EXTERNAL GEOGRAPHIC EFFECTS UPON REAL ESTATE PRICES

by

Reuben Hooley

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geoinformatics and Geospatial Intelligence

Committee:

____________________ Dr. Ruixin Yang, Thesis Director

____________________ Dr. Anthony Stefanidis, Committee Member

Kevin M. Curtin ____________________ Dr. Kevin M. Curtin, Committee Member

Sheryl L. Beach ____________________ Dr. Sheryl Beach, Department Chairperson

____________________ Dr. Richard Diecchio, Interim Associate Dean for Student and Academic Affairs, College of Science

____________________ Dr. Peggy Agouris, Acting Dean, College of Science

Date: December 5, 2013 ____________________

Fall Semester 2013
George Mason University
Fairfax, VA

External Geographic Effects upon Real Estate Prices

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

By

Reuben Hooley
Master of Science
George Mason University, 2013
Bachelor of Science
Oral Roberts University, 2004

Director: Ruixin Yang, Associate Professor
Geography and Geoinformation Science

Fall Semester 2013
George Mason University
Fairfax, VA

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

EXTERNAL GEOGRAPHIC EFFECTS UPON REAL ESTATE PRICES

Reuben Hooley, M.S.

George Mason University, 2013

Thesis Director: Dr. Ruixin Yang

This thesis evaluates the influence of external factors on real estate prices in Washington, DC. Most peer reviewed real estate studies compare internal factors of housing units like flooring or number of bathrooms rather than external influences like Interstate access. This study looks at nine different external factors in general categories of zoning regulations, proximity to undesirable environmental influences, proximity to desirable transportation conveniences and the influence of defined growth districts. This study generated measurable interval and ratio dimensions from data provided by the District of Columbia government for each of the nine variables and compared it to the most significant internal factor: price per floor space. This study establishes the significance levels for each of the nine dimensions with arithmetic mean trend analysis and regression analysis. Then through Pearson correlation, Apriori association analysis and Decision Stump classification, this thesis found that proximity to historical districts has the most significant influence on real estate prices.

1. INTRODUCTION

Real estate theory has developed multiple hypotheses about the influence on real estate prices. The well-known saying that location matters seems true, yet many cannot scientifically prove why. There are many statistical allusions as to why; however, statistics are generally combined with personal knowledge about a region. Often the inside knowledge and experience of realtors gives the best solution, since too often the curse of dimensionality changes the basis of statistical real estate analysis. This thesis uses a multiple layer data mining approach to analyze external influences on real estate.

Most studies within real estate analysis focus on the particularities of property features rather than the particularities of external influences. The distance a person has to commute to work often plays a role in his choice of a home. Furthermore, access to conveniences or to entertainment venues also plays a role. Many studies often avoid external factors. Far too often research is also done with too much estimation at a global statistical level. There are local house to house influences that cause price fluctuation and a statistical analysis should consider these fluctuations.

This thesis takes a holistic approach to analyze the external factors influencing real estate prices at a local house to house level. This thesis uses the most significant internal factor of price per area within a residential unit as the basis for the analysis. Within this thesis 1,753 different real estate properties within Washington, DC were analyzed. External factor influence on these properties are statistically analyzed throughout this thesis. The external factors are subdivided into five categories: zoning influence, pollutant influence, transportation influence, business influence and historical influence. Within the realm of these five categories, correlation, Apriori association and Decision Stump classification are used to show which dimension has the greatest influence on real estate pricing.

This thesis compares the five external influence categories above to the baseline dimension of price per area of a residential unit. Chapter 2 discusses current studies that have analyzed these external categories on an individual basis. Chapter 3 discusses external influence data within the five broad categories for the Washington, DC area. Chapter 4 discusses the detailed data transformation involved within this thesis as well as the data mining approach of this thesis. Chapter 5 evaluates the spatial and attribute distribution of the nine dimensions established from the five external categories discussed.

Chapter 6-8 analyzes the calculations from Chapter 5 with correlation, an Apriori association rule model and a Decision Stump classification model. These data are evaluated with Pearson correlation in Chapter 6 in a matrix of cross correlation measures. Chapter 7 uses Apriori Association Rule analysis to determine if there is any significant rules between multiple dimensions. A Decision Stump classification model in Chapter 8 evaluates each individual dimension's influence to determine which has the most influence on the independent variable of price per area. Thus, this approach utilizes data mining to evaluate the external dimension that has the most influence on property price.

2. LITERATURE REVIEW

Several techniques are used to analyze real estate data within geography. Most studies follow a housing centric perspective where actual housing variables (like number of bathrooms) were used to determine their effect on price. Huang and Kennedy (2008) used their Hidden Markov Model data mining approach to evaluate such parameters. They found significant results at the local house to house level with a very low error rate. However, their technique did not evaluate external factors. This thesis seeks to evaluate housing in a similar fashion in relation to external factors such as zoning, nearby hazards, transportation access, proximity to business centers and proximity to historical areas.

Land values are often affected by neighboring land values. It is difficult to analyze these factors at a theoretical level. Local statistic calculations on such factors are one approach. A study in Ontario Canada shows that spatial autocorrelation is a preferred method over Kriging in regards to spatial land price indexes since Kriging reduces the accuracy of the data so that it is only relevant at a large scale (Spinney and Scott 2011). Certainly, neighborhood price influences occur within a city due to the “quality or quantity of neighborhood amenities” (Spinney and Scott 2011). A neighborhood to neighborhood price comparison is necessary within a city’s housing price index.

Spatial variables could be logically analyzed within a spatial model; however, there is no consideration for non-spatial attributes. Huang and Kennedy (2008) utilized a “doubly embedded stochastic method based on probability theory” within their Hidden Markov Model in order to address these hidden factors that affect the spatial distribution of a land price index. Hidden factors could be modeled in a similar way within clustering applications that consider both spatial distances and non-spatial attributes as a series of attributes. Distance measurements between real estate properties could be considered within a model to enhance its spatial component.

Unlike most spatial statistical calculations where only one attribute is considered spatially, this thesis’s approach allows for a comparison of multiple spatial attributes. The spatial external variables are based on five subcategories: governmental zoning influences, undesirable factor influences, transportation conveniences, local business influences and the historical influences.

2.1. Governmental Zoning Influence on Real Estate Prices

Governmental manipulation of land use zones plays a role in the eventual property value of an area. Zoning law influence on the distribution of businesses and residential communities could have a positive or negative effect on housing prices. A study found considerable governmental influence in Beijing (Ding 2013). This thesis hypothesizes that US restrictions will have significant influence on the distribution of property value in the Washington, DC metropolitan area due to similar restrictions on building height and land use types.

Property value is affected by zoning laws. A paper on land value distribution found that Beijing's planned central business district had significantly higher prices, since during the last few decades the city has contracted many companies to build very elaborate apartment communities in that district (Han 2004). This created a significant rise in property value in northeastern Beijing between the second and third ring road (Han 2004). The distribution of property value within the Washington, DC area is expected to be affected in a similar way, since there are zoning restrictions on residential usage, commercial usage and industrial usage.

Floor area is a real estate dimension whose average price gradually changes as distance from the center of a city increases. Zoning laws are influenced by these trends; however, in some cases governments may want to change the natural distribution of housing characteristics within an area. With regard to floor space, one study found significant relation between distance from the center of a city and average floor space within their geographically weighted model (Helbich et al. 2013). This thesis attempts to further validate these suspicions through data mining techniques for the District of Columbia in order to determine the magnitude of zoning law effects.

Within real estate analysis, height limitations also affect housing price. A study on height restrictions in downtown Beijing showed a decrease in land value within the center of a city and an increase in value along the edges of a city (Ding 2013). This same phenomena will be tested with data for the Washington, DC area.

2.2. Noise and Pollution Influence on Real Estate Prices

Another external factor influencing home value is the proximity to environmentally hazardous facilities. The effect of hazardous facilities is difficult to determine due to multiple external factors. A study in France found correlation between hazardous facilities and economic deprivation according to the Townsend index over a provincial size region (Viel et al. 2011). However, a more detailed data correlation will generate more precise rules governing the distribution of population in regards to hazardous facilities. A study by census tract in New Jersey also derived a correspondence between areas with low average income and high density of polluting facilities (Mennis 2005). This thesis attempts to confirm this relation of distance from a hazardous facility and property price within Washington, DC.

Noise factors play a role as well—even when the nature of the noise provides minor transportation convenience. A study in Spain found that proximity to railways has an effect of about -4.9% of the price variable within their SAR-QUEEN model—confirming suspicions that the negative effects of noise levels will overtake the positive effects of transportation conveniences (Ibeas et al. 2012). Their result confirmed the results in earlier papers that found that noise from railways reduces the price of nearby properties. Within Washington, DC railroad presence does not equate to transportation convenience since there are very few access point to the railroad. This study analyzes railway noise influence on housing price for the District of Columbia.

2.3. Transportation Convenience Influence on Real Estate Prices

The relation of work access and land prices is covered significantly in peer reviewed literature. A study in Ireland utilized Newton-Raphson's optimization method to compare the work commute distance to a person's income level; this paper basically confirmed the bid rent theory hypothesis that distance to work inversely affects the amount of rent per housing unit area (O’Kelly and Niedzielski 2012). A paper in Slovenia determined there was significant correlation between the average land price in a district and that district’s access to improved roads (Lisec et al. 2008). Distance to improved roads gives insight to the average commuting time for people who live in the area.

A study in Tokyo utilized Kriging to show the effects of commuter railways on land price from a multivariate perspective. Within the study's Geographically Weighted Regression (GWR) model, the authors derived that accessibility to railroads had a direct effect upon land use type and population density within the area (Tsutsumi et al. 2011). Within this thesis, Metrorail in Washington, DC is analyzed.

The model by Tsutsumi et al. (2011) did have a slightly high error rate (18.3% on average); however, this is likely due to the spatial level of Kriging utilized. Unlike the study by Ibeas et al. (2012) where proximity to railroads had a negative effect on price due to noise, railways in Japan provide a significant transportation convenience. This positive shift in land pricing effect is likely caused by the accessibility and interconnectedness of commuter railways and subway systems in Japan. These positive and negative influences are further evaluated in this thesis on the District Columbia’s Metrorail and commercial railway. These transportation conveniences likely overtake the negative effects from noise.

2.4. BID Influence on Real Estate Prices

The analysis of business locations provides a twofold effect in real estate purchases. Work accessibility is naturally driven by the locations of business centers. Also, the accessibility to consumer goods and social services is driven by locations of business centers in a general sense. Investment in an area is also revealed indirectly—showing possible corporate manipulation of the economic growth of an area, if it occurs.

Recreation areas reveal the sociological influence upon geography. For example, an area’s attraction, governmental policy, land supply and popular demand were used to evaluate the distribution of “peri-urban” recreation areas in Beijing (Liu et al. 2010). This categorization shows that classifying different types of recreations facilities in an ordinal manner allows for enhanced spatial analysis information from generalization methods.

Access to social areas have an influence on the distribution of population within a city as well. For example, a paper analyzing travel routes between a traveler's foursquare "check-ins" and their home found that lengthy routes were better explained statistically (Noulas et al. 2012). The author’s notion of rank distribution with regards to travel routes is not the best descriptive statistic for this type of analysis. A version of spatial autocorrelation or correlation may give more significant findings. This thesis tests business district access within Washington, DC where most of these social areas are likely located.

Also, capital infusion in an area has an effect upon real estate development and population distribution. A paper studying countries within the eastern bloc in Europe found a general trend between the location centrality and investment from neighboring regions (Petrakos 2001). The study found that there was westerly-easterly trend within the analysis, where it appeared that foreign investment within the western part of a nation eventually inspired investment in the eastern part of the nation (Petrakos 2001). This thesis tests business district distribution and its effect upon real estate price within the District of Columbia.

Business distribution influences commuter destinations as well as access to goods and services. A study in Spain found a significant relation between distance to the Central Business Unit (CBU) of the city and property price; there was a reduction of 0.5-1.1% of price per minute of travel time (Ibeas et al. 2012). Accessibility to business districts is a factor that is further researched within this thesis for Washington, DC.

2.5. Real Estate Prices within Historical Areas

Within real estate development, the proximity to historical areas is always a factor. These areas are usually remodeled and visually appealing, thus increasing the price value of those areas. Historical refurbishing of an area is pleasing and often drives property value up. Han (2004) found there was a significant price increase in the historic quarter in Jakarta according to his interpolated TIN model. He also found that in all of the cities he analyzed, there was low price fluctuation within the historic area (Han 2004). This thesis looks at historical district data in Washington, DC.

Time also plays a factor in real estate development. A study in Savannah, Georgia also found that home price has a tendency to increase when the age of the home is beyond a century (Winson-Geideman et al. 2011). Certainly with regards to historicity, there is an effect upon population distribution as well as price.

2.6. Categories and Methods Utilized

This literature review revealed five general categories of external pricing influences. They are zoning regulations, negative environmental influences, transportation conveniences, business district accessibility and historic area proximity. Of the methods utilized to analyze spatial interaction, Huang and Kennedy’s (2008) notion of hidden spatial factors within a Hidden Markov Model seemed the most intuitive since it considers multiple distance dimensions. This thesis plans to research the five external factors discovered with Washington, DC data with a model that explores multiple distance attributes.

3. DATA

This thesis evaluates real estate listing data for the Washington, DC area. These listings are compared with the five external factors of governmental zoning influences, negative environmental influences, transportations conveniences, business district influences and historical district influences.

Data for the external comparison layers are provided by District of Columbia Office of the Chief Technology Officer (http://octo.dc.gov/DC/OCTO/). Their website provides shapefiles for the current location of police stations, schools, and facilities with toxic waste. They also have current road, railway and Metrorail station shapefiles as well as area shapefiles showing historical districts, business districts and zoning details. Current real estate listings were downloaded from www.realator.com in the Washington, DC area; there were 1,753 listings. All of this data was acquired on June 15, 2013.

The listing price, square footage of the real estate unit and location were downloaded for each of the 1,753 real estate listings. In order to create a single internal dimension for analysis, the price was divided by the square footage of the unit giving the price per area attribute as shown in Table 1.

Table 1. Real Estate Pricing Example.

Lat	Lon	Price	Sq Ft	Price Per Sq Ft
38.9006	-77.0114	$427,165	1740	$245.50
38.9027	-77.0184	$913,961	1890	$483.58
38.9009	-77.0135	$464,640	780	$595.69
38.9079	-77.0190	$282,601	1410	$200.43
38.9076	-77.0128	$820,414	1990	$412.27
38.9061	-77.0149	$294,006	980	$300.01

The zoning data for this study consisted of a polygon spatial data structure with zoning keys similar to that shown in Figure 1 and zoning descriptions from a separate table with the same zoning keys as shown in Table 2. Table 2 shows an example of how three attributes were generated based upon the zoning description. An attribute for the Floor Area Ratio (FAR), maximum allowable height and maximum residential usage were generated.

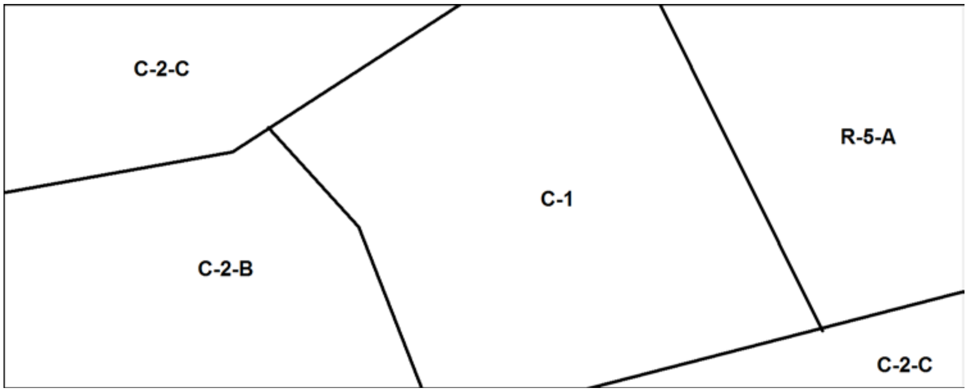


Figure 1. Zoning Spatial Boundaries Example.

The Washington, DC zoning laws are influenced by an extra factor. Height of a building is restricted generally to the length of the street segment in front of it due to the Heights of Buildings Act of 1910. This dimension was added as an extra attribute to consider along with FAR and allowed maximum residential usage within Table 2. According to the Washington, DC Office of Zoning, areas are limited to certain types of development based on a sliding scale. A zoning law will clarify whether an area is allowed 80 percent residential usage and 100 percent all other usages.

It should be noted that generally there are three types of maximum allowable zoning usages that this study encountered in the District of Columbia: allowable residential usage, allowable commercial usage and allowable industrial usage. These percentages often overlap to allow for building flexibility. For example, some zones may have a maximum residential usage of a 100% and a commercial usage of 60%. Washington, DC has 35 complex zoning definitions; however, all of the zones encountered for properties within this analysis have a maximum residential usage, a maximum FAR and a maximum building height.

Obviously, much like zoning influence in Beijing discussed earlier, zoning preferences will have considerable influence on property price in Washington, DC. Maximum residential usage amounts, building height restrictions and floor area restrictions are transposed from the actual zone regulation for each District of Columbia zone as shown in Table 2. These attributes are related to the actual spatial data structure in Figure 1, which is used further in this thesis.

Table 2. Zoning Regulations and Corresponding Attributes Example.

Zone Code	Zone Regulation	FAR Limit	Height Limit	Residential Usage %
C-1	Permits matter-of-right neighborhood retail and personal service establishments and certain youth residential care homes and community residence facilities to a maximum lot occupancy of 60% for residential use and 100% for all other uses, a maximum FAR of 1.0, and a maximum height of three (3) stories/forty (40) feet. Rear yard requirements are twenty (20) feet; one family detached dwellings and one family semi-detached dwellings side yard requirements are eight (8) feet.	1	40 ft	60%
C-2-B	Permits matter-of-right medium density development, including office, retail, housing, and mixed uses to a maximum lot occupancy of 80% for residential use and 100% for all other uses, a maximum FAR of 3.5 for residential use and 1.5 FAR for other permitted uses, and a maximum height of sixty-five (65) feet. Rear yard requirements are fifteen (15) feet; one family detached dwellings and one family semi-detached dwellings side yard requirements are eight (8) feet.	3.5	65 ft	80%
C-2-C	Permits matter-of-right higher density development, including office, retail, housing, and mixed uses to a maximum lot occupancy of 80% for residential use and 100% for all other uses, a maximum FAR of 6.0 for residential and 2.0 FAR for other permitted uses, and a maximum height of ninety (90) feet. Rear yard requirements are fifteen (15) feet; one family detached dwellings one family semi-detached dwellings side yard requirements are eight (8) feet.	6	90 ft	80%

Data for railroads were downloaded from the Washington, DC GIS website as well. The data consisted of a linear topographic data structure of the centerline of all railroads in Washington, DC. A series of 500m distance rings were generated around the Railroad centerlines throughout all of Washington, DC in Figure 2 below. This ring layer will be used for distance analytics further in the thesis.

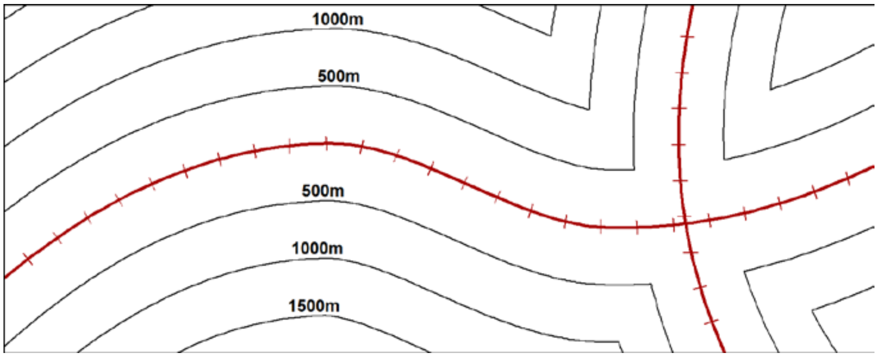


Figure 2. Railroad Distance Rings Example.

The hazardous facilities tracked by the EPA within Washington, DC were downloaded as well from the District of Columbia government website. The spatial data structure had geographic coordinates as well as information about the particular site as shown in Figure 3. This thesis considers the distance between a real estate property and these facilities as separate dimension—these distance calculations will be discussed later.

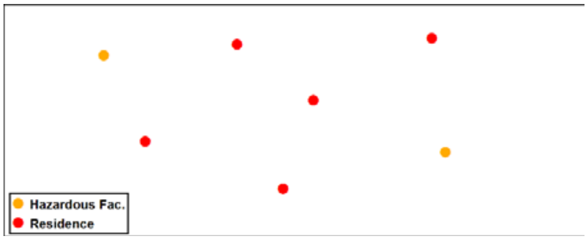


Figure 3. Hazardous Facility Distribution Example.

