baltimorecity.gov/) and Baltimore Neighborhoods Initiative Alliance (BNIA) (http://bniajfi.org/).

This paper describes a housing typology classifier trained on derived features which are geographical scale agnostic. The classifier is trained on relatively small US census tracts and used to classify relatively larger Community Statistical Areas (CSAs). We then discuss correlation analysis of CSA "Vital Signs" with home prices in Baltimore City from 2011 to 2014 and utilizing the coefficients to aggregate vital signs data to compute corresponding vital signs category scores. This score is a manifestation of population perception of corresponding features. We further perform clustering of CSAs based on vital signs and vital signs categories to visually explore home price variations in otherwise comparable CSAs. The terms CSA, community and neighborhood are used interchangeably in this work.

2 DATASETS

In this section we briefly describe the datasets used in this work from two sources: OpenBaltimore[1] and Baltimore Neighborhoods Initiative Alliance (BNIA)[2].

2.1 Housing Market Typology [1]

This data set comprises of housing market typology, an ordinal scale for the residential real estate market, assigned to 710 census tracts for Baltimore City. This data is available for the period 2010-11 and is provided on OpenBaltimore[3]. These typology labels are assigned by the the responsible government agency primarily based on the median residential property sales prices for a tract and the number of residential properties sold in a year.

2.2 Vital Signs '14 [2]

The Vital Signs data set is created and maintained by the Baltimore Neighborhoods Initiative Alliance in collaboration with the University of Baltimore. This dataset is created using data points for Baltimore City from various government and non-governmental agencies and organizations at the city, state and national level. It provides data recorded for 55 Baltimore City Community Statistical Areas (CSAs) from 2011 through 2014. The vital indicators are grouped into 8 sections or categories.

[1] www.data.baltimorecity.gov
[2] www.bniajfi.org
[3] www.data.baltimorecity.gov

3 OUTLINE OF ANALYTICS PIPELINE

In this section we provide an overview of the analytics pipeline used in this work. Figure 1 shows the analytics pipeline. It performs four broad tasks:

1. Typology assignment for Baltimore CSAs: A decision classifier is trained using the 2011 Housing Market Typology dataset. It is used to classify CSAs using the VitalSigns housing category vitals on a year-to-year basis.

2. Selecting key vitals per category: Pearson correlation analysis of vitals with CSA typology is done per Vital Signs category. High correlation vitals are identified based on a correlation threshold.

3. Category impact measurement on home prices: Correlation-weighted aggregation of vitals per category are analyzed for correlation with CSA typology year-on-year.

4. CSA clustering analysis: CSAs are clustered based on key vitals and typology variance within clusters identified.

Additionally, the geometry of geographical regions, represented as Esri Shapes[3], are made available as a shape-file in the Vital Signs dataset. These shape files are utilized for visualizing geographical regions.

4 METHOD & OBSERVATIONS

This section discusses the analytics methods in detail alogn with their performance and observations thereof.

4.1 Typology Prediction for Vital Signs CSAs

The granularity of the Vital Signs dataset is at the Baltimore City CSA level. The 2011 Housing Market Typology labels for Baltimore City are available at census tracts granularity. CSA and census tract boundaries do not necessary align which makes the task of mapping census tract labels to CSAs non-trivial.

We model this task as that of classification. We train a classifier on the 2011 housing typology data and use the trained classifier to classify the CSAs. One requirement for