The distance to Metrorail can also be measured by the distance to tunnel entrances. A polygon spatial data structure representing the tunnel entrances for Washington, DC Metrorail was downloaded from the District of Columbia website. Since the polygons representing these tunnels were relatively small in comparison to the size of the District of Columbia, the centroid of each polygon was generated for each Metrorail tunnel as shown in Figure 4. This data will be compared with the locations of real estate properties later in this study.

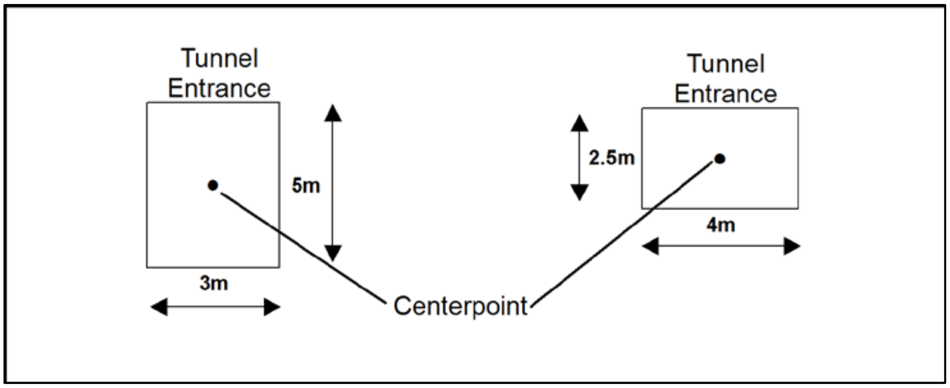


Figure 4. Metrorail Tunnel Example.

Topological data for the US interstates was downloaded from the Washington, DC government website. Polygonal distance ranges were generated from this data in a series of 500m intervals in a way similar to the example in Figure 5. This data will be compared to housing data later in this thesis.

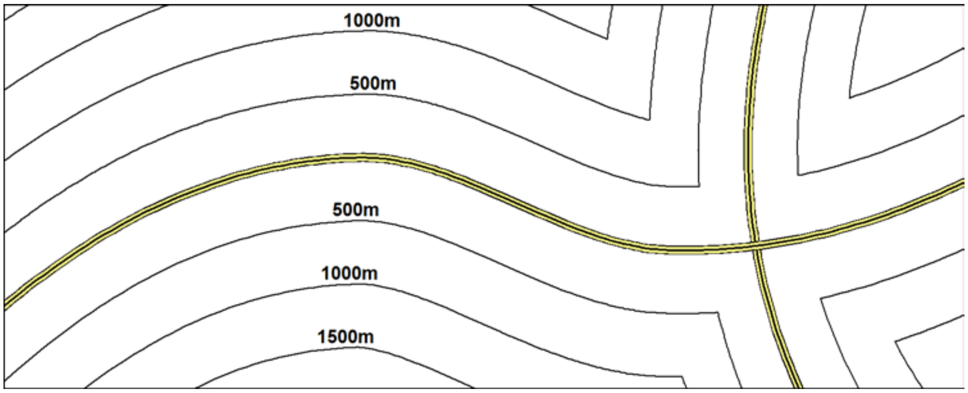


Figure 5. Distance from Interstates Example.

Official District of Columbia business district and historical district boundaries were downloaded from their website. Polygonal distance rings were generated from the business and historical districts separately at 500m intervals. An example is shown in Figure 6. These distance ranges will be compared with the housing data downloaded later in the thesis.

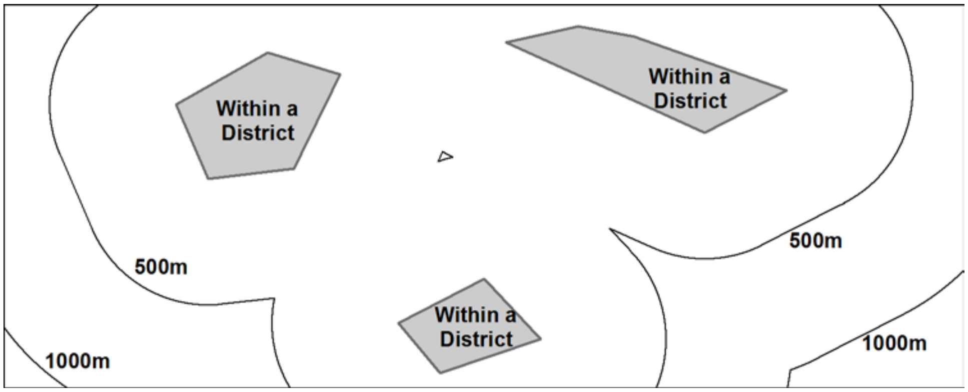


Figure 6. Business and Historical District Distance Rings Example.

This study compares nine dependent variables from the five categories of zoning laws, undesirable influences, transportation conveniences, business influences and historicity to an independent variable based upon price per area of the unit. According to Huang and Kennedy's (2008) analysis, unit floor space is one of the greatest factors on real estate price as opposed to lesser things such as access to a garage. An external dimension analysis derived upon this one internal dimension gives the most accurate external factor comparison possible between multiple units with different features.

4. METHODS

Throughout this analysis, the nine external factors are compared to price per area. They are compared based upon the geographic nature of each external factor. These nine factors fit within the five categories mentioned in Chapter 2. This thesis evaluates the zoning dimensions of FAR, building height and residential usage; the undesirable influences of railroads and hazardous facilities, the transportation influences of Metrorail and interstates; the distance to business districts; and the distance to historical districts.

4.1. Study Limitations

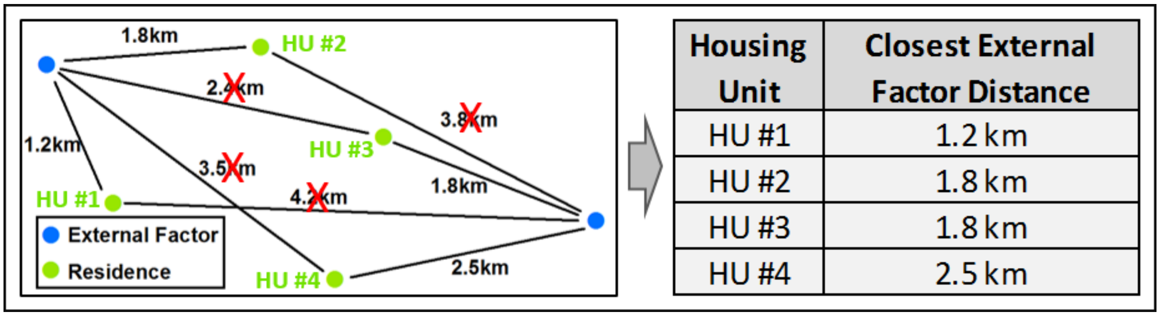
This study recognizes that there are some limitations by using only external influences to evaluate property price; however, this thesis does use the most influential internal dimension found in real estate analysis of property price per floor area (Huang and Kennedy 2008). This study also incorporates a large dataset with a large sampling distribution that represents varying internal dimensions such as the existence of a garage or the number of bathrooms.

This study also focuses on the most static external dimensions. Proximity to good schools or bad crime areas was ruled out due to the variable nature of these dimensions. Evaluation of locational relation to good schools is verbosely vague, since it would have to evaluate the question of what determines a good school. Also, crime statistics of an area vary from year to year.

4.2. Methods for Each of the Four Data Types

Within each of these five categories there are four different types of external influence data that are treated the same way throughout this thesis: point data, line data, polygonal data that encompasses all the housing locations and polygonal data that does not encompass all the housing locations.

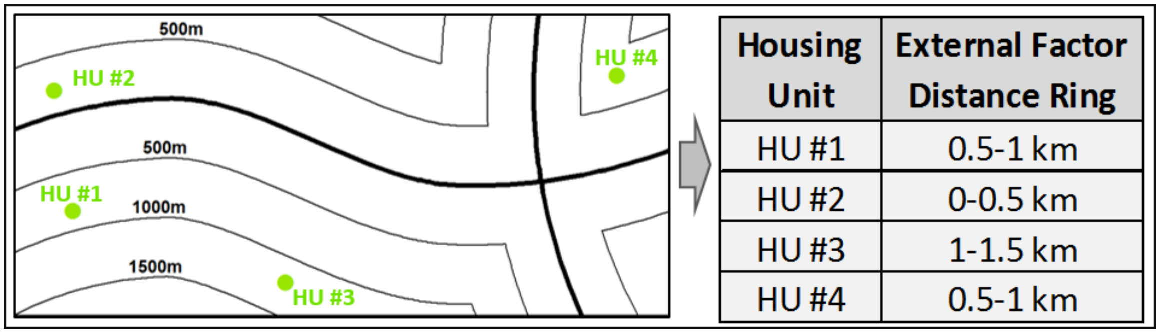
Exact distance was calculated between point information and housing locations within an equidistant projection. This was done for the Metrorail Station comparison to housing units and the hazardous facility comparison to housing units. Lines between all point locations and each housing location were generated, the distance was calculated, and within a database the distance information was convolved to the lowest distance found by housing location. As shown in Figure 7, only the closest distance between an external feature and a housing unit was added as an attribute within the database.



Housing Unit	Closest External Factor Distance
HU #1	1.2 km
HU #2	1.8 km
HU #3	1.8 km
HU #4	2.5 km

Figure 7. Closest Facility Distance Calculation.

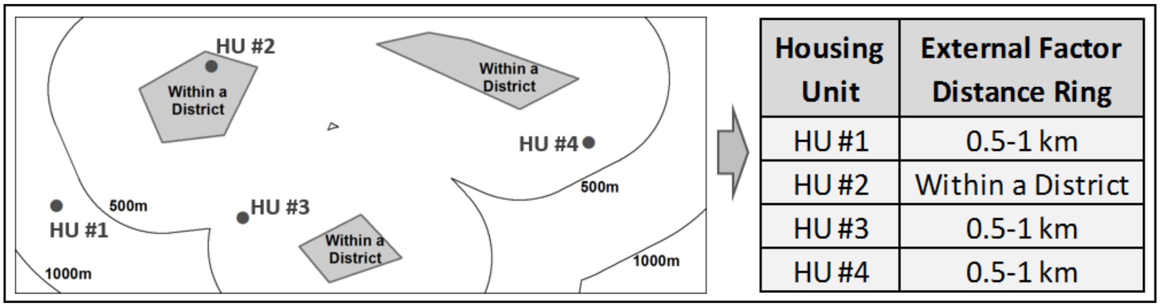
As discussed in Chapter 3, distance zones were created for line data and polygon data that did not encompass the whole map. These distance polygons were generated at 500 m intervals. These distance polygons were intersected with the housing information point layer so that the distance range information could be joined to the housing information for further analysis within a table. Figure 8 shows how the 500 m distance rings for Interstates and Railways were added to a database through a spatial intersect operation in ArcGIS.



Housing Unit	External Factor Distance Ring
HU #1	0.5-1 km
HU #2	0-0.5 km
HU #3	1-1.5 km
HU #4	0.5-1 km

Figure 8. Polyline Features Distance Rings Calculation.

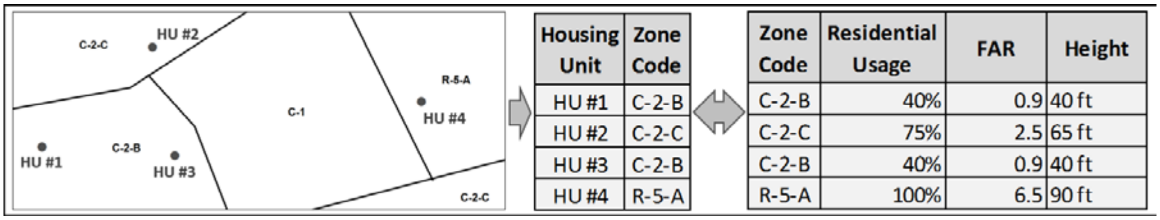
Distance zones were created around the polygonal data that were not zoning layers as discussed in Chapter 3. These external factors are business districts and historical districts. A distance attribute of zero was given for values that were within the polygon, whereas the corresponding numerical distance value was given for each of the ranges. The numerical values for the polygonal category are treated as interval data, so an arbitrary zero value will fit within a data mining model. Figure 9 shows how a spatial intersect operation was performed between the distance rings of the polygonal data and the actual housing units.



Housing Unit	External Factor Distance Ring
HU #1	0.5-1 km
HU #2	Within a District
HU #3	0.5-1 km
HU #4	0.5-1 km

Figure 9. Polygon Features Distance Rings Calculation.

Lastly for data within the polygonal category that encompasses all the housing locations, the polygon layer was intersected with the housing information layer so that the polygonal attribute could be joined with the housing information within a data table. This was done for the three zoning law attributes of FAR, Residential Usage and Height. Below Figure 10 highlights this spatial relationship.



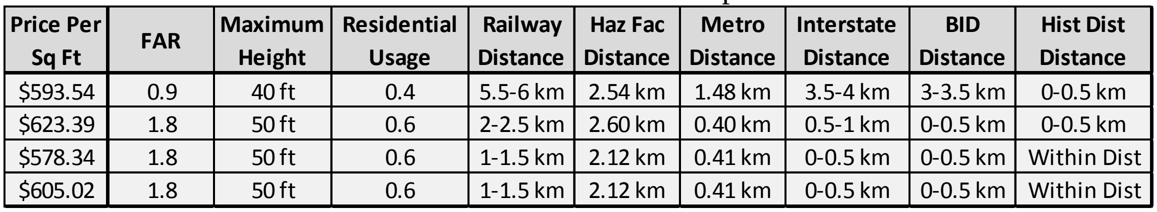
Housing Unit	Zone Code
HU #1	C-2-B
HU #2	C-2-C
HU #3	C-2-B
HU #4	R-5-A

Zone Code	Residential Usage	FAR	Height
C-2-B	40%	0.9	40 ft
C-2-C	75%	2.5	65 ft
C-2-B	40%	0.9	40 ft
R-5-A	100%	6.5	90 ft

Figure 10. Zoning Spatial Intersection Method.

Once all these preliminary calculations are completed, the data appears as such in Table 3 below. These spatial dimensions allow data mining principles to discover interrelations within the data. The data in the first column of Table 3 represents the price per area of a real estate unit. It serves as the base dimension for multiple analysis processes discussed in this thesis. The other nine columns represent the spatial transform of the data discussed earlier. Namely, they represent the measure of the external influence of the corresponding feature on the particular real estate unit.

Table 3. Data Table Example.



Price Per Sq Ft	FAR	Maximum Height	Residential Usage	Railway Distance	Haz Fac Distance	Metro Distance	Interstate Distance	BID Distance	Hist Dist Distance
$593.54	0.9	40 ft	0.4	5.5-6 km	2.54 km	1.48 km	3.5-4 km	3-3.5 km	0-0.5 km
$623.39	1.8	50 ft	0.6	2-2.5 km	2.60 km	0.40 km	0.5-1 km	0-0.5 km	0-0.5 km
$578.34	1.8	50 ft	0.6	1-1.5 km	2.12 km	0.41 km	0-0.5 km	0-0.5 km	Within Dist
$605.02	1.8	50 ft	0.6	1-1.5 km	2.12 km	0.41 km	0-0.5 km	0-0.5 km	Within Dist

4.3. Initial Categorization Methods

This thesis is based on a series of analyses that build upon one another. The methods for these analysis layers are explained here to reiterate the purpose of each layer. The first layer of analysis considers the data distribution of each of the nine dimensions that are being studied. The first layer of analysis also considers the spatial distribution of the data and its relation to the original external feature that is being studied.

Regression or the arithmetic mean for each categorical unique value are used for each data series to determine the trend. If a category like FAR only had five possible values like 1.0, 2.5, etc, the arithmetic mean was used within each category to determine trend. This is beneficial over regression when it is possible since it is deterministic. This method is used for seven dimensions measured by discrete values and regression is used for the two dimensions measured with continuous values. Regression was used for the minimum distance to a Metrorail station or a hazardous facility categories. Regression was needed for these two categories since the data was virtually continuous, so a stochastic method was necessary. The continuous data method was utilized to reduce error within the first layer of analysis. Figure 11 shows an example of the difference in categorization between the two methods.

Trend analysis is conducted on the distribution of the nine dimensions to determine if there is any observable decrease or increase of average price values as the dimension value increase. For the two dimensions analyzed with regression, an inflection of the quadratic regression trend line was observed. The quadratic form of regression was chosen over the linear and cubic forms since the quadratic curve form fit the data better in both cases.

Within the trend analysis section, a significant distance (Or value) was determined. In most categories, this was decided at the first inflection point within the average values as the distance or value increases. The significant distance was determined with the inflection of the regression curve as well. Figure 12, shows this in further detail.

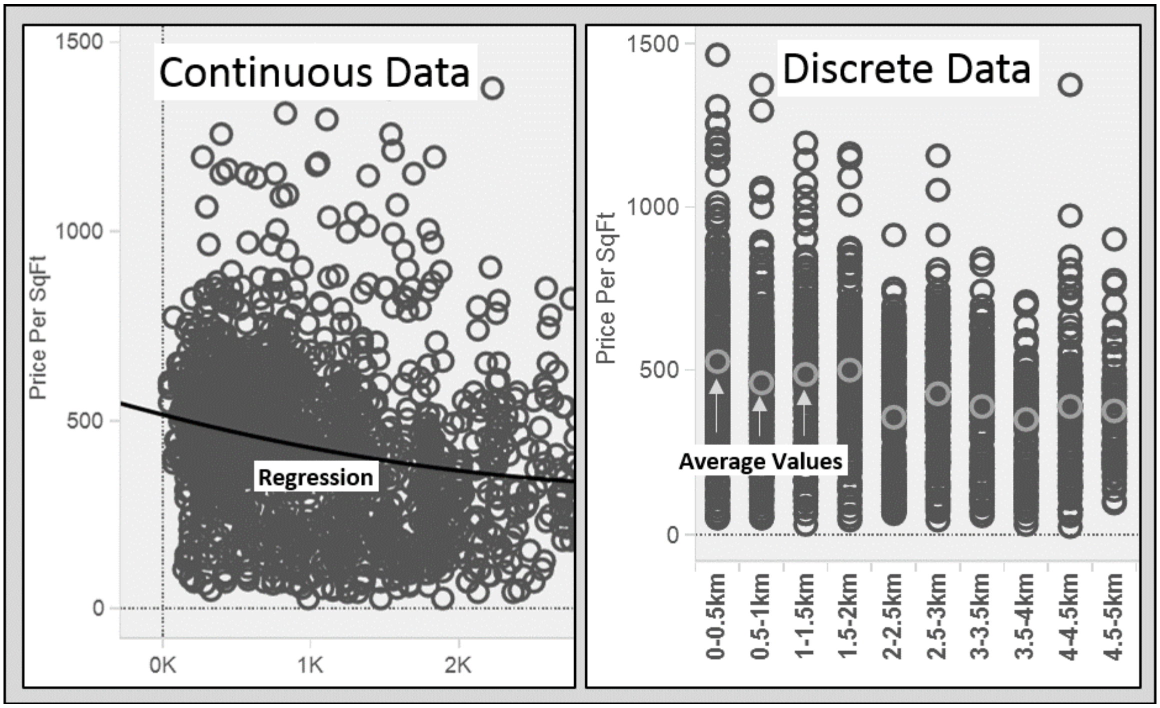
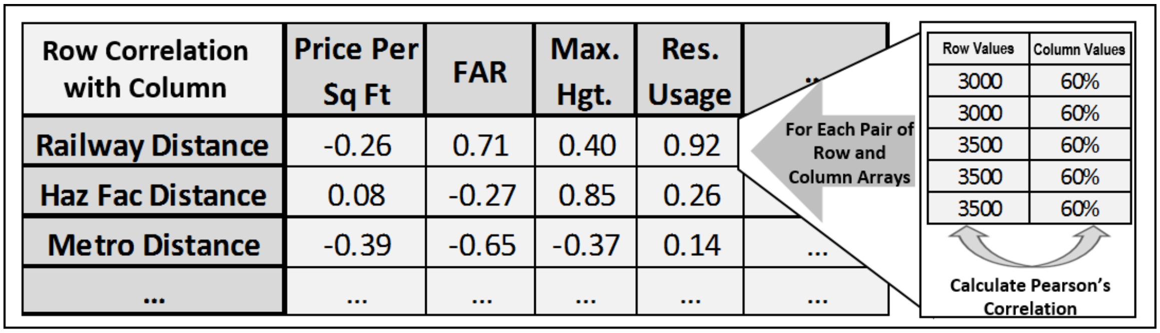


Figure 11. Data Distribution Example.

The significant influence distance will be needed for the next stage of analysis since the upper bound distance may affect the true inherit environment rule in relation to housing price. There may be instances where an external factor like railroads may be miles away from the property in question, so the rule attempting to be validated by the model would no longer be applicable.

The next layer of analysis utilizes Pearson correlation to minimize the cross dimensional influence within the nine dimensions being analyzed. A secondary objective of correlation analysis is to determine which of the nine layers has the greatest correlation magnitude; however, the direct relation will be analyzed later in the third analysis level with data mining models. This stage of analysis was also conducted in two phases. The first stage looks at all global values for the data, whereas the second considers only the significant distances (or values). These two methods consider the correlation between every pair of dimensions utilized within this analysis within a cross correlation matrix similar to Figure 12.



Row Correlation with Column	Price Per Sq Ft	FAR	Max. Hgt.	Res. Usage	...
Railway Distance	-0.26	0.71	0.40	0.92	
Haz Fac Distance	0.08	-0.27	0.85	0.26	
Metro Distance	-0.39	-0.65	-0.37	0.14	...
...	...	...	...	...	...

Row Values	Column Values
3000	60%
3000	60%
3500	60%
3500	60%
3500	60%

For Each Pair of Row and Column Arrays
Calculate Pearson's Correlation

Figure 12. Cross Correlation Matrix: Correlation between Each Dimension Pair (The Inset Shows the Arrays of Values for each Dimension that is being Correlated)

The global level of correlation analysis determines whether there are any dimensions that have a significant relation, so that the spurious dimensions can be negated within the data mining processes that follow. The local level of correlation analysis with significant distances considers a second aspect of possible interrelation between the dimensions.

If a series of properties were within a close distance of one of the external dimensions like hazardous facilities and if that same series of properties were in close proximity to say railroads, then local correlation evaluation would consider railway and hazardous facilities proximity to each other in only the cases of that particular housing subset. In these series of local correlation subsets, there may be a relation that should be considered for the larger model.

The third layers of analysis involve two different data mining principles. A two pronged data mining approach was used in this thesis in order to determine if there is any significance from a combination of external dimensions and from a single external dimension. A breath first search algorithm known as the Apriori method was used to analyze multiple dimensions. This model is commonly used to derive the marketing probably of something like the chances a family will buy beer when they also buy milk and diapers. A decision tree model that only goes to one leaf level known as the Decision Stump method was used to analyze single dimensional influence within the data.

Multiple dimension data mining in this context considers the idea, for example, that if there is a significant number of condos that happen to have a FAR of 6.5 as well as a distance between 0.5 and 1.0 km to a Metrorail station that happen to be within the same pricing range (like $350 - $450 per square feet). This significance is tested with Apriori association rules in Waikato Environment for Knowledge Analysis (WEKA). For a single external dimension, Decision Stump classification was utilized within WEKA to determine the most significant data break.

The Apriori association rule model requires that the continuous dimensions were first generalized into categories such as 500m distance ranges similar to those of the other dimensions. That is, all the data were converted to nominal data with a filter within WEKA. Then it essentially iterates through all possible combinations of dimension values and returns the probability of when certain values are true. Some outlier property prices will be excluded for this portion of the model; however, the outliers are considered in the Decision Stump model.

The flow of the Apriori Association model is essentially equivalent iterating through an itemset of the pricing categories and each of the external dimensions and calculating the support for each. The ratio of an itemset value within an itemset is known as support in data mining. If the selection of a pair of dimensions had a high confidence that was more expected from the total support for a subset of those dimensions, then it was considered significant. If the pair of dimensions were say the number of apples and oranges, the confidence of the rule in this case would be the ratio where there were say 2 apples and 3 oranges within 50 different customer shopping bags. Confidence in this case is simply the support where both values occur. The model in WEKA in the thesis will need to use a very low minimum support and minimum confidence metric in order to evaluate second and third less significant confidence rules within the data. This secondary evaluation was utilized to incorporate rules that are significant over multiple dimensions.

It should be noted that the Apriori model not only iterates through every dimension, it iterates through every possible subset of 2 to 10 dimensions as well. There are also some considerations to be taken into account for false positives in situations where a more significant rule created a less significant rule that considered more dimensions. These rules will be ruled out so that only the most significant rule is shown for the final result. Also, correlation is considered for this model. If one dimension is highly correlated to another dimension within the model, then it is very likely that two rules would be generated with similar percentages. The lesser significant rule was ruled out in this case as well.

The Decision Stump model provides a second data mining approach to the data. This approach considers all the original data elements unlike the Apriori approach. Decision Stump looks for the most significant division within the data. This thesis utilized this method on the entire dataset in order to find the most significant dimensional division out of the nine dimensions. Then it further evaluates the rest of the dimensions with the most significant dimension and follow on significant dimension removed. In this way the progressive dimension significance is found with accuracy rate.

Lastly, the results from the two data mining approaches are evaluated geographically to double check for any analogous results. Thus, the approach of this thesis involves three layers of analysis with a two prong data mining approach within the third layer. Figure 13 shows the methodology flow of this thesis.

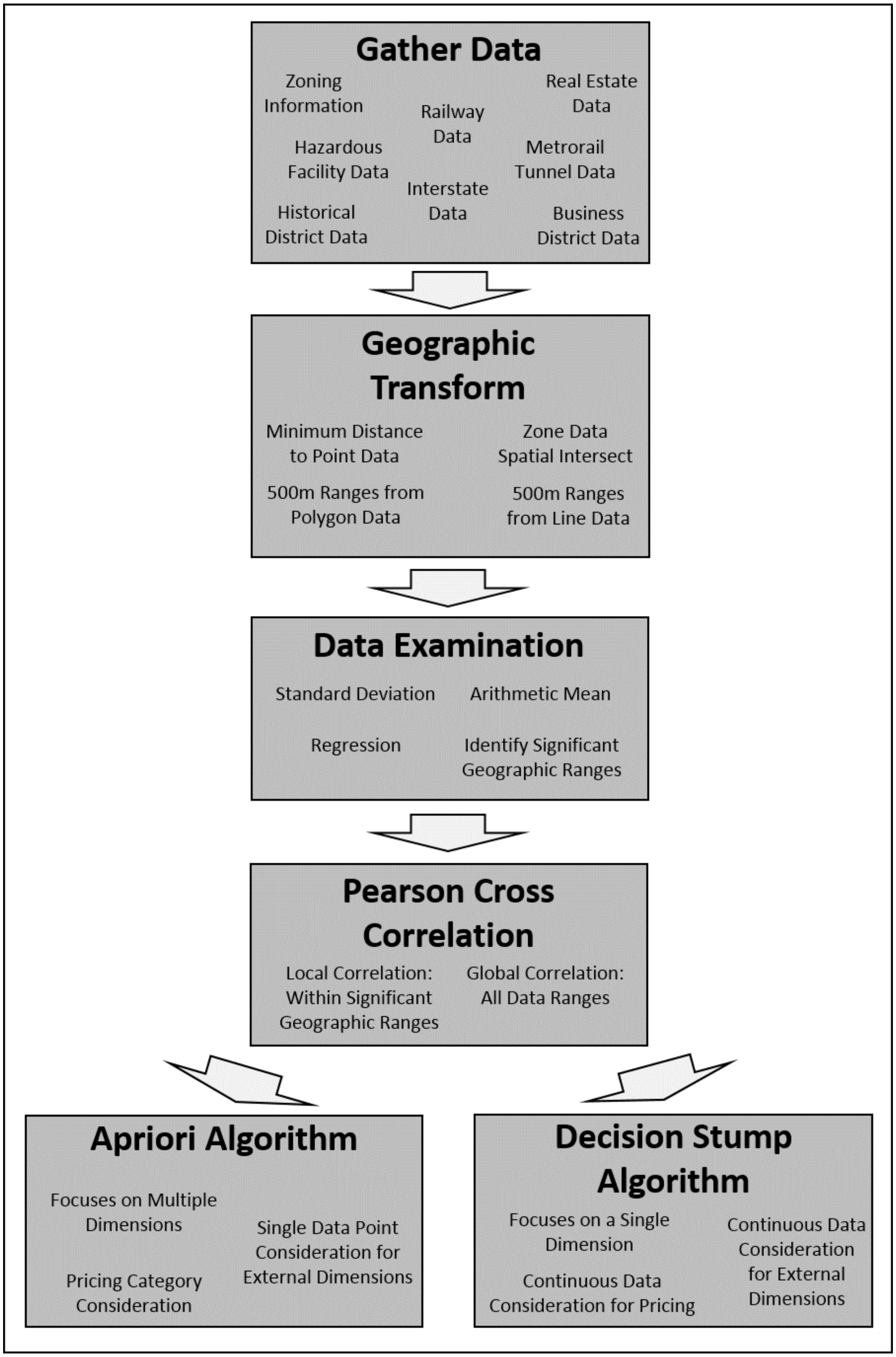

Gather Data
Zoning Information
Real Estate Data
Railway Data
Hazardous Facility Data
Metrorail Tunnel Data
Interstate Data
Historical District Data
Business District Data
Geographic Transform
Minimum Distance to Point Data
Zone Data Spatial Intersect
500m Ranges from Polygon Data
500m Ranges from Line Data
Data Examination
Standard Deviation
Arithmetic Mean
Regression
Identify Significant Geographic Ranges
Pearson Cross Correlation
Local Correlation: Within Significant Geographic Ranges
Global Correlation: All Data Ranges
Apriori Algorithm
Focuses on Multiple Dimensions
Single Data Point Consideration for External Dimensions
Pricing Category Consideration
Decision Stump Algorithm
Focuses on a Single Dimension
Continuous Data Consideration for External Dimensions
Continuous Data Consideration for Pricing

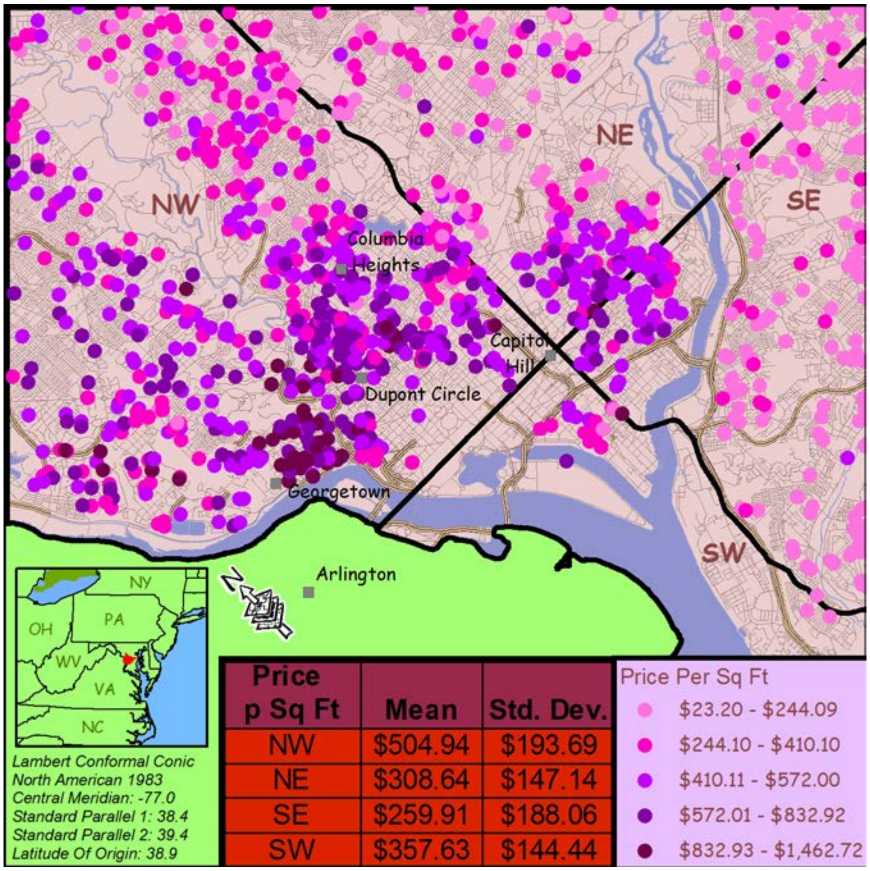
Figure 13. Data Process Flow Chart.

5. EVALUATION OF THE FIVE CATEGORIES

The nine layers generated for these five categories of analysis should have significant relation with the average price per square foot of real estate. The five influences governing these layers all have significant contribution from the academic community in regards to property price distribution. This Chapter's analysis determines the significance levels of these layers to find which has the most influence in the District of Columbia.

Some of the external influence dimensions should have a direct relation to price whereas others should have a negative relation. Floor area ratio, height restrictions, maximum allowable residential usage, distance from noisy railways and distance from hazardous facilities should be inversely related to property value, whereas the average distance to Metrorail, Interstates, business areas and historical areas should be directly related to property value. This thesis evaluates these criteria from peer reviewed literature with data for Washington, DC.

An initial look at the distribution of the property prices per area in Figure 14 shows high price clustering near Georgetown, DuPont Circle and Columbia Heights. There are also significantly higher prices around Capitol Hill. Properties within the northwest sector of Washington, DC are clearly more expensive; their average price is almost at least $150 more per square foot than any other sector of Washington, DC.



Price p Sq Ft	Mean	Std. Dev.
NW	$504.94	$193.69
NE	$308.64	$147.14
SE	$259.91	$188.06
SW	$357.63	$144.44

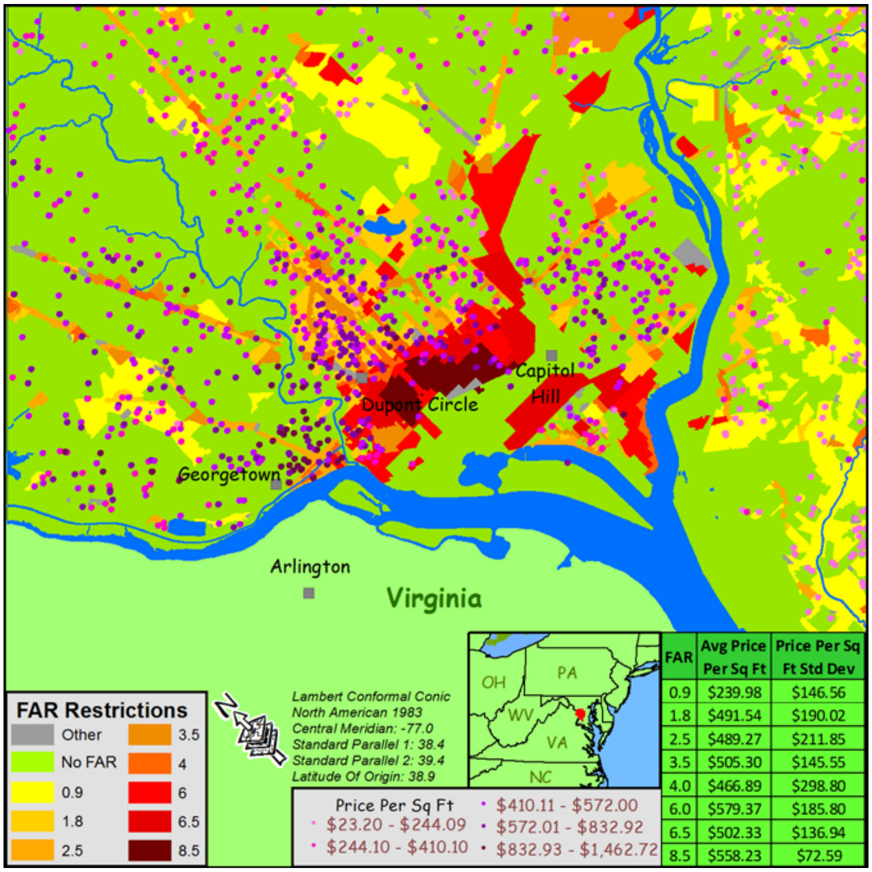
Figure 14. Price per Unit Area Geographic Distribution.

This Chapter of analysis looks at the geographic distribution of real estate properties within each of the nine dimensions. Furthermore, the pricing distribution is analyzed from a statistical standpoint. In each section there will be an evaluation of the significant distance (or value) range of the particular dimension under question.

5.1. Floor Area Ratio Categorization and Initial Statistics

Within Washington, DC some zones have a Floor Area Ratio (FAR), which is the ratio of the number of square feet within the building versus the amount of land used for that building. It was included within this analysis, though there are some units that do not have a FAR limitation. For the purposes of this section, only properties with FAR values were evaluated. This limitation in many story buildings limits the number of units allowed within a building and can affect real estate pricing.

Figure 15 shows the statistical calculations of the housing prices based on seven different ranges of FAR values. FAR restrictions affect properties when a restriction of 1.8 and higher is induced. This higher than 1.0 FAR restriction is generally for multiple floor buildings whereas the 0.9 FAR restriction areas occur on parcels that have more open land—such as single family home communities. Though FAR likely does not affect single family housing, 0.9 FAR homes were still included within this analysis since it limits construction of multistory buildings within those zones.



FAR	Avg Price Per Sq Ft	Price Per Sq Ft Std Dev
0.9	$239.98	$146.56
1.8	$491.54	$190.02
2.5	$489.27	$211.85
3.5	$505.30	$145.55
4.0	$466.89	$298.80
6.0	$579.37	$185.80
6.5	$502.33	$136.94
8.5	$558.23	$72.59

Figure 15. Housing Price Constraints from FAR.

Figure 16 shows the pricing distribution for each FAR category. Zones indicating 1.0, 3.0 or 5.0 FAR were ruled out for this portion of analysis since they were not represented by many homes within the data; however, the data are still considered by the correlation and data mining sections. This category was one of two categories that did not have a significant trend. A person could argue that the distribution shows a slight increase in average value for properties with high FAR values (6.0, 6.5 and 8.5); however, the range of values within each FAR value category indicate there are likely other factors effecting pricing.

Since all this data is garnered from zoning laws, the maximum allowed height for a building or maximum allowed residential usage may be the more significant dimension. Though the distribution of price oscillates, there is an observable increase in price as FAR increases—especially when a few minor values for FAR of 1.0, 3.0 and 5.0 are ruled out due to insufficient data. No holistic trend appears within the data though.

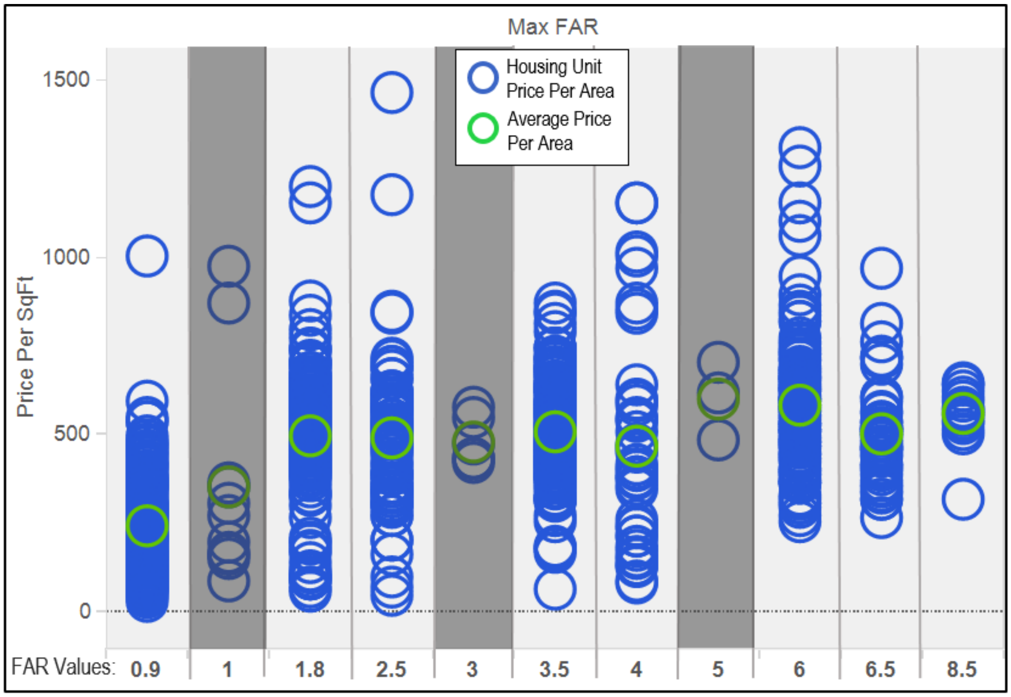


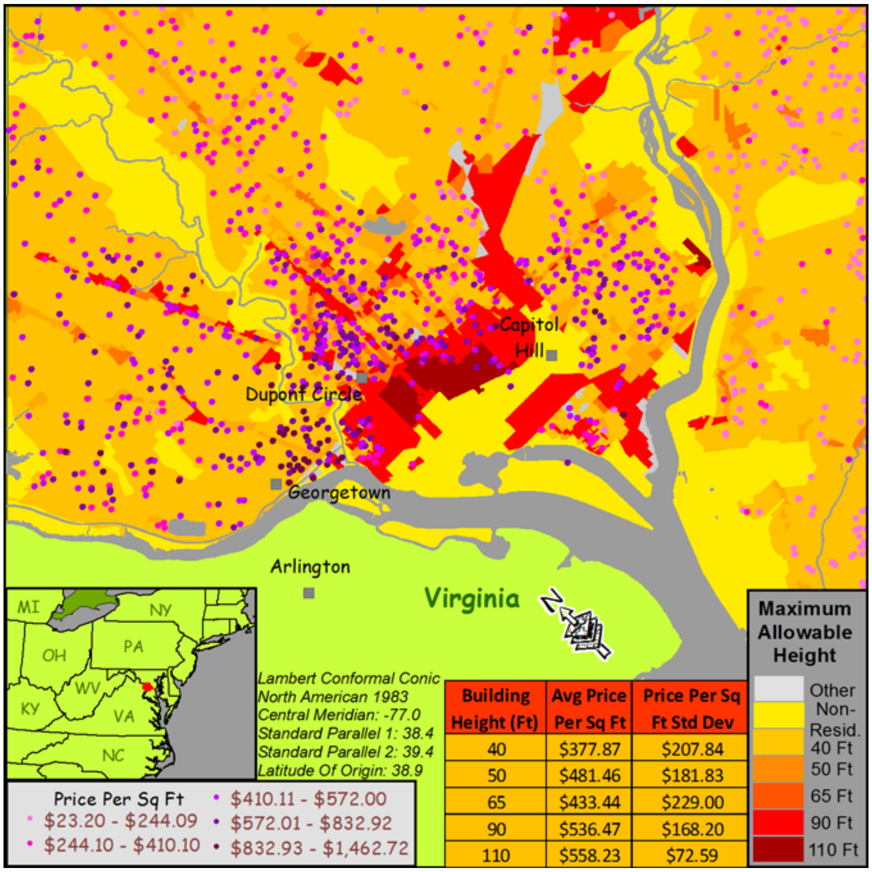
Figure 16. Pricing Distribution from FAR.