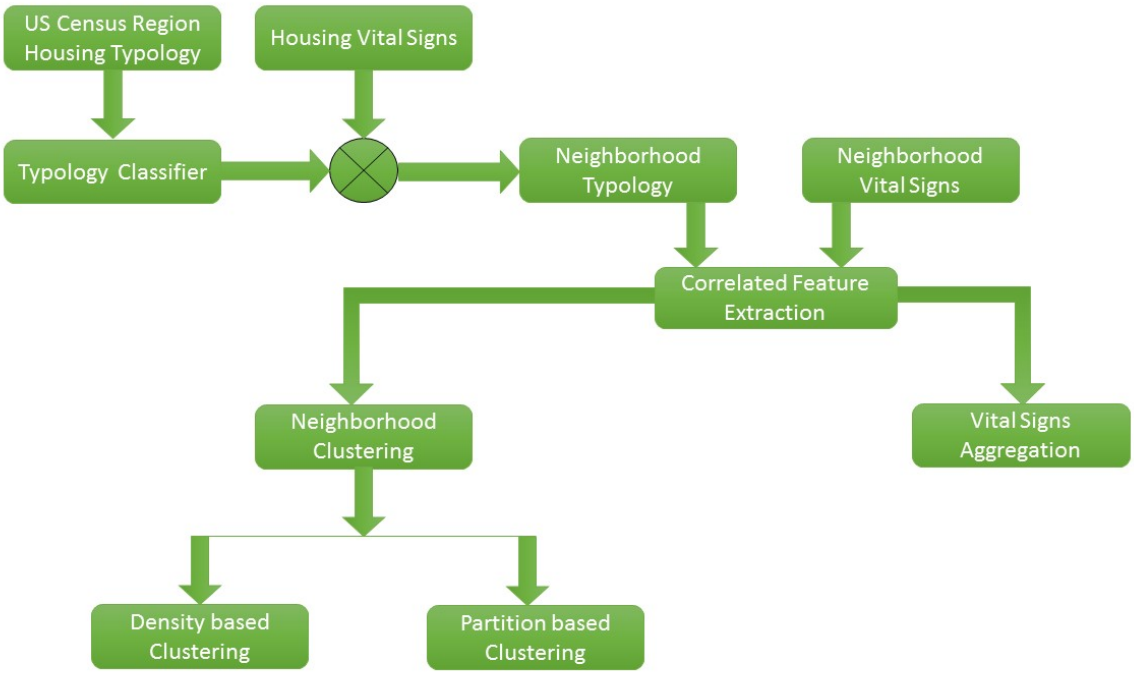


Figure 1: Analytics Pipeline

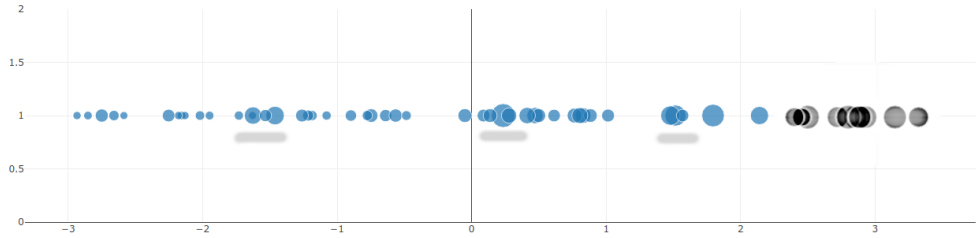


Figure 2: tSNE Based Neighborhood Similarity Plot for "Arts & Culture", "Education" and "Economic & Workforce Development" Vital Signs Categories

this approach to succeed is that predictors used to model the classifier must be available for the CSAs as well. Another requirement is that the predictors should not be affected by the size of the geography they are observed for. The latter guarantees the same characterization by predictors when observed for CSAs which are significantly larger in size and for census tracts which are relatively smaller. If this doesn't hold true, such affected predictors may skew the model and result in incorrect classification for the larger sized CSAs.

A new feature is constructed to avert any side effects of changes in geographical area and residential units density while manifesting all the information important to the classification task. The new feature is defined as:

$$Residential\ Units\ Per\ Sq.Mi\ Sale = \frac{unitsPerSquareMile}{sales20092010} \quad (1)$$

This constructed feature contains all the required information normalized by area. The feature can also be computed for each CSA using the Vital Signs Housing dataset. This enables consistent characterization of census tracts and city CSAs by our new constructed feature.