The relation of this category with other dimensions studied may reveal hidden information about a combination of dimensions within the Association analysis portion of this thesis, since Association rules attempt to find valuation from a group of dimensions as well as single dimensions. This dimension will be further analyzed in the next few Chapters—though the lack of FAR data for all units could cause some issues.

5.2. Height Restriction Categorization and Initial Statistics

The characteristics of the multiple zones used within Washington, DC have many variations. The most notable distinction is the restriction of building height. In general, a building cannot be taller than the length of the street block in front of it; this limits the building height to 110 ft throughout the district. Also, within this portion of analysis there was a small representation for residential units with 45, 60 and 70 ft height restrictions—thus, they were not included in the attempt to determine significant height ranges. This data are still considered in the correlation and data mining sections of this thesis however.

The spatial representation of housing values in Figure 17 indicate that zones representing the 40 ft category incorporate much of the land in the District of Columbia. Since the 40 ft category contains single family homes and townhouses whereas the 50, 65 and 90 ft categories allow for high rise buildings, the lower pricing for the 40 ft category is expected. The 40 ft category is generally for areas with single family homes and could be misleading within a pure building height comparison analysis. However, these values are still considered since this part of the zoning law limits a zone from building high rise apartment buildings and changes valuation in the long term.



Building Height (Ft)	Avg Price Per Sq Ft	Price Per Sq Ft Std Dev
40	$377.87	$207.84
50	$481.46	$181.83
65	$433.44	$229.00
90	$536.47	$168.20
110	$558.23	$72.59

Figure 17. Housing Price Constraints from Height.

Figure 18 shows the price distribution for each height category according to current zoning laws. The range of the data is very large for each of the significant categories. It is interesting to find that the average price for the 65 ft category is lower than the price for 50 and 90 ft categories, which is outside any suspected trends. This may suggest that high rise condo communities over 9 stories and low rise condo communities less than 5 stories are more desirable than medium rise condo communities with 6 to 9 stories. The variation at 65 ft and the height category's pricing distribution indicates that the height of the entire building has little influence on the pricing of a single unit; however, there may be influence from height and FAR restrictions together.

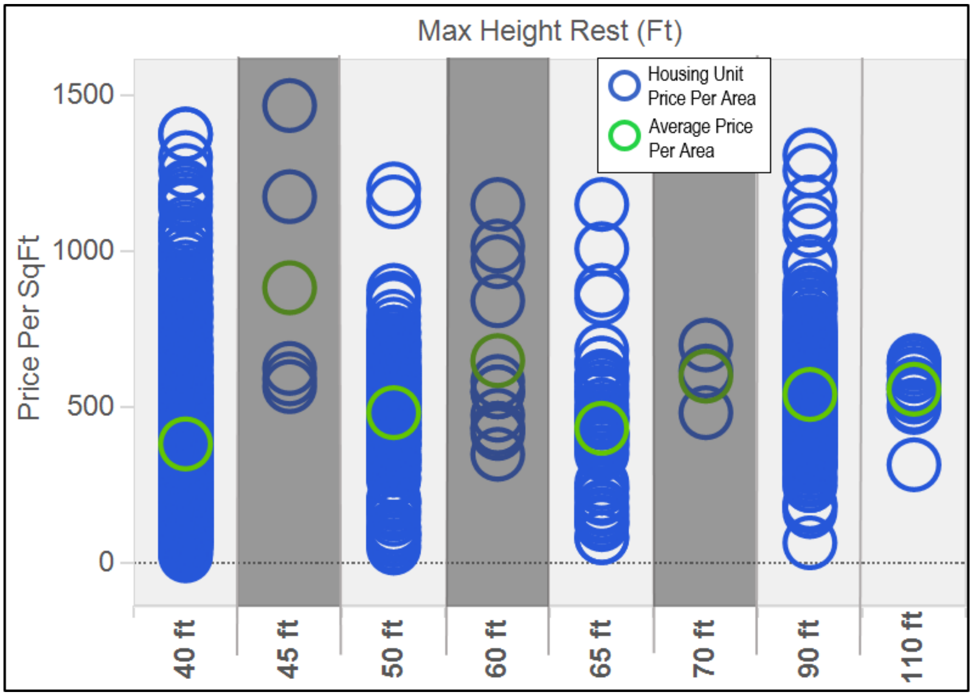


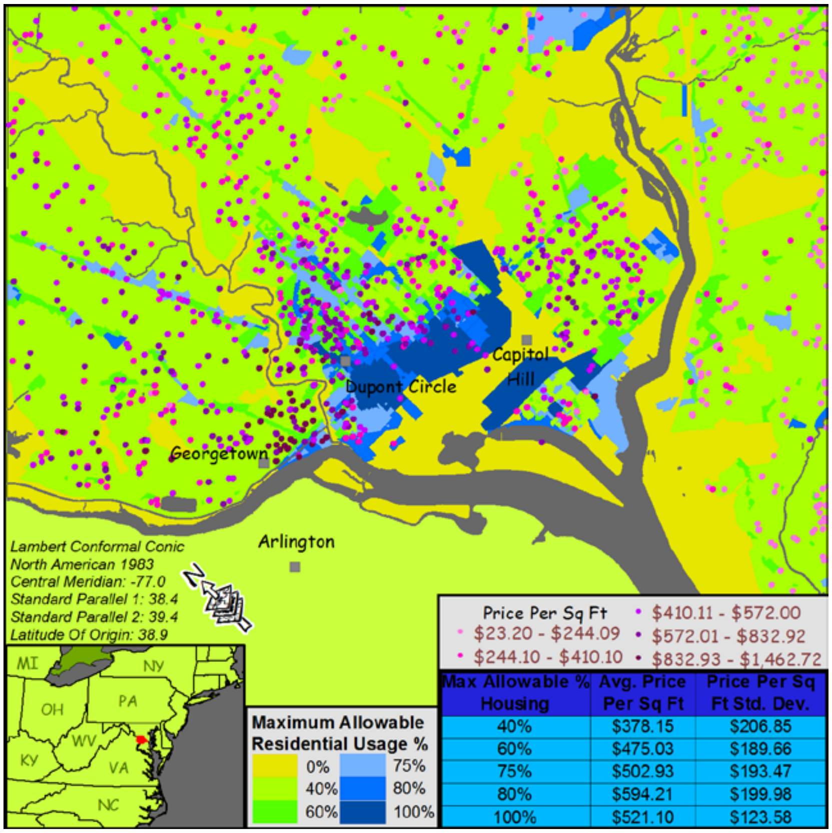
Figure 18. Pricing Distribution from Height.

This dimension appears to have some correlation with price for the 65, 90 and 110 ft categories. Even though there is a possibility that there may be some influence from this dimension and from FAR, it appears that these dimensions only affect the overall floor space within a building rather that the price per area within a single real estate unit within the building.

5.3. Maximum Residential Usage Allowed Categorization and Initial Statistics

The next zoning dimension for consideration is maximum allowable residential usage, which indicates the amount of allowed usage for residences versus all other allowable usages (like commercial or industrial) within a regulation zone. The maximum allowable residential usage values are the most directly related to property price over the other two categories.

As the map shows in Figure 19, residential usage increases towards the center of Washington, DC. Pricing variations, thus, would theoretically increase in accordance with the bid-rent hypothesis. The data distribution in Figure 20 may speak to this, however, spatially one can notice a preponderance higher valued properties in range of the greater residential usage values. This may be caused by the zoning of properties within the 80%-100% category in DuPont Circle and the area around Capitol Hill.



Max Allowable % Housing	Avg. Price Per Sq Ft	Price Per Sq Ft Std. Dev.
40%	$378.15	$206.85
60%	$475.03	$189.66
75%	$502.93	$193.47
80%	$594.21	$199.98
100%	$521.10	$123.58

Figure 19. Housing Price Constraints from Residential Usage.

Figure 20 shows the distribution of housing values within each residential usage category. The data are clustered in each category and prove to be a better measure of the effect of zoning laws upon property price than FAR and height restriction. Only with the 100% allowable usage category does the price fall slightly. This may be due to allowed mixed usages within the 100% residential zones within current zoning laws within Washington, DC. This dimension is analyzed further in the next few Chapters.

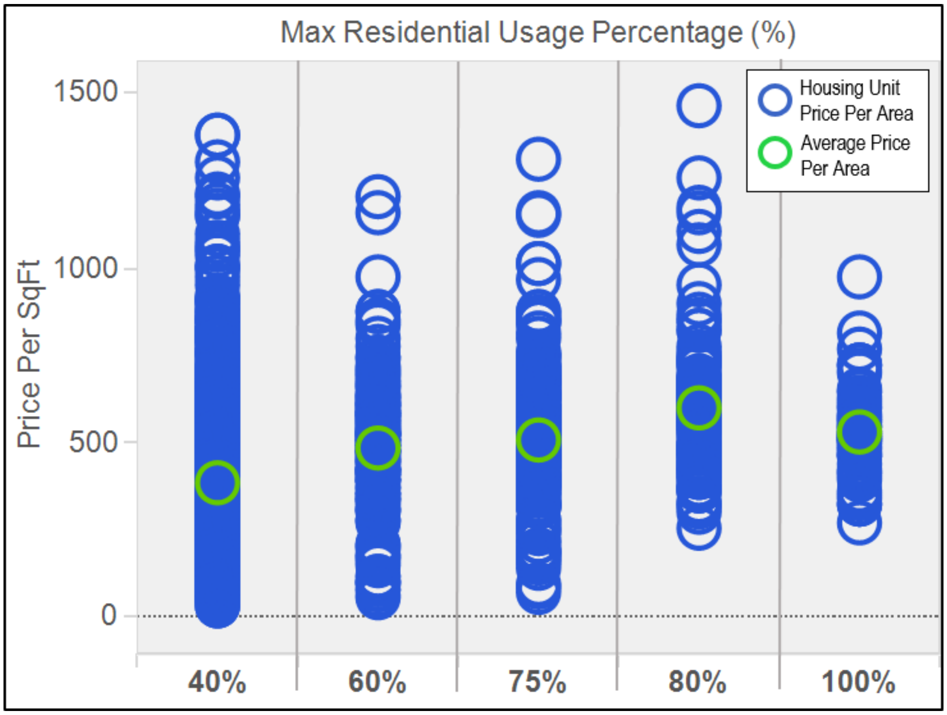


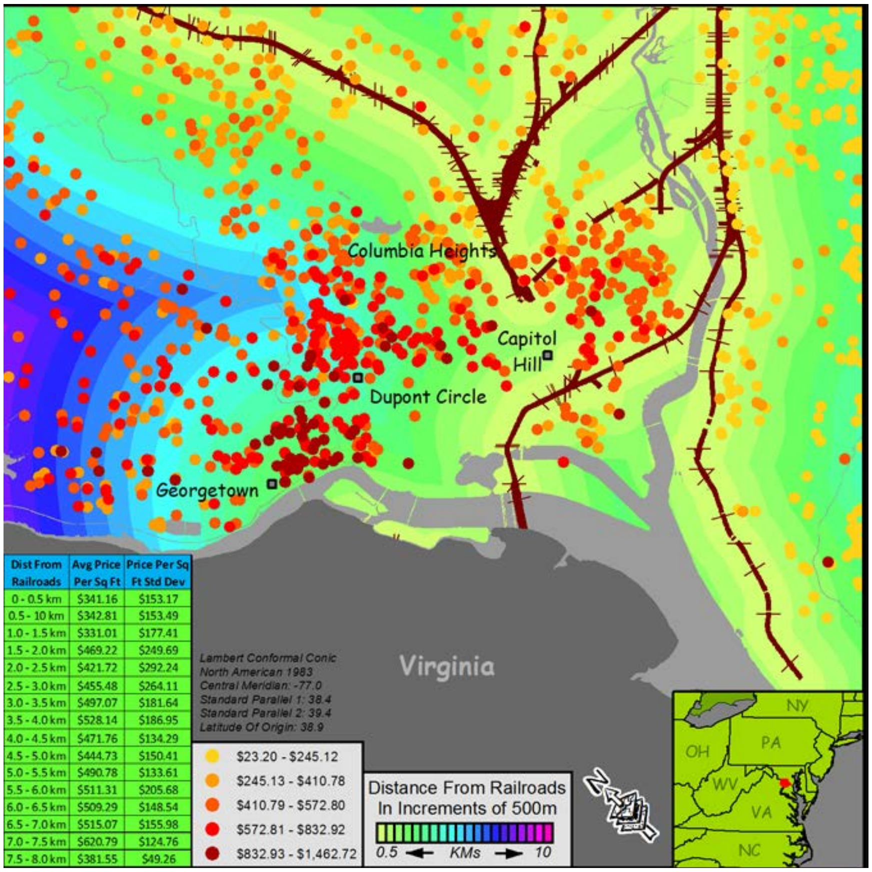
Figure 20. Pricing Distribution from Residential Usage.

An initial look at the three zoning dimensions in terms of categorization shows that allowable residential usage is the most related zoning dimension to housing price. It was also shown to have significance relation to housing price between 40% and 80%. The results of this initial level of analysis will be further evaluated within the next two layers of analysis.

5.4. Rail Influence Categorization and Initial Statistics

A number of studies discussed in section 2.2 state that the negative noise effects from railways outweigh the positive transportation effects upon property price. Ironically, the reverse effect occurs within the analysis by Tsutsumi et al. (2011)—though this may be due to the fact that railways allow for a greater transportation convenience in Japan. Within this study the noise effects of commercial railroads are analyzed separately from the transportation benefit Washington, DC Metrorail.

As seen within Figure 21, the railroads are closer to most of the properties within the lower price categories. This conclusion appears significant at lower distance levels; however, above the 2 km distance level the data do not appear directly related. It should be noted that some of the properties east of Capitol Hill are in an area surrounded by railroads on 2 sides and may influence the data slightly.



Dist From Railroads	Avg Price Per Sq Ft	Price Per Sq Ft Std Dev
0 - 0.5 km	$341.16	$153.17
0.5 - 10 km	$342.81	$153.49
1.0 - 1.5 km	$331.01	$177.41
1.5 - 2.0 km	$469.22	$249.69
2.0 - 2.5 km	$421.72	$292.24
2.5 - 3.0 km	$455.48	$264.11
3.0 - 3.5 km	$497.07	$181.64
3.5 - 4.0 km	$528.14	$186.95
4.0 - 4.5 km	$471.76	$134.29
4.5 - 5.0 km	$444.73	$150.41
5.0 - 5.5 km	$490.78	$133.61
5.5 - 6.0 km	$511.31	$205.68
6.0 - 6.5 km	$509.29	$148.54
6.5 - 7.0 km	$515.07	$155.98
7.0 - 7.5 km	$620.79	$124.76
7.5 - 8.0 km	$381.55	$49.26

Figure 21. Housing Price in Proximity to Railroads.

Within Figure 22, there is a natural bifurcation within the data between the distance categories of 1-1.5 km and 1.5-2 km. Generally, there should be a price increase trend from 0-1.5 km; however, there is a slight decrease within the 1-1.5 km category. This decrease is likely due to almost no upper range outliers as is evident with the 0-0.5 km and 0.5-1 km ranges. Thus, this thesis assumes that distance from railroads is only significant at the 0-1.5km range, since price fluctuates after 1.5 km and appears to have no relationship.

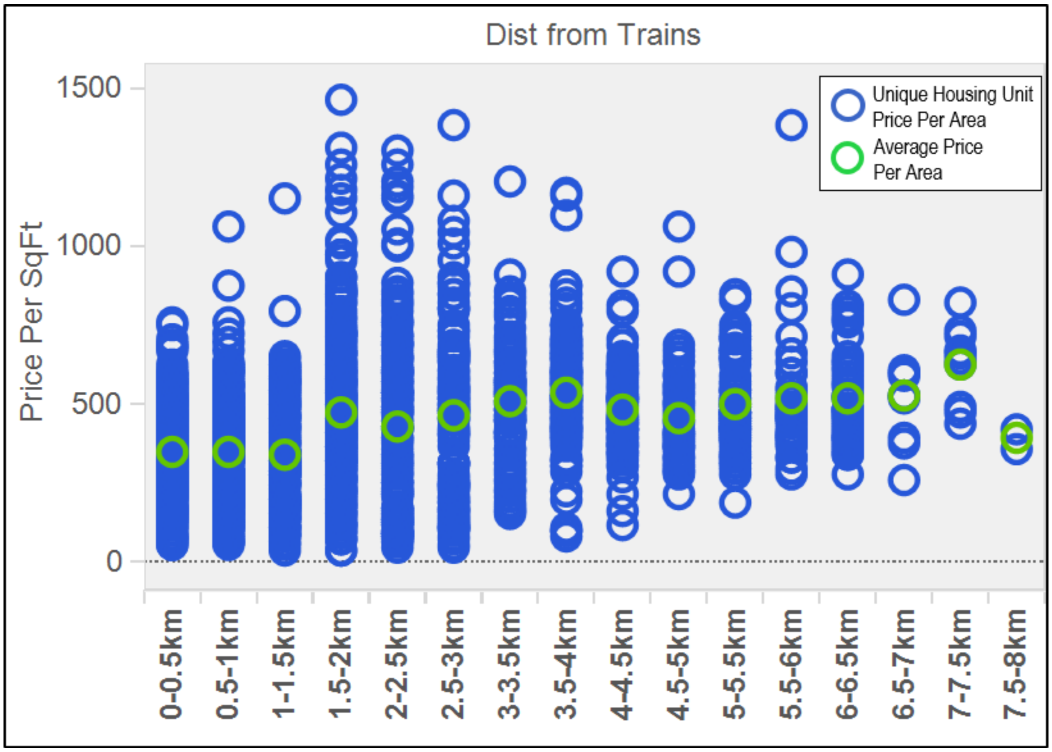


Figure 22. Railway Influence on Pricing Distribution.

The variations seen in the data may be due to the geographical layout of the railroads in relation to the real estate properties. The railroads are interwoven over the eastern portion of the city and may explain why there is a slight decrease in average price within the 1-1.5 km category. The full effectiveness of this dimension will be further assessed in the next few Chapters.

In this case, this dimension is directly influential within a range of 1.5 km of a railroad in a negative manner. This result confirms the study by Ibeas et al. (2012) in Spain that noted almost a 5% decrease in price.

5.5. Pollution Influence Categorization and Initial Statistics

Pollution obviously affects a home buyer's decision making. As Mennis (2005) showed within New Jersey, there are direct influences on property price at the local level due to pollution. This thesis compares distance from hazardous facilities and real estate pricing to confirm these results in Washington, DC. This initial level of analysis found a significant relation that is further analyzed within correlation and classification portion of this thesis.

The proximity to hazardous facilities study shown in Figure 23 had interesting results; for the most part, there was a noticeable difference in property price between lower and middle distance ranges. However, distance from a facility is likely not a factor with the higher distance ranges, since pollution is probably not a consideration to a future home owner at long distance ranges. It is possible that there are some indirect influences from zoning restrictions, since the hazardous facilities within this study mostly appear in less populous areas.

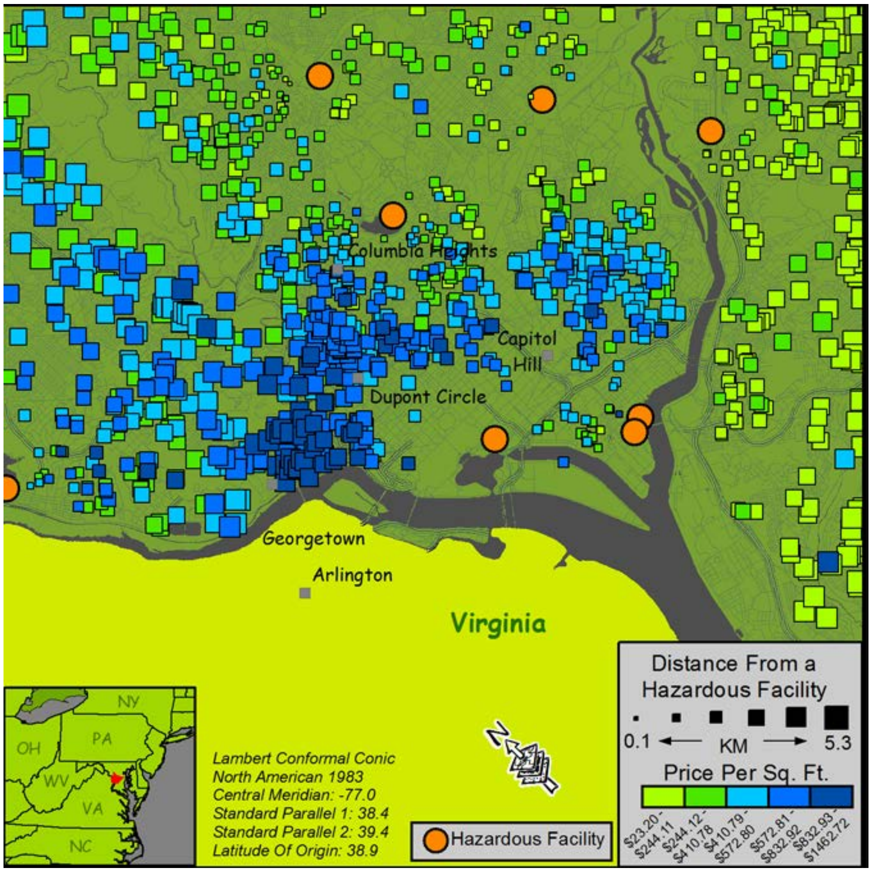


Figure 23. Housing Price in Proximity to Hazardous Facilities.

Regression analysis in Figure 24 reveals a consistent trend within the data between 0-3 km distances. This confirms the local nature of this variable. Data are clustered around a second order polynomial representation within the 0-3 km range, whereas the data are more distributed after 3 km. This is almost twice as far than railroad’s negative influence.

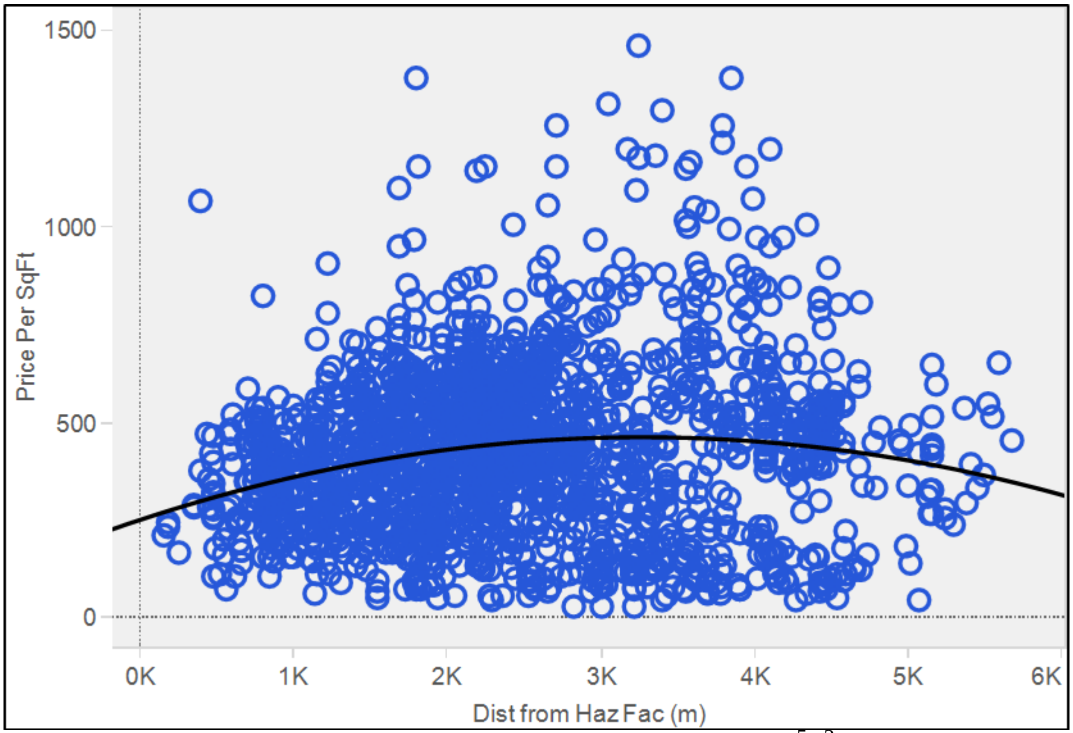


Figure 24. Distance from Hazardous Facilities ($y = -2.0*10^{-5}x^{2} + 0.13x + 250$).

These initial results confirm Mennis's (2005) suspicions about hazardous facilities' negative effect on price. Unlike, FAR or height restrictions, this dimension has a clear negative effect on pricing. The magnitude of this relation is further analyzed within the correlation portion of this thesis.

5.6. Metrorail Accessibility Categorization and Initial Statistics

Transportation convenience allows for business infrastructure improvement and influences property price in a positive manner. As shown by Tsutsumi et al. (2011) residential units near public transportation gives greater access to a city, which drives property value up. Within this thesis, the influence of the distance to Metrorail access tunnels is statistically analyzed.

Convenience to the Metrorail is another thing that a person looking for a new home may consider, since it would provide commuting convenience as well as social convenience. Figure 25 shows a significant number of high priced properties near a Metrorail station. There are also a few high priced properties away from Metrorail stations in Georgetown. This dimension will probably increase price for lower distances; however, the Georgetown properties may affect the data distribution.

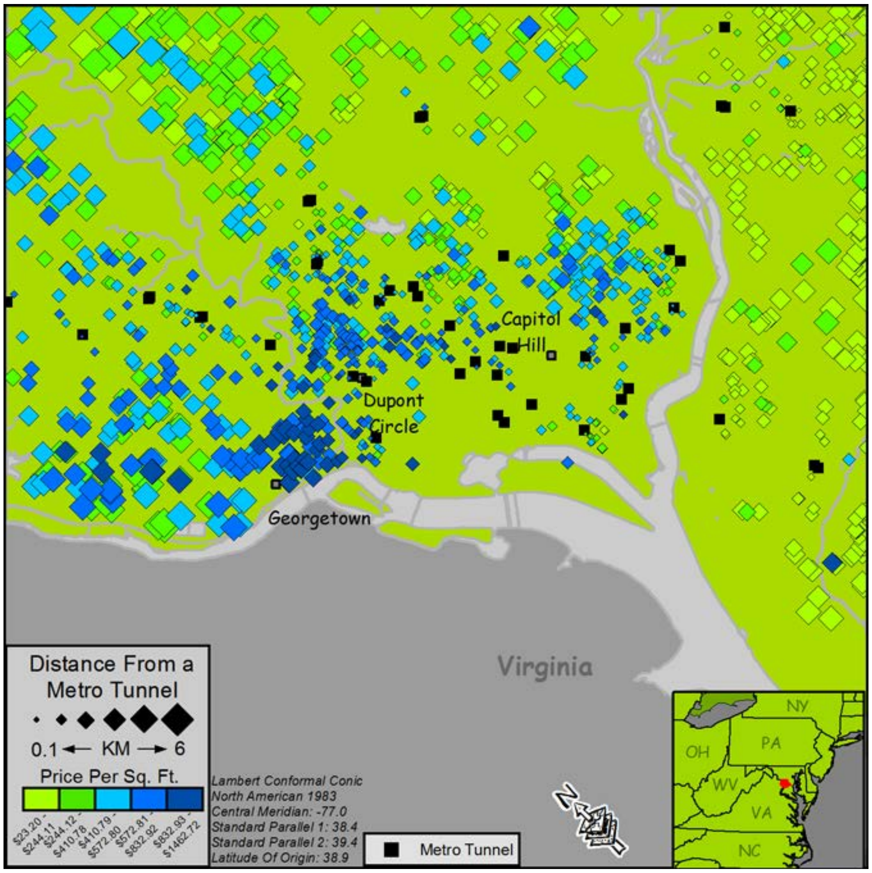


Figure 25. Housing Price in Proximity to Metrorail Tunnels.

Figure 26 shows a price decrease between 0-3.5 km. The data are clustered around the second order polynomial within this range, after which the data becomes more sporadic. These significance ranges will be further considered within the correlation portion of this thesis; however, it is likely that this dimension is influenced by its proximity to other dimensions analyzed within this study.

There are a significant number of upper bound outliers within 0-3.5 km. These properties are likely in the area northwest of Georgetown where there are a significant number of properties within the upper pricing categories within Figure 26. Outliers like these show that Metrorail accessibility is only significant at closer ranges. Generally speaking, people would choose to use alternative means of transportation at a certain critical distance.

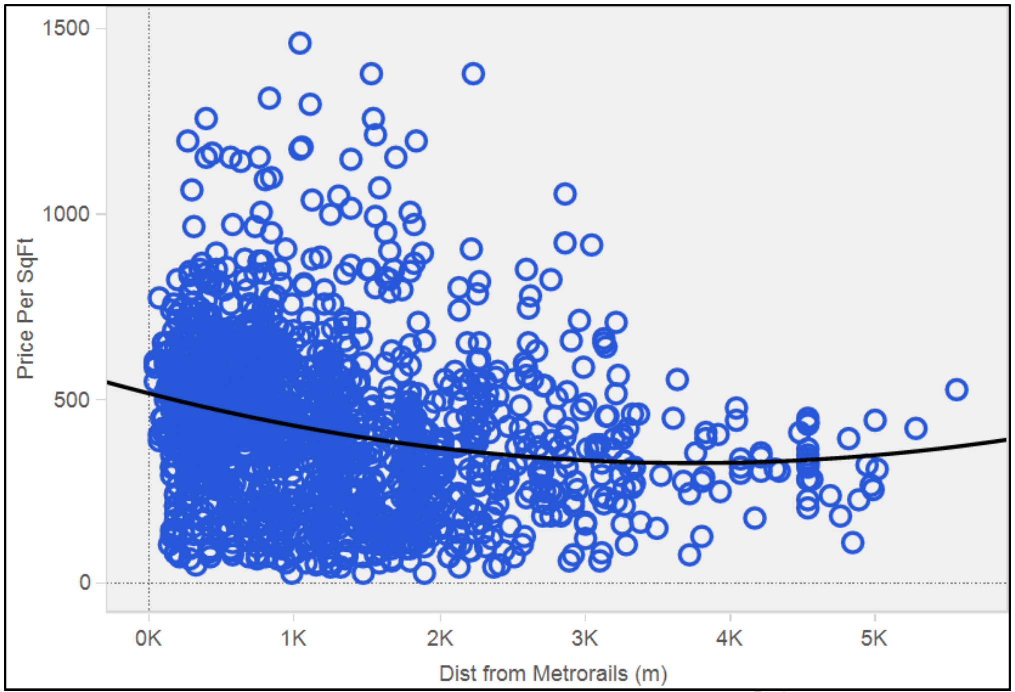


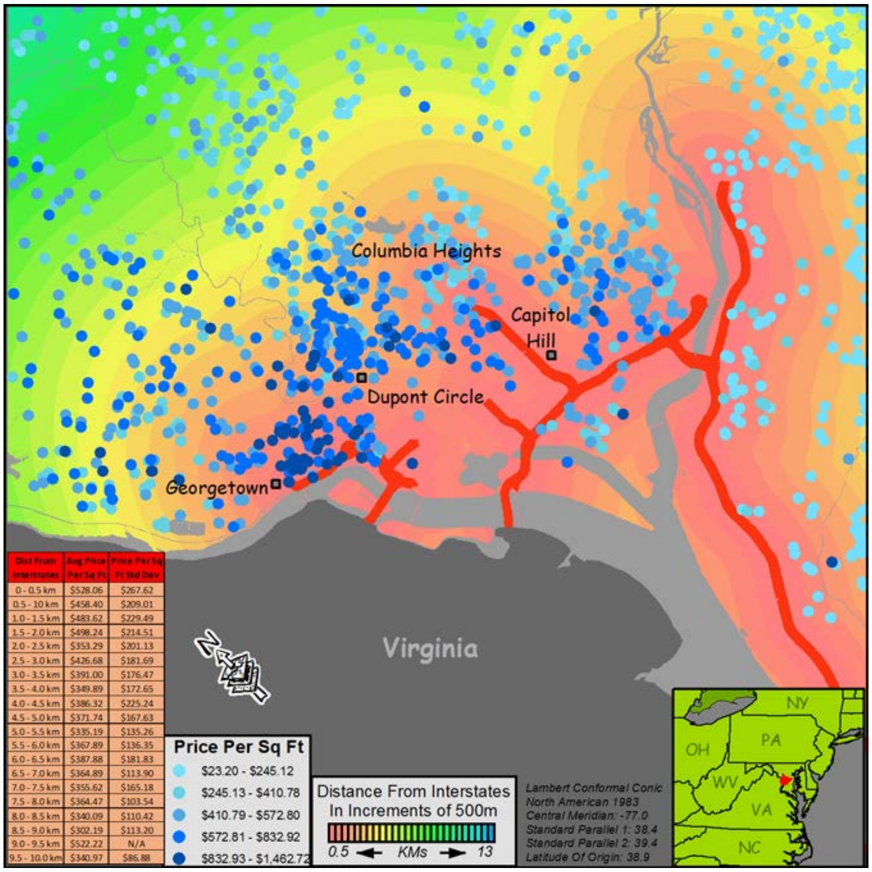
Figure 26. Distance from Metrorail Stations ($y = 1.3*10^{-5}x^{2} – 0.10x + 514$).

Since Metrorail only operates in certain areas of Washington, DC, its influential effectiveness is limited. In some cases from a subjective standpoint, Metrorail access would not be a consideration of a home buyer. The effective range of this dimension will be assessed at 0-2 km—though it is apparently it has minor influences at greater ranges. The data, however, in Figure 26 shows more outliers after 2 km.

5.7. Interstate Accessibility Categorization and Initial Statistics

Interstate accessibility allows for better transportation convenience and increases the value of a residential unit. O’Kelly and Niedzielski (2012) recently confirmed bid rent theory postulates based on commuter travel data. This study will look at access to Interstates to determine the impact on the Real-Estate market and determine the general influence range of Interstates.

Figure 27 shows that proximity to Interstates has an increasing effect upon price likely due to driving convenience. There is general decrease in price per area within the results; however, there are some variations within the data between progressive categories. This dimension likely only matters to a home owner within a few kilometers from an Interstate. It will be interesting to see how this dimension relates to other dimensions used at the end of this thesis.



Dist From Interstates	Avg Price Per Sq Ft	Price Per Sq Ft Std Dev
0 - 0.5 km	$528.06	$267.62
0.5 - 10 km	$458.40	$209.01
1.0 - 1.5 km	$483.62	$229.49
1.5 - 2.0 km	$498.24	$214.51
2.0 - 2.5 km	$353.29	$201.13
2.5 - 3.0 km	$426.68	$181.69
3.0 - 3.5 km	$391.00	$176.47
3.5 - 4.0 km	$349.89	$172.65
4.0 - 4.5 km	$386.32	$225.24
4.5 - 5.0 km	$371.74	$167.63
5.0 - 5.5 km	$335.19	$135.26
5.5 - 6.0 km	$367.89	$136.35
6.0 - 6.5 km	$387.88	$181.83
6.5 - 7.0 km	$364.89	$113.90
7.0 - 7.5 km	$355.62	$165.18
7.5 - 8.0 km	$364.47	$103.54
8.0 - 8.5 km	$340.09	$110.42
8.5 - 9.0 km	$302.19	$113.20
9.0 - 9.5 km	$522.22	N/A
9.5 - 10.0 km	$340.97	$86.88

Figure 27. Housing Price in Proximity to Interstates.

Figure 28 shows a general decrease in average price value from Interstates. The oscillation in average values after 1 km indicates that this dimension is only significant between 0-1 km. However, the average home values are significantly lower after the 2 km break. This may indicate other factors that may be influencing the data within this case; however, the 0-0.5 km and 0.5-1 km category definitely appear directly correlated with price.

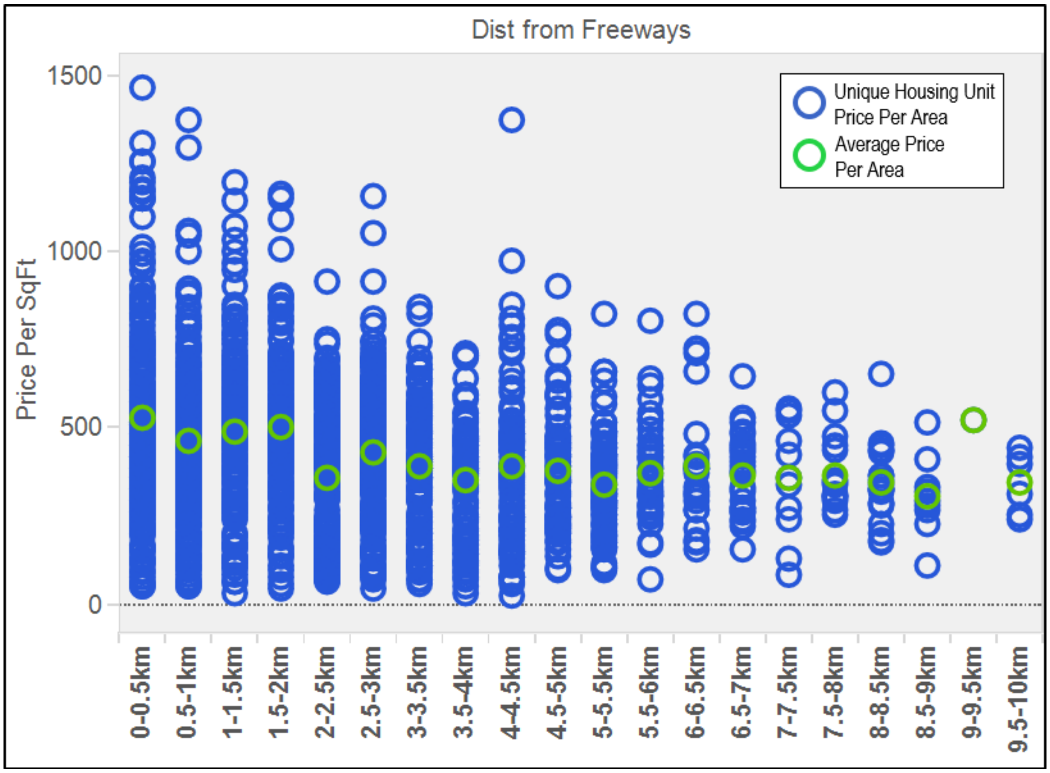


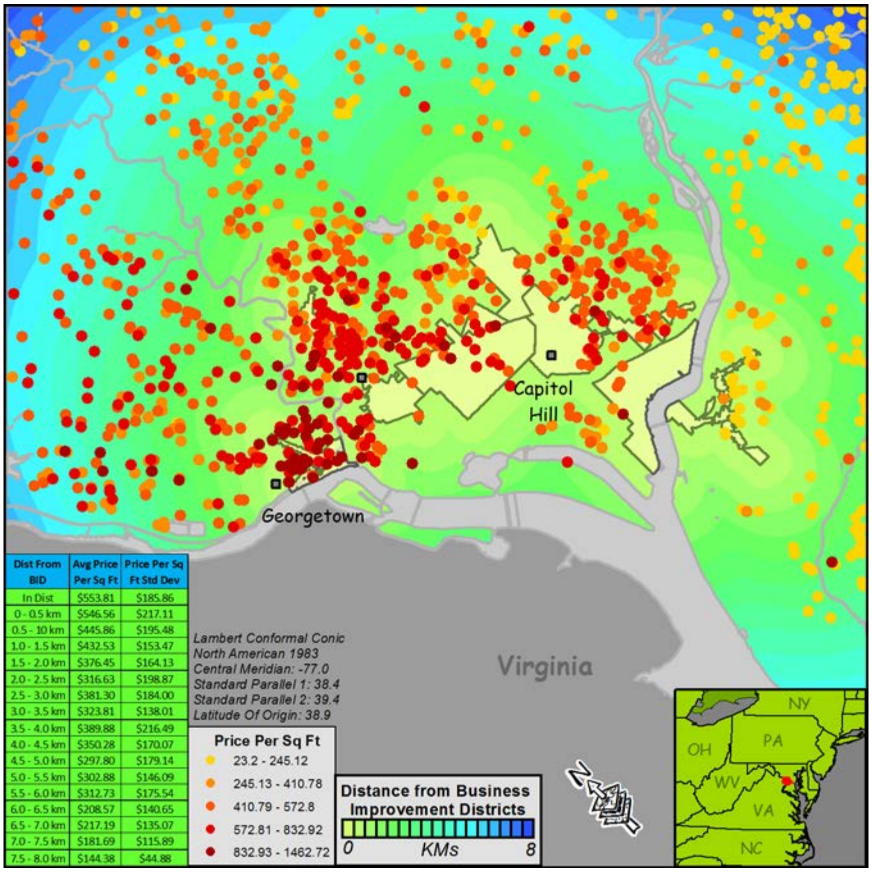
Figure 28. Interstate Influence on Pricing Distribution.