A decision tree typology classifier is trained on the two features: (a) median sales price (b) residential units per square mile per sale The classifier has a training set accuracy of 100% and 10-fold cross-validation accuracy of 94%(+/-5%). This classifier is used to classify the CSAs year-on-year for the 2010-2014 period.

4.2 Vital Signs Correlation with Typology and Category-level Aggregation

Housing typology assignment for CSAs enables us to leverage the Vital Signs data for further analyses. Correlation coefficients of vital signs with home prices are reflective of

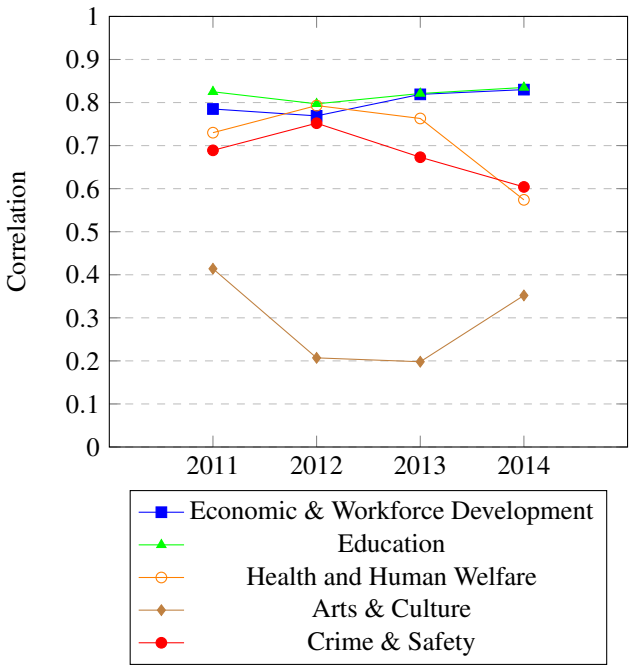


Figure 3: Vital Signs Categories Correlation w/ Median Sales Price

the perception of the vital signs by the city population. Further, these correlation coefficients computed on a year-on-year basis from 201 through 2014 allow us to investigate the trends of these perceptions.

These correlation coefficients are used to find a weighted mean of negatively and positively correlated vital signs in each vital signs category. Finally we compute a vital signs category score which is a combination of the two means. This score is a value scaled between 0 and 10 both inclusive and is a manifestation of the perception of the category as it relates to home prices by the population.

The line plot in Figure 3 shows the impact of some vital signs categories based on the correlation of our category score with the median sales prices for the 2011-2014 period. We observe that "Economic & Workforce Development" and "Education" are perceived of greater importance than other factors. The "Arts & Culture" vital signs trend shows that public perception of it's importance is on the rise again. Perceived importance of "Crime & Safety" and "Health & Human Welfare" are on a subtle decline. Causality of these trends require additional data and investigation but the scoring method herein can certainly facilitate such analysis and is also a value add by itself for the end user.

Figure 2 visualizes Baltimore CSAs for the year 2014 based on a user selection of "Arts & Culture", "Education" and "Economic & Workforce Development" Vital Signs categories. The size of each point is proportional to the neighborhood's mean home sales price. This visual element uses the category scores for these 3 categories and applies t-Distributed Stochastic Neighbor Embedding (t-SNE) [4] to the data. tSNE helps bring similar neighborhoods together in the reduced dimensionality space (in this case 2D). Points towards the far right have a higher composite score while those to the left have a lower score. In this case the neighborhood ordering is driven by the education score and the neighborhoods to the far right (marked in black) have education scores > 6.5. The regions of the plot that are underlined in gray show disparity in home prices for neighborhoods with similar scores in the selected categories. This visualization can be further extended to 3 dimensions or quadrant based visualization for different visual analysis use cases.