There is some information within the data that indicates a direct relationship between distance from Interstates and property value. A new home buyer may take into consideration accessibility to interstates; thus, a significance level for properties within 0-1 km corroborates that idea. Also like Metrorail access, the positive effects from transportation convenience appear to outweigh the negative effects from Interstate noise in the District of Columbia.

5.8. Business District Proximity Categorization and Initial Statistics

Peer review literature shows that social and business infrastructure within society influences property price distribution. Several studies upon business expenditures within a region have shown considerable influence on neighboring real estate prices. This thesis finds significant relation between distance from business districts and property price.

Figure 29 shows that price dramatically decreases as distance from a Business Improvement District (BID) increases within the first few kilometers. Outside of that range the value per area begins to fluctuate. The influence of business center proximity may be similar to other strictly local influences determined within this study. As shown in Figure 29, there are many high price properties near the BID in Georgetown and near DuPont Circle. These data are analyzed more thoroughly later in this thesis.



Dist From BID	Avg Price Per Sq Ft	Price Per Sq Ft Std Dev
In Dist	$553.81	$185.86
0 - 0.5 km	$546.56	$217.11
0.5 - 10 km	$445.86	$195.48
1.0 - 1.5 km	$432.53	$153.47
1.5 - 2.0 km	$376.45	$164.13
2.0 - 2.5 km	$316.63	$198.87
2.5 - 3.0 km	$381.30	$184.00
3.0 - 3.5 km	$323.81	$138.01
3.5 - 4.0 km	$389.88	$216.49
4.0 - 4.5 km	$350.28	$170.07
4.5 - 5.0 km	$297.80	$179.14
5.0 - 5.5 km	$302.88	$146.09
5.5 - 6.0 km	$312.73	$175.54
6.0 - 6.5 km	$208.57	$140.65
6.5 - 7.0 km	$217.19	$135.07
7.0 - 7.5 km	$181.69	$115.89
7.5 - 8.0 km	$144.38	$44.88

Figure 29. Housing Price in Proximity to BID.

Figure 30 shows the variation in price per distance category from BID. The average price decreases consistently until the data changes from the 2-2.5 km category to the 2.5 km-3 km category. After the 2.5 km break, the average price fluctuates showing that distance from a BID within this study is only significant within 0-2.5 km.

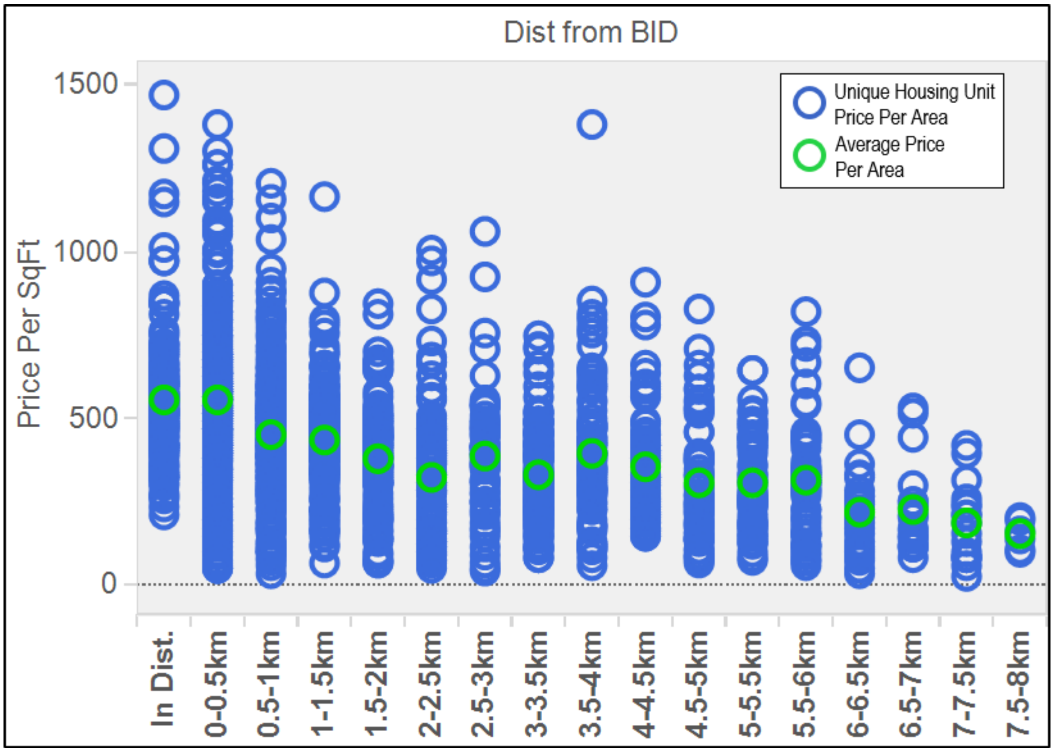


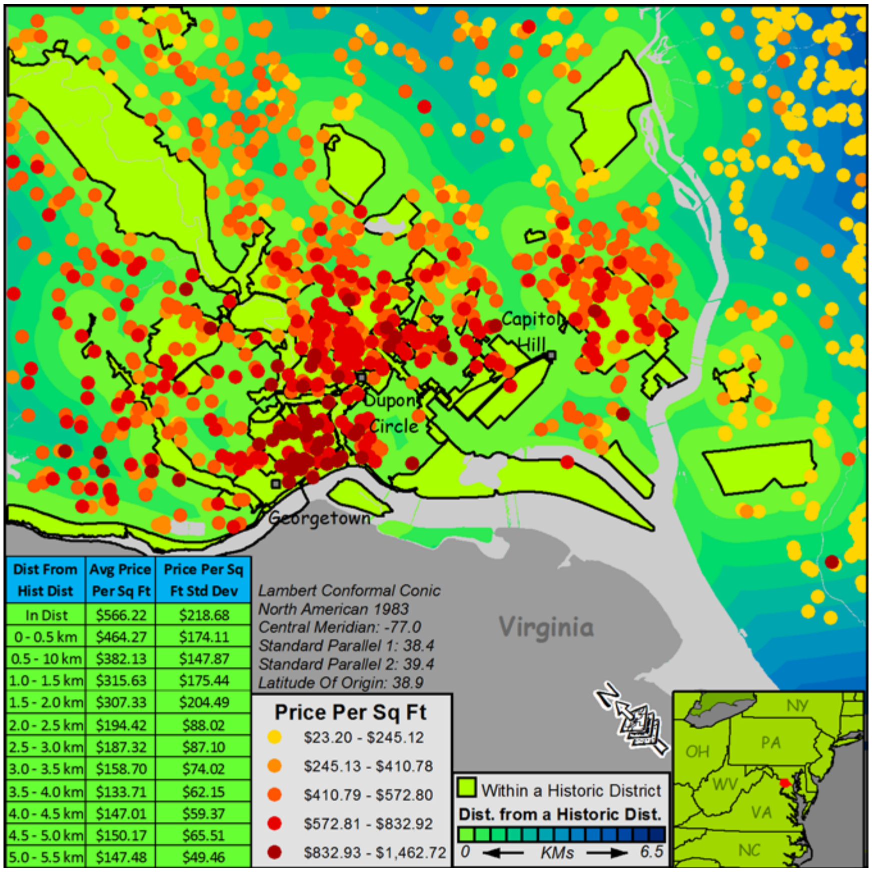
Figure 30. BID Influence on Pricing Distribution.

This dimension appears to have greater influence than the other attributes found within this study on real estate price. However, there were some variations after 2.5 km. The correlation analysis portion will give further insight into the relation of these influences.

5.9. Historical Influence Categorization and Initial Statistics

Washington, DC is rich with historical districts. This study encountered 57 historic zones. As Han (2004) noted within Beijing and Jakarta, historical districts within an area has considerable influence on pricing when looking at cross-sections of a city. This study will attempt to determine if these historical influences have more influence than the other factors discussed within this thesis.

As expected, Figure 31 shows there is a direct decrease in home value as distance from a historical district increases. Historic districts in this study incorporate high value properties in Georgetown, DuPont Circle and in the area north of Capitol Hill. Also, throughout the spatial distribution, most of the second and third level pricing category properties appear within 1.0 km of a historic district. This is one of the few dimensions encountered so far that appear to correspond directly to unit floor area price.



Dist From Hist Dist	Avg Price Per Sq Ft	Price Per Sq Ft Std Dev
In Dist	$566.22	$218.68
0 - 0.5 km	$464.27	$174.11
0.5 - 10 km	$382.13	$147.87
1.0 - 1.5 km	$315.63	$175.44
1.5 - 2.0 km	$307.33	$204.49
2.0 - 2.5 km	$194.42	$88.02
2.5 - 3.0 km	$187.32	$87.10
3.0 - 3.5 km	$158.70	$74.02
3.5 - 4.0 km	$133.71	$62.15
4.0 - 4.5 km	$147.01	$59.37
4.5 - 5.0 km	$150.17	$65.51
5.0 - 5.5 km	$147.48	$49.46

Figure 31. Housing Price in Proximity to Historical Districts.

Pricing data in Figure 32 decreases for each progressive pricing category between 0-4 km. Also of significance, there is almost a $100 per square foot variation between properties within a historic district and properties within 500m of a historic district. These two pricing attributes shows that distance from historical districts is a highly influential factor significant between 0-4 km.

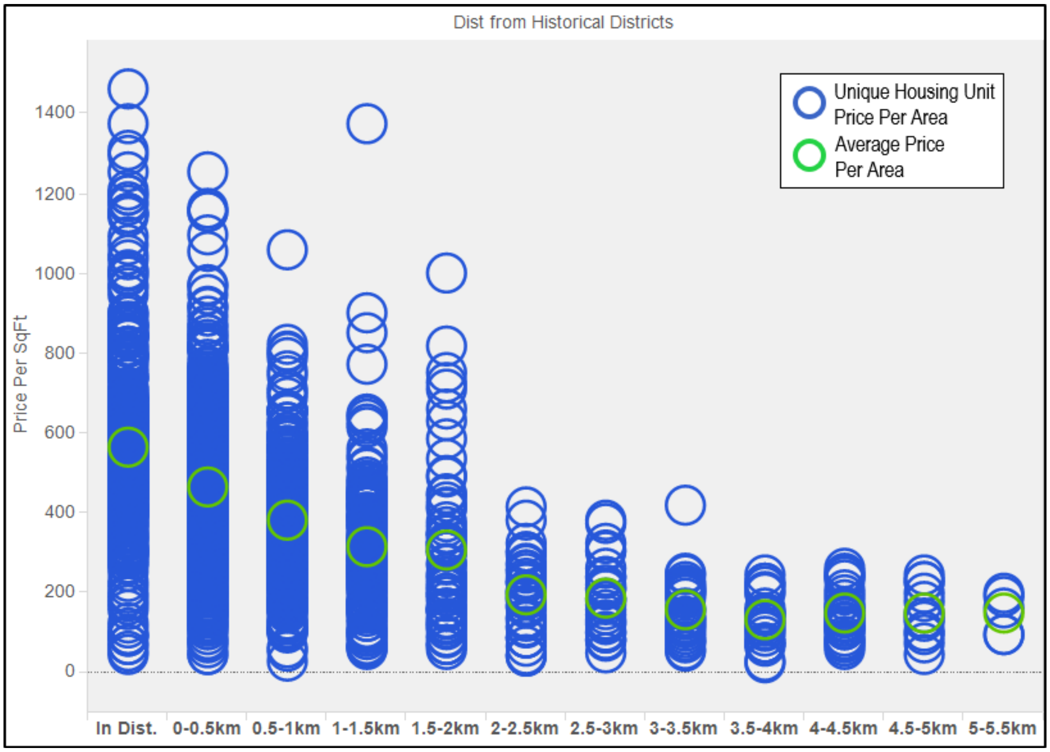


Figure 32. Historic Districts Influence on Pricing Distribution.

This study will carefully consider the correlation of this layer with the zoning layers. This dimension is derived based on government designated historic districts and it may correspond a little to the other zoning factors covered within this thesis. If one carefully chooses the dimensions for the classification model with regards to intermediate correlation, the model will yield more accurate rules.

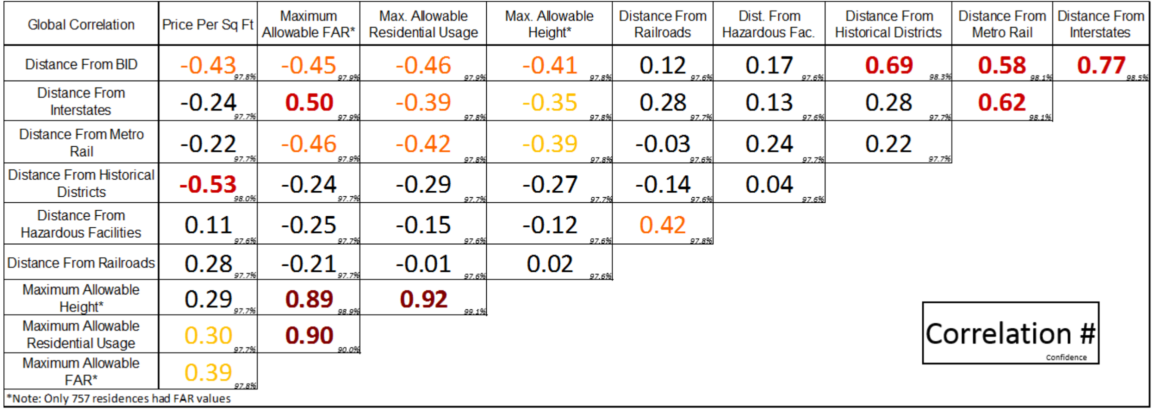
6. CORRELATION OF THE EXTERNAL INFLUENCE DIMENSIONS

Within this Chapter, Pearson correlation calculations are generated to compare the nine dimensions studied. This is done for two purposes: to remove redundant information from the data mining models and to evaluate statistical dependencies within the data. Two versions of the model are created at the global level and local level. The global level includes all ranges of the data analyzed. The local level includes only the significant ranges identified in Chapter 5.

Global correlation comparison in Table 4 shows that the three zoning law dimensions are correlated with at least a 0.89 magnitude with more than 95% confidence. This makes senses since this data comes from the same zoning laws. Consideration will be used when handling these dimensions within the data mining sections, the most applicable of these three dimensions will trump the other two dimensions within that stage of analysis.

There was also high magnitude between interstates and BID with a 98.5% confidence in Table 4. Business districts would likely need the most improved roads possible, so this high magnitude correlation makes sense. There are no other correlation values that are greater than a 0.75 magnitude between evaluation dimensions at the global scale. These relations are further evaluated with local correlation calculations.

Table 4. Global Correlation Comparison between each Dimension.



Global Correlation	Price Per Sq Ft	Maximum Allowable FAR*	Max. Allowable Residential Usage	Max. Allowable Height*	Distance From Railroads	Dist. From Hazardous Fac.	Distance From Historical Districts	Distance From Metro Rail	Distance From Interstates
Distance From BID	-0.43 97.8%	-0.45 97.9%	-0.46 97.9%	-0.41 97.8%	0.12 97.6%	0.17 97.6%	**0.69** 98.3%	**0.58** 98.1%	**0.77** 98.5%
Distance From Interstates	-0.24 97.7%	**0.50** 97.9%	-0.39 97.8%	-0.35 97.8%	0.28 97.7%	0.13 97.6%	0.28 97.7%	**0.62** 98.1%	
Distance From Metro Rail	-0.22 97.7%	-0.46 97.9%	-0.42 97.8%	-0.39 97.8%	-0.03 97.6%	0.24 97.7%	0.22 97.7%		
Distance From Historical Districts	**-0.53** 98.0%	-0.24 97.7%	-0.29 97.7%	-0.27 97.7%	-0.14 97.6%	0.04 97.6%			
Distance From Hazardous Facilities	0.11 97.6%	-0.25 97.7%	-0.15 97.6%	-0.12 97.6%	0.42 97.8%				
Distance From Railroads	0.28 97.7%	-0.21 97.7%	-0.01 97.6%	0.02 97.6%					
Maximum Allowable Height*	0.29 97.7%	**0.89** 98.9%	**0.92** 99.1%						
Maximum Allowable Residential Usage	0.30 97.7%	**0.90** 90.0%							
Maximum Allowable FAR*	0.39 97.8%								

Correlation #
Confidence

*Note: Only 757 residences had FAR values

It should be noted though that there are a few more high magnitude relationships within the data at the global scale in Table 4. There is a group of 0.5-0.8 magnitude correlation calculations with a 95% confidence between BID, Interstates, Historical Districts and Metrorail. There is also a group of 0.3-0.5 magnitude correlation measures with a 95% confidence between the zoning dimensions and the distances from BID, Interstates and Historical Districts. The models in Chapter 7 and 8 take into account the cross correlation influences found in this portion of the analysis.

As discussed in Chapter 5, seven of the nine dimensions had a significant relation to average unit price. Most of the dimensions had a significant distance or significant value from category arithmetic mean trends or from regression analysis. Table 5 shows these values, which will be used within Table 6 that shows local correlation comparison.

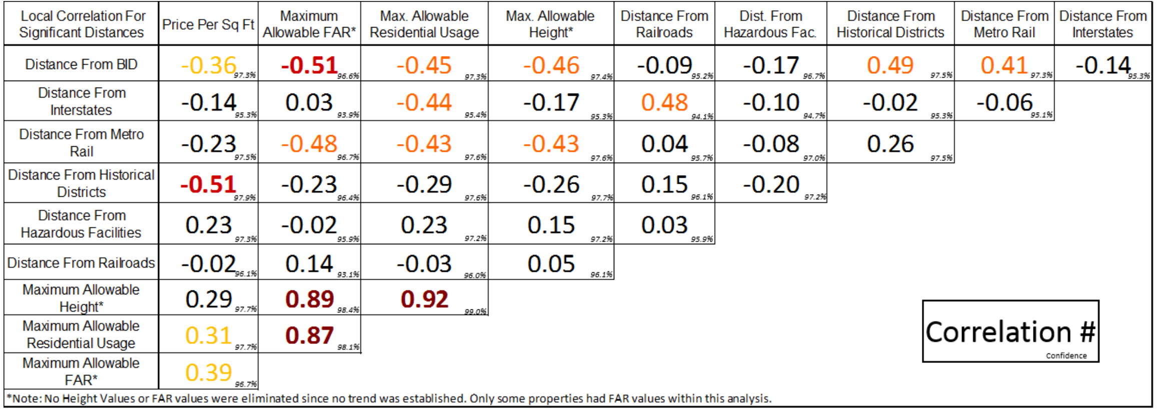
Table 5. Comparison of Significant Distance Ranges within each Dimension.

Dimension With Partial Spatial Significance	Significant Ranges	Ranges Ruled Out
Maximum Residential Usage Allowed	40%-80%	100%
Distance from Railroads	0-1.5 km	1.5-8 km
Distance from Hazardous Facilities	0-3 km	3-6 km
Distance from Metrorail	0-2 km	2-6 km
Distance from Interstates	0-1 km	1-10 km
Distance from BID	0-2.5 km	2.5-8 km
Distance from Historical Districts	0-4 km	4-5.5 km

Local correlation negates some of the higher magnitudes found between the evaluated dimensions within the global correlation calculations. As shown by Table 6 for each correlation comparison, each pair of dimensions had at least 94% confidence. The evaluations between Railroads and Interstates and between hazardous facilities and Interstates both had confidence levels that were slightly below the statistical standard of 95%; however, this occurs only for 2 of the 45 correlation calculations within Table 6.

Local correlation analysis actually reduced railroad distance and price correlation magnitude significantly at a 96.1% confidence. The lack of local correlation shows evaluation within the local ranges for railroads is virtually insignificant with Washington, DC data. Within both correlation evaluations, distance from historical districts has the greatest magnitude.

Table 6. Local Correlation Comparison between each Dimension.



Local Correlation For Significant Distances	Price Per Sq Ft	Maximum Allowable FAR*	Max. Allowable Residential Usage	Max. Allowable Height*	Distance From Railroads	Dist. From Hazardous Fac.	Distance From Historical Districts	Distance From Metro Rail	Distance From Interstates
Distance From BID	-0.36 97.3%	**-0.51** 96.6%	-0.45 97.3%	-0.46 97.4%	-0.09 95.2%	-0.17 96.7%	0.49 97.5%	0.41 97.3%	-0.14 95.3%
Distance From Interstates	-0.14 95.3%	0.03 93.9%	-0.44 95.4%	-0.17 95.3%	0.48 94.1%	-0.10 94.7%	-0.02 95.3%	-0.06 95.1%	
Distance From Metro Rail	-0.23 97.5%	-0.48 96.7%	-0.43 97.6%	-0.43 97.6%	0.04 95.7%	-0.08 97.0%	0.26 97.5%		
Distance From Historical Districts	**-0.51** 97.9%	-0.23 96.4%	-0.29 97.6%	-0.26 97.7%	0.15 96.1%	-0.20 97.2%			
Distance From Hazardous Facilities	0.23 97.3%	-0.02 95.9%	0.23 97.2%	0.15 97.2%	0.03 95.9%				
Distance From Railroads	-0.02 96.1%	0.14 93.1%	-0.03 96.0%	0.05 96.1%					
Maximum Allowable Height*	0.29 97.7%	**0.89** 98.4%	**0.92** 99.0%						
Maximum Allowable Residential Usage	0.31 97.7%	**0.87** 98.1%							
Maximum Allowable FAR*	0.39 96.7%								

Correlation #
Confidence

*Note: No Height Values or FAR values were eliminated since no trend was established. Only some properties had FAR values within this analysis.

Within both global and local correlation evaluations, only distance from historic districts and average price maintains a significant correlation over a magnitude of 0.5—except in the case of correlation between zoning dimensions. It should be noted that there are multiple pairs of dimensions with a magnitude of 0.4 or greater in both local and global correlation comparisons.

Beside the interrelation of zoning dimensions, there is a significant relation between the zoning dimensions and BID distance, the zoning dimensions and Metrorail distance, BID distance and historic district distance, and BID distance and Metrorail distance. This BID influence on zoning laws is inherently inevitable since stricter zoning laws are generally used within business district centers and are close to a Metrorail.

As one can see from the residential unit price column, distance from historic districts has the highest magnitude, which was followed by distance from BID and zoning law influences. Further analysis within a Decision Stump Classifier Model will validate these results.

7. ASSOCIATION RULES FOR THE EXTERNAL INFLUENCE DIMENSIONS

Correlation reveals interesting relationships within the data; however, in order to determine if there is more information from a combination of dimensions, Apriori Association Rules were generated. Association rules are used within market analysis to predict the likelihood a customer may buy certain goods based on other goods he bought.

Within this thesis, the same concept is applied to ranges of the external dimension data to predict housing price with one exception: all secondary rules with lower support were still retained. This would show a combination of dimensions that predicted a certain price category that might be ruled out in some cases when there would be a single dimension rule with higher confidence.

Since the average price of a real estate unit is slightly spurious within this study, association rules were only generated for property values in $100 categories from the properties that are within the $50 - $650 per square ft range. Figure 33 shows the categories utilized within this portion of the analysis. Data generalization of the pricing category for this model allows for cross dimension analysis. The support for each itemset item is shown in Figure 33.

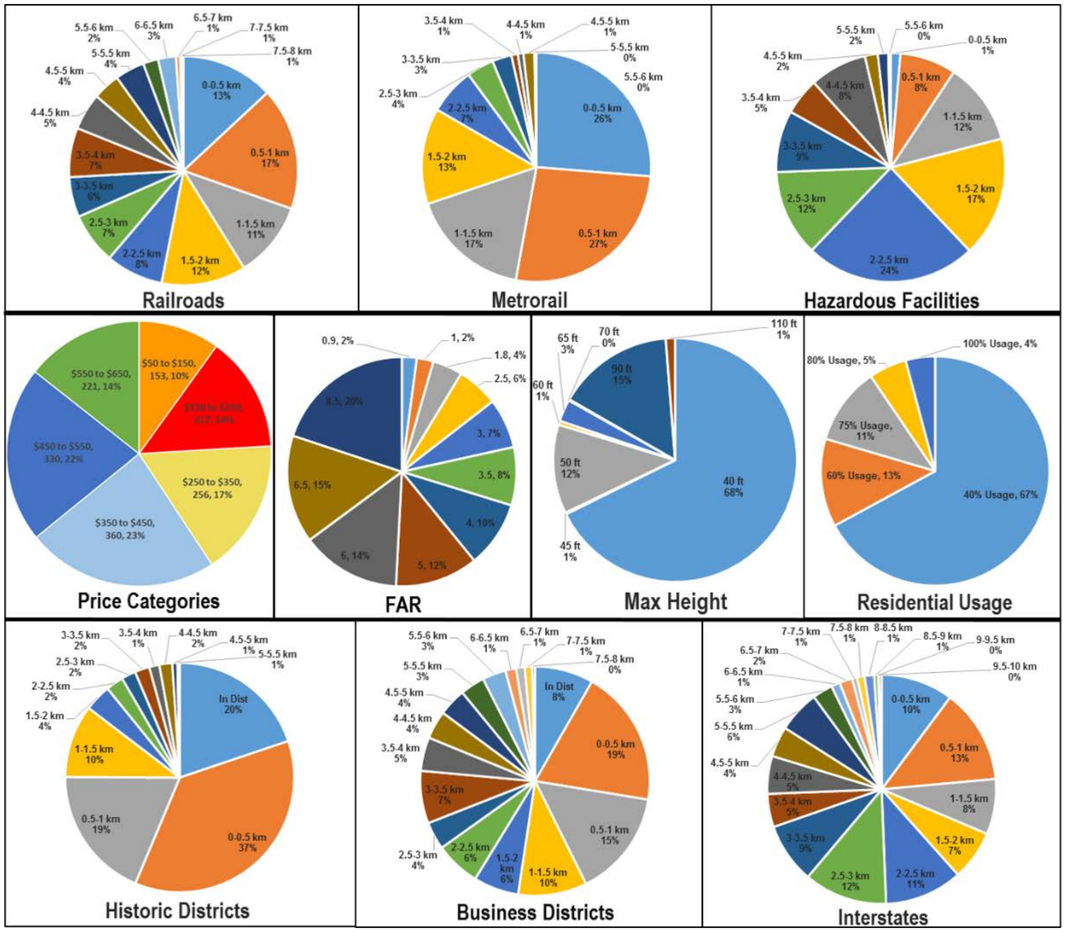


Figure 33. Categories used in Apriori Association Rule Analysis.

In order for data to be incorporated within an Apriori association rule model, it first had to be converted into nominal form with a filter function within WEKA. Before the filter was applied, the dimensions with continuous data for Metrorail and Hazardous Facilities were categorized into 500m intervals in order to be consistent with the other distance dimensions.

The Antecedent Support is the total support for the all dimensions that make up the antecedent. The antecedent is the left hand portion of the rule (LHS). The Consequent

Support within Table 7 is simply the proportion of the overall data that falls within one of the six price categories as shown in Figure 33 above. The consequent is the right hand side (RHS) of the rule.

The Rule Order Strength indicates if the rule was the one with the most confidence for that particular LHS. Theoretically, there are six possible LHS since there are only six possible outcomes for the RHS for this particular analysis. Thus, the Rule Order Strength indicates if it was the first LHS with the highest confidence generated for the model or the second LHS with the second highest confidence and so forth. Second and third order antecedents are interesting for this analysis since they can indicate a price category close to the first order antecedent.

The lift for the rule indicates the significance of the rule versus the ratio of when the rule should hypothetically occur naturally. In other words, it is the ratio of the support of both the RHS and the LHS over the product of the individual support values for RHS and LHS separately.

The model in WEKA was adjusted to its lowest allowable limit to return all rules that were greater than a 10% confidence and a 0.05 support. There were 1,693 rules generated based on different divisions within the external dimensions. 25 significant rules were found concerning property price since the model generated consequents that did not involve the pricing category.

Highly correlated zoning dimension rules were reduced to one rule within the results. Rules that predicted multiple dimensions along with price were ruled out since there was a stronger rule only for the price. Rules with no FAR were ruled out since this study does not want to look at no FAR instances within that category. Combination rules between the highly correlated zoning law dimensions were ruled out since they were overshadowed by a stronger single dimension zoning law rule. Also, rules with a lift value below one were ruled out since they are not significant.

The significant results found from Association Analysis in Table 7 shows that there was only one significant rule found between multiple dimensions that predicted price. The lack of cross dimension rules within the data emphasizes the other data mining approach that will be used in Chapter 8 since it is a technique that emphasizes single dimension influence.

Table 7. Significant Apriori Association Rules.

Antecedent	Antecedent Support	Price Consequent	Consequent Support	Rule Confidence	Rule Order Strength	Lift
Hist Dist: Within a District	20%	$450 to $550	22%	35%	First	1.62
Height: 90 ft	15%	$450 to $550	22%	34%	First	1.57
Hist Dist: 0.5-1 km	19%	$350 to $450	23%	32%	First	1.35
Business Dist: 0-0.5 km	19%	$550 to $650	14%	31%	First	2.18
Hist Dist: Within a District	20%	$550 to $650	14%	30%	Second	2.12
Metrorail: 0-0.5 km	26%	$450 to $550	22%	30%	First	1.39
Business Dist: 0-0.5 km	19%	$450 to $550	22%	29%	Second	1.33
Hist Dist: 0-0.5 km Height: 40 ft	25%	$350 to $450	23%	29%	First	1.24
Hist Dist: 0-0.5 km	37%	$450 to $550	22%	28%	First	1.30
Metrorail: 0.5-1 km	26%	$450 to $550	22%	28%	First	1.29
Hist Dist: 0-0.5 km	37%	$350 to $450	23%	28%	Second	1.20
Haz Fac: 1.5-2 km	17%	$350 to $450	23%	28%	First	1.18
Haz Fac: 1.5-2 km	17%	$450 to $550	22%	27%	Second	1.24
Metrorail: 0-0.5 km	26%	$350 to $450	23%	25%	Second	1.09
Metrorail: 0.5-1 km	26%	$350 to $450	23%	23%	Second	1.04
Resid Use: 40%	67%	$350 to $450	23%	23%	First	1.02
Height: 40 ft	68%	$350 to $450	23%	23%	First	1.01
Metrorail: 0.5-1 km	26%	$550 to $650	14%	21%	Third	1.56
Haz Fac: 1.5-2 km	17%	$550 to $650	14%	21%	Third	1.46
Metrorail: 0-0.5 km	26%	$550 to $650	14%	21%	Third	1.46
Height: 40 ft	68%	$250 to $350	17%	20%	Second	1.20
Resid Use: 40%	67%	$250 to $350	17%	20%	Second	1.19
Height: 40 ft	68%	$150 to $250	14%	19%	Third	1.34
Resid Use: 40%	67%	$150 to $250	14%	19%	Third	1.33
Hist Dist: 0-0.5 km	37%	$550 to $650	14%	18%	Third	1.23

The final results had one interesting rule for multiple dimensions. It combined 0.5-1 km from historical districts with a 40 ft height restriction and predicted a $350-$450 price per square foot with a confidence of 29% and a lift of 1.24. Considering the support for height at 40 ft is 68% and the support for 0-0.5 km of historical districts is 37%, the combined support for the rule is 25%.

This rule confidence of 29% was higher than the confidence for 0-0.5 km distance from a historical district predicting $350-$450 at 28% and the confidence for height at 40 ft predicting $350-$450 at 23%. The 1% difference in confidence between the multiple dimension rule and 0-0.5 km from a historical district rule shows that the multiple dimension rule is almost marginalized by a single dimension; however, the multiple dimension rule is still slightly more significant. Figure 34 shows the geographic distribution of properties with a 40 ft height limitation, a 0-0.5 km range from a historic district and a pricing category of $350-$450. These homes are distributed throughout the map, so this rule is not influenced by a group of homes in one area.

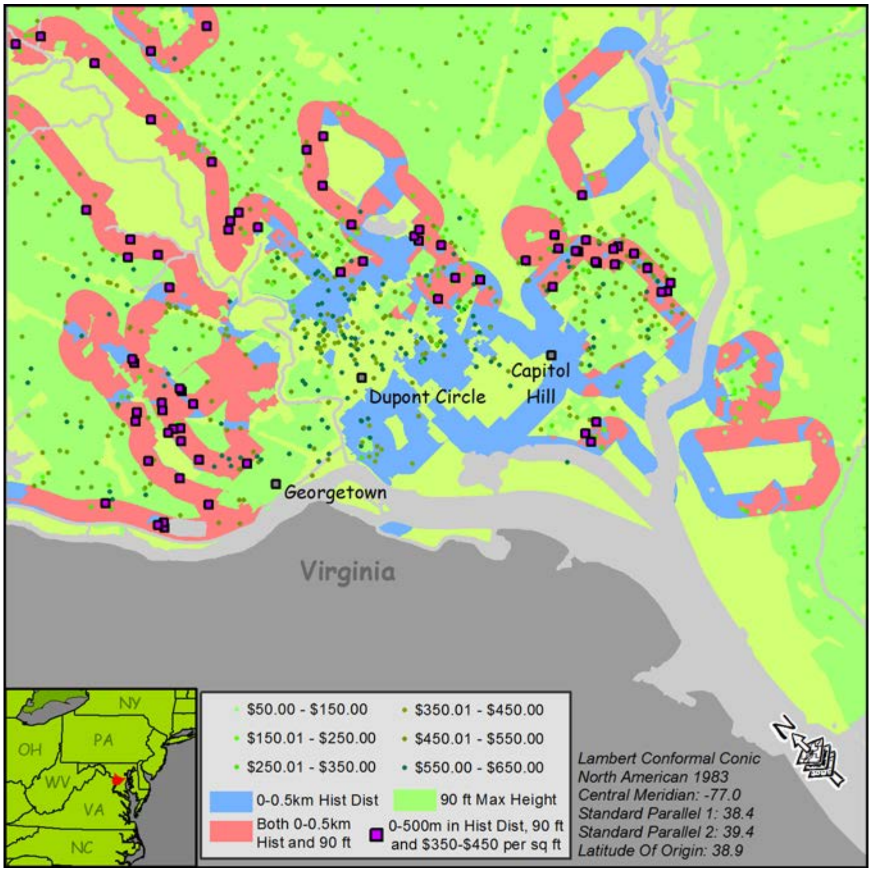


Figure 34. Historical Districts (0-0.5 km) and Height (90 ft) Association Rule.

When a home was within historical districts, 35% of the time it was within the $450-$550 pricing category—this was the rule with the highest confidence. This was followed by height at 90 ft, which was within $450-$550 pricing category at a 34% confidence rate. These rules were followed by a slight drop in confidence rate for 0.5-1 km from historical districts at 32% confidence, 0-0.5 km from business districts at 31% confidence, a second rule for within a historical district at 30% confidence, 0-0.5 km from a Metrorail at 30% confidence, 0-0.5 km from a business district at 29% confidence and then finally by the rule discussed in Figure 34 with 29% confidence.

It should be noted that the first rule for 0-0.5 km from business districts and the second rule for within a historical district both had lift values higher than 2. This may be due to a lower number of RHS instances for the $550-$650 pricing category; however, the greater than 0.5 gain in lift may indicate that these two rules are slightly more significant than the previous rules.

Apriori Association Analysis shows that there are multiple significant dimensions that slightly influence the price of a real estate unit. The model found 25 rules with RHS for the pricing category that had a lift greater than 1. Generally speaking these confidence rates were 5-15% higher than the support rates for the rule dimension or the pricing category dimension. These rules may be subtle, but they do give a scientific basis to argue that these factors do affect real estate pricing.

It should be noted that some rules have a confidence higher than 60%, when two pricing categories are merged. Homes within a historical district have a confidence of 65% predicting the $450-$650 pricing categories, while having a support of 20% for the within a historical district category and a support of 36% for the $450-$650 pricing range. Also, business districts within a distance range of 0-0.5 km are found for the $450-$650 pricing range with a confidence of 60% for a 36% pricing category support. This difference between rule confidence and pricing range support is very significant and indicates why historical districts and business districts had a high price average for low distances within Chapter 5.

This Chapter finds that distance from historical districts is the most significance over any multiple dimension rule. The only multiple dimension rule found with this analysis involved historical districts anyways. Competing rules that are almost as significant concern lower distance ranges for historical districts, business districts, Metrorail, hazardous facilities categories and the highest value in the height category. These single dimensions are analyzed more appropriately with the Decision Stump model. The strength of the historical district dimension may be due to the depth of the historical history within the Washington, DC area.

8. DECISION STUMP CLASSIFICATION

This portion of analysis involves a data mining technique in WEKA that evaluates the significance of single dimensions within the data. It essentially evaluates each single dimension by sorting it and iterating through each value. It utilizes a threshold to divide the data into two parts to evaluate it. Through this method the model determines the best binary split within the data to predict price.

For example, at one point the model will consider the pricing distribution of property values between a Metrorail data break of say less than 1.2 km and greater than 1.2 km. The average price may be higher with the greater than 1.2 km category. The model would thus test to see where the average price of the two subsets differs the most. It also returns the accuracy rate that considers when the wrong values appear on the wrong side of the data division. The model considers the data distribution for all nine external dimensions used within this analysis.

After the data were incorporated within the Decision Stump Classifier Model, the distance to historical locations was the most significant attribute discovered with a break at 0.5 km. The model estimated with an 87.5% relative absolute accuracy rate that when historical distance was 0.5 km or less, the price per square foot within the unit was on average $504 per square foot. Also the model estimated within the same error threshold when distance from a historical district is 0.5 km or greater, price was on average $307 per square foot.

This shows almost a $200 difference in average price. This also corresponds to the greatest correlation magnitude generated within the correlation matrix when compared with price. The 0.5 km division within the data fits nicely within the 0-4 km range of significance established by Chapter 5. Comparison to the data distribution shown in Chapter 5 shows a significant change in average price from the Within a District category to the 0-0.5km category as well as between the 0-0.5km category and the 0.5-1 km category.

The map below in Figure 35 shows a fairly consistent change in values in comparison with the 0.5 km and 4 km rings shown. However, it should be noted that the model still had 12.5% error. However, considering the large range of pricing standard deviation for each of the values within the data, this value is very low.

Figure 35 has almost all the homes in the highest pricing category within the 0.5 km range around historical districts. Also, visually speaking, most of the lowest pricing category homes are outside that 0.5 km range. It should be noted that the significant distance range found at 4.0 km is almost off of the map. The values for the 4.0-5.5 km categories only came from the eastern corner of Washington, DC.

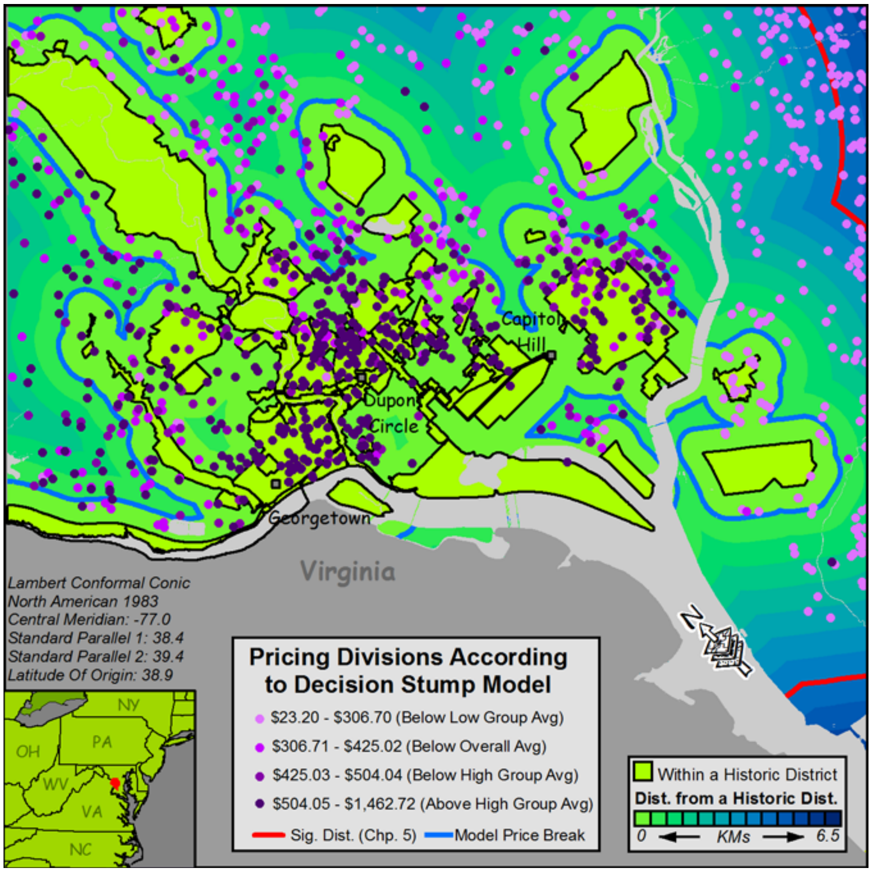

Capitol Hill
Dupont Circle
Georgetown
Virginia
Lambert Conformal Conic
North American 1983
Central Meridian: -77.0
Standard Parallel 1: 38.4
Standard Parallel 2: 39.4
Latitude Of Origin: 38.9
NY
OH
PA
WV
VA
NC
Pricing Divisions According to Decision Stump Model
● $23.20 - $306.70 (Below Low Group Avg)
● $306.71 - $425.02 (Below Overall Avg)
● $425.03 - $504.04 (Below High Group Avg)
● $504.05 - $1,462.72 (Above High Group Avg)
Sig. Dist. (Chp. 5)
Model Price Break
🟩 Within a Historic District
Dist. from a Historic Dist.
0
KMs
6.5
N

Figure 35. Historic District Influence from Decision Stump Results.