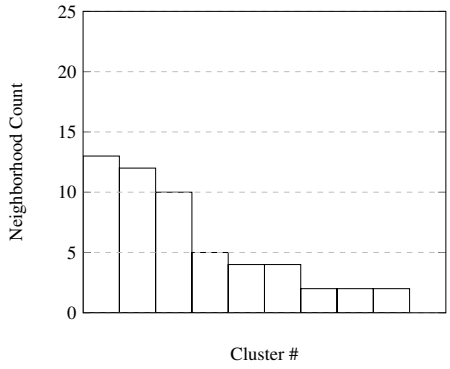
Outlier CSA	Typology
Inner Harbor/Federal Hill	5
Downtown/Seton Hill	5
North Baltimore/Guilford/Homeland	5
Canton	5
Mount Washington/Coldspring	5
Cross-Country/Cheswolde	5
South Baltimore	5
Fells Point	5
Greater Roland Park/Poplar Hill	5

Table 1: Outlier CSAs Identified by DBScan w/ Market Typology

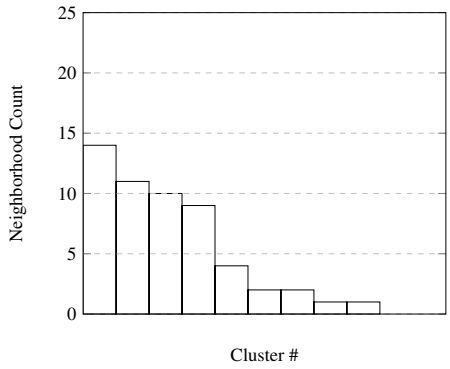
4.3 Neighborhood Clustering

This section explores clustering of CSAs to assess similarities between them based on vitals and consistency of typology labels per cluster. There are a wide variety of multi-dimensional clustering algorithms which can be broadly classified into partition based and density based clustering approaches. Either of these classes can be utilized individually or in a combination to perform hybrid hierarchical clustering. In this work we utilize partition based and density based clustering to understand the cluster structure in our data. We apply three different clustering algorithms and discuss cluster size, membership and distribution of typology within each cluster.

For partition based clustering we use K-means++[5] and spectral clustering[6]. For density based clustering we apply DBScan[7]. DBScan and spectral clustering approaches identify arbitrarily shaped clusters while K-means identifies more restricted cluster shapes based on euclidean radii from centroids identified for each cluster.



(a) KMeans++



(b) Spectral

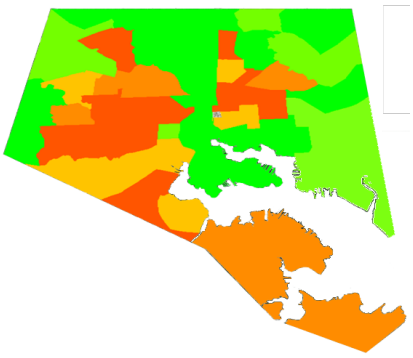
Figure 4: Partition-based Clusters Stats

Both K-means and spectral clustering require the desired number of clusters as an input parameter. We set the number of clusters to twice the number of typology levels i.e 10 clusters. This choice of number clusters is made to help identify substructures within typology groups and achieve better resolution at a finer granularity to assess typology consistency within clusters. Fig. 4 shows the cluster size distributions for K-means++ (Fig. 4a) and Spectral Clustering (Fig. 4b). K-means++ partitions the CSAs into 9 clusters while spectral clustering partitions them into 8 clusters. The sizes of clusters in both cases are fairly similar. Variations in cluster sizes and membership can be attributed to spectral clustering's ability to identify abstract shaped clusters rather than just spherical ones.

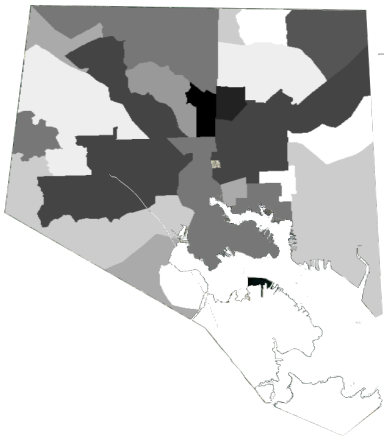
The clusters identified in both cases for 2011 show variance in typology within them. This is evidence that CSAs with comparable key vitals have a disparity in home prices and consequently typology. Fig.5 shows the 2011 Baltimore CSA Typology Map (fig.5a) and Clusters Map (fig.5b). It is apparent that several CSAs with high ordinality typologies (5-deep green and 4-lighter green) are clustered with those with lower typology instead of with each other. This visual element shows disparity in home prices for comparable CSAs.

We also perform clustering using DBScan. The input parameters are features related to the desired density level needed to qualify clusters as valid. This clustering technique generates a group of outliers which do not qualify for membership in any identified density clusters based on the input parameters. Outliers are characterized by unexpected values in one or more dimensions which situates them apart from high density regions.

We perform some preliminary analysis to help us identify best values for the input parameters based on the distribution of the input data set. DBScan generates one high density cluster and one outlier group. This suggests that CSAs are predominantly distributed in the key vitals space with uniform density except for a few outliers. The outliers are listed in Table 1 with their typology. All outliers are CSAs with typology 5. These are prime neighborhoods of Baltimore and seem to have exceptional values for one or more key vitals which sets them apart from the rest of the CSAs. A deeper investigation of vitals which are responsible for this behavior is to be pursued in the future.



(a) CSA Typology Map



(b) CSA Clusters Map

Figure 5: CSA Typology and Clusters Visualization Using Google Maps Pro

5 ONGOING AND FUTURE WORK

Exploring causality in addition to correlation is a planned future area of contribution. Additionally, incorporating crowdsourced data such as reviews, complaints and ratings in various communities is an area to be explored. Development of a web-based interactive platform for analysis of communities based on custom category and importance selection is ongoing. The interface will include density-based clustering based on user selected categories for on-demand, custom cluster analysis. A human-in-the-loop workflow to enable fine grained correlation analysis at the level of vital signs within each category and custom category score generation capability is planned for development in the future.

6 CONCLUSION

In this work we built a multi-label, geographical area scale agnostic, decision tree classifier to assign housing market typology to Baltimore City CSAs. The city population's perception of importance of community development categories was quantified using correlation of Vital Signs with median home prices for communities and used to compute aggregated Vital Signs category scores. These scores are shown to enable visual filtering of similar communities with price disparity based on custom category selection. Further, category importance trends are charted year-on-year.

Lastly, application of various clustering approaches is described to enable identification of housing typology disparity in comparable communities. Density-based clustering also helps identify outlier communities with exceptional Vital Signs values.

The methods, insights, ongoing and future work discussed in this paper make a strong case for enabling interactive, visual exploration of geographical neighborhoods to enable informed decision making for home buyers, investors and policy makers alike.

References

[1] Baltimore City Planning Department, "2011 housing market typology." `https://data.baltimorecity.gov/Housing-Development/2011-Housing-Market-Typology/782b-zpd7`, 2016. Accessed: 2016-09-15.

[2] Baltimore Neighborhood Indicators Alliance - Jacob France Institute, "Vital signs 14." `https://www.bniajfi.org/vital_signs`, 2016. Accessed: 2016-09-15.

[3] Environmental Systems Research Institute, Inc., "Esri shapefile technical description," *Comput. Stat*, vol. 16, pp. 370–371, 1998.

[4] L. van der Maaten, G. Hinton, and Y. Bengio, "Visualizing data using t-sne," 2008.

[5] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," in *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[6] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pp. 849–856, MIT Press, 2001.

[7] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," pp. 226–231, AAAI Press, 1996.