Further evaluation into the same model with the distance to historical locations dimension removed found that distance from business districts as the second most significant factor on unit price. This makes sense, since the business district distances and the historic area distances were correlated to a high magnitude of 0.69 within the global correlation model in Chapter 6. However, this relation is believed to be due to close geographic proximity between historical districts and business districts.

Within an 89.2% relative absolute accuracy rate, the model found that when the distance from business was 2 km or more, the price per square foot was on average $328 and when the distance was less than 2 km, the price was on average $502 per square foot. This was a difference of $174 per square foot. It should be noted here that the error rate generated is based upon cross dimension calculations, so a higher accuracy rate in this context is due to the fact that a dimension was removed from the model. Although, it is still interesting to consider the subsequent relations within the data.

Figure 36 shows the pricing distribution based around the overall average value of $425.02 per square foot generated by the model. As shown in the map, most of the values above the average are within the distance bound, with the exception of some properties in the northwest of the city.

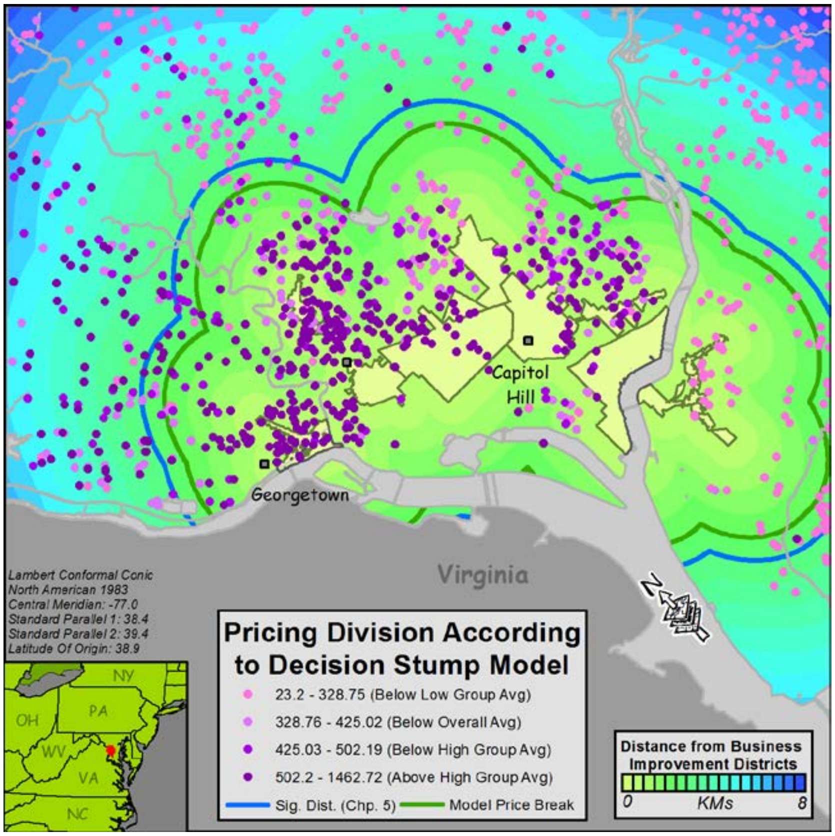

Capitol Hill
Georgetown
Virginia
N
Lambert Conformal Conic
North American 1983
Central Meridian: -77.0
Standard Parallel 1: 38.4
Standard Parallel 2: 39.4
Latitude Of Origin: 38.9
NY
PA
OH
WV
VA
NC
Pricing Division According to Decision Stump Model
● 23.2 - 328.75 (Below Low Group Avg)
● 328.76 - 425.02 (Below Overall Avg)
● 425.03 - 502.19 (Below High Group Avg)
● 502.2 - 1462.72 (Above High Group Avg)
Sig. Dist. (Chp. 5)
Model Price Break
Distance from Business Improvement Districts
0
KMs
8

Figure 36. Business District Influence from Decision Stump Results.

As further dimensions were removed from the model, distance from railroads was the third most significant with a 95.1% relative absolute error. Units within 1.5 km of a railroad were on average $339 per square foot and units more than 1.5 km from a railroad were on average $475 per square foot. This has a difference of $136 in average prices. Below is a map showing the variations of these pricing constraints. It should be noted that this distribution is not as significant as the earlier distributions since there are greater value outliers within this distribution. Of note some of the upper bound property prices in Figure 37 are likely affected by their proximity to DuPont Circle and Georgetown.

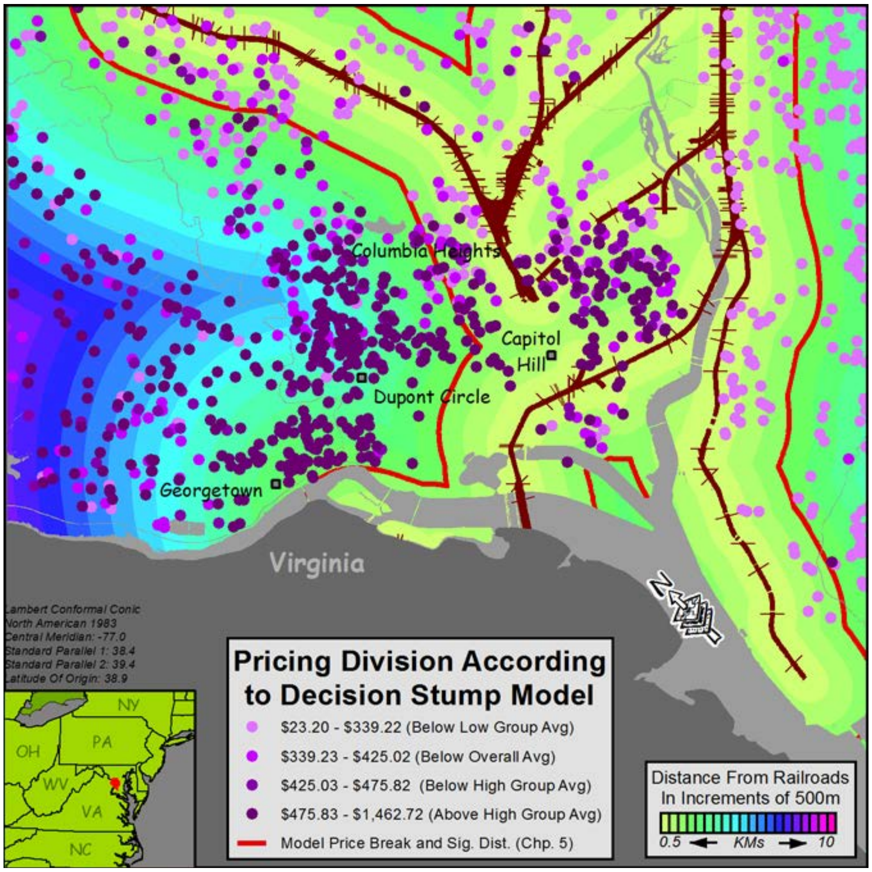

Columbia Heights
Capitol Hill
Dupont Circle
Georgetown
Virginia
N
Lambert Conformal Conic
North American 1983
Central Meridian: -77.0
Standard Parallel 1: 38.4
Standard Parallel 2: 39.4
Latitude Of Origin: 38.9
NY
PA
OH
WV
VA
NC
Pricing Division According to Decision Stump Model
● $23.20 - $339.22 (Below Low Group Avg)
● $339.23 - $425.02 (Below Overall Avg)
● $425.03 - $475.82 (Below High Group Avg)
● $475.83 - $1,462.72 (Above High Group Avg)
Model Price Break and Sig. Dist. (Chp. 5)
Distance From Railroads In Increments of 500m
0.5 KMs 10

Figure 37. Railway Influence from Decision Stump Results.

The fourth most significant dimension was height restrictions. Units restricted to a max height of 40 ft were less than $378 per square foot and units with a height restriction taller than 40 ft were $515 or more per square foot. This is a difference of $137, which is virtually equivalent with the calculation for railroads.

Figure 38 shows the distribution. Most of the property prices in the top category were within 50, 65, 90 and 110 ft zones except near the Georgetown area where most of those high property prices are within the 40 ft category. Also of note, almost all of the property prices in the lowest category are within the 40 ft category. The Georgetown properties are likely affecting the strength of this model.

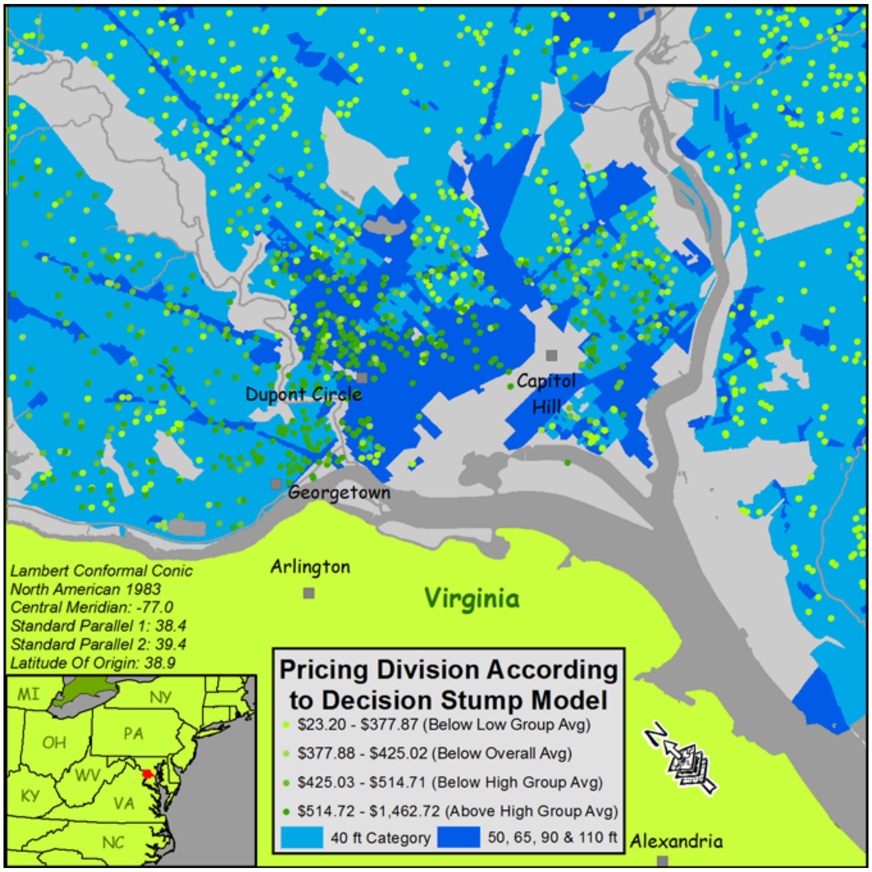

Dupont Circle
Capitol Hill
Georgetown
Arlington
Virginia
Alexandria
Lambert Conformal Conic
North American 1983
Central Meridian: -77.0
Standard Parallel 1: 38.4
Standard Parallel 2: 39.4
Latitude Of Origin: 38.9
MI
NY
OH
PA
WV
KY
VA
NC
Pricing Division According to Decision Stump Model
● $23.20 - $377.87 (Below Low Group Avg)
● $377.88 - $425.02 (Below Overall Avg)
● $425.03 - $514.71 (Below High Group Avg)
● $514.72 - $1,462.72 (Above High Group Avg)
40 ft Category
50, 65, 90 & 110 ft
N

Figure 38. Building Height Influence from Decision Stump Results.

The fifth most significant dimension after other zoning dimensions were ruled out due to high correlation was Metrorail distance. The price averaged $474 per square foot above 1132 m and $359 below that distance, which was a $115 difference in price.

Figure 39 shows the pricing distribution from this model. Most of the properties within the highest pricing category are inside the model price break line; however, there is a cluster of a few properties outside the break line within northwest Washington, DC. Also of note, most of the properties on the southeast side of the city are within the lower price category on both sides of the break distance line.

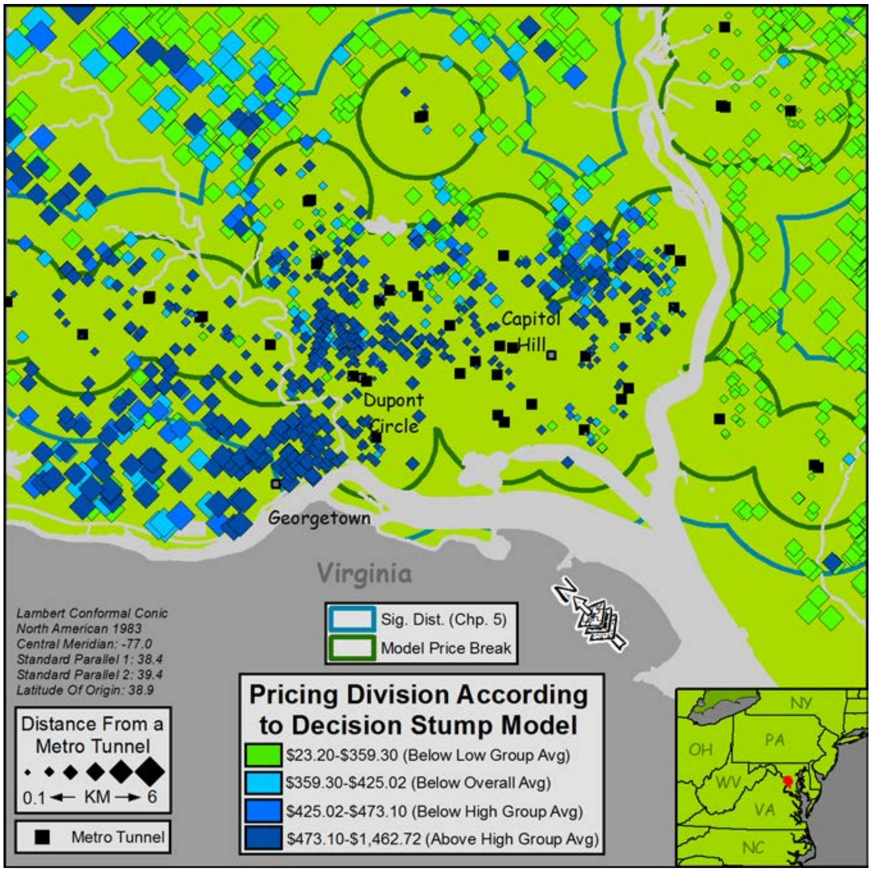

Capitol Hill
Dupont Circle
Georgetown
Virginia
Lambert Conformal Conic
North American 1983
Central Meridian: -77.0
Standard Parallel 1: 38.4
Standard Parallel 2: 39.4
Latitude Of Origin: 38.9
Sig. Dist. (Chp. 5)
Model Price Break
Pricing Division According to Decision Stump Model
$23.20-$359.30 (Below Low Group Avg)
$359.30-$425.02 (Below Overall Avg)
$425.02-$473.10 (Below High Group Avg)
$473.10-$1,462.72 (Above High Group Avg)
Distance From a Metro Tunnel
0.1 ← KM → 6
Metro Tunnel
N
NY
PA
OH
WV
VA
NC

Figure 39. Metrorail Influence from Decision Stump Results.

Thus, the social ordering of importance found by this Chapter is first proximity to historical areas, followed by influences by business districts, followed by distance to undesirable railroad noise influences, followed by influence of zoning laws like height restriction and lastly by the transportation influence of Metrorail. These general categories could likely be applied to other cities to determine consistency within the general rules found. Smaller distance ranges could be evaluated as well.

9. DISCUSSION: HISTORICAL AREAS ARE THE MOST SIGNIFICANT

The results show that there are some hidden general societal rules governing real estate prices. Generally speaking, a home buyer will seek properties with more prestige and furthermore choose locations that allow more access to conveniences for work and sociological improvement. Lesser subtle things like noise factors probably play a more secondary role within the thinking of an average home buyer.

The results of this analysis revealed that external factors upon real estate prices play a very significant role. As a study showed earlier on train accessibility and real estate price in Tokyo, these external factors are confined to the cultural conveniences within a city (Tsutsumi et al. 2011). Categorization of these cultural influences allows a generalized approach to scientifically analyze these influences.

Decision Stump modeling was able to directly consider the influence of these dimensions with a relatively high accuracy rate. With the data for Washington, DC, this thesis found only one significant interrelated dimensional influences. Thus, the results of the single dimensional method are magnified. Historical Districts not only had the most relevant pricing distribution of all the dimensions evaluated in Chapter 5, it also dominated the analytics used within every method of this thesis.

The effect of business districts on price may be influenced by historical districts, since as noted in Chapter 6 there is significant correlation between business districts and historical districts at a global and local level. An argument that railroad influence should be considered as the second greatest influence could be argued here; however, the magnitude for the correlation values were within the 0.45-0.8 range and could be argued to not be significant enough.

Though if a person were to consider this argument, he would also have to consider the 0.4-0.55 magnitude between zoning laws, business districts and Metrorail at the local level. Thus, the most significant results from this analysis under this argument are historical districts and then railroads. This thesis, however, argues against this case since the correlation between some of these elements and historical districts have a lower magnitude—there is no clear disconcerting correlation relations to negate some other dimensions as with the zoning dimensions case.

Apriori Association rule modeling found significant predictions for homes within a district and for homes 0-0.5 km from business district. These two rules had over 60% confidence for two combined pricing categories. These dimensions were the most significant dimensions in the Decision Stump model as well. Also of note, the most significant rule confidence for multiple dimensions also involved historical districts.

It is interesting that the second strongest rule in Apriori Association rule modeling was with the height dimension for the 90 ft category. Zoning influences were not as influential in the Decision Stump model. This may be due to the uneven data distribution shown in Chapter 5. Also, it is probably true that there are some significant factors within the height category since single data points within the height category do give higher confidence rates. As Chapter 7 shows, the combination of the height category with the historical category provided a higher confidence than the single dimensions in one case.

The pricing division found for historical districts is definitely more significant than all the other dimensions analyzed within this study. There was a difference of $197 for the average pricing of the division between the two historical distance categories. The only other category that was close to this division was business districts with a difference of $174 per square foot. Whereas, all of the following categories were below $150.

This range of pricing for historical districts is definitely apparent in the data distribution shown in Chapter 5. Most of the other dimensions had apparent outliers within their distribution, whereas historic districts have very little outliers. Also, most of the other dimensions didn’t directly decrease or increase in average price as significantly as historical districts, which had a significant range of 4 km that was continuing to decrease for some of the categories after the 4 km break. This 4 km variation is also simply due to the geographic extent of the data used within this study, so the influence may even be more significant.

Further comparative research into cultural categorization generalization may give further insight for the influence of real estate prices within modern cities. Data generalization of internal housing conveniences may give further insight as well. Certainly, limiting the categories to specific generalities of historical influence, business influence, undesirables influence, governmental influence and transportation will give further benefit to the basis of real estate analysis.

REFERENCES

Han, S. S. 2004. Spatial Structure of Residential Property-Value Distribution in Beijing and Jakarta. *Environment and Planning,* A(36): 1259–83.

Helbich, M., W. Brunauer, J. Hagenauer, and M. Leitner. 2013. Data-Driven Regionalization of Housing Markets. *Annals of the Association of American Geographers,* 103(4): 871–89.

Huang, R., and C. Kennedy. 2008. Uncovering Hidden Spatial Patterns by Hidden Markov Model. In: International Symposium on Geographic Information Science. *GIScience,* LNCS 5266: 70–89.

Ibeas, A., R. Cordera, L. dell'Olio, P. Coppola, and A. Dominguez. 2012. Modelling Transport and Real-Estate Values Interactions in Urban Systems. *Journal of Transport Geography,* 24: 370–82.

Lisec, A., S. Drobne, and M. Bogataj. 2008. The Influence of the National Development Axes on the Transaction Value of Rural Land in Slovenia. *Geodetski Vestnik,* 52(1): 54–68.

Liu, J., R. Wang, and T. Chen. 2010. Factors of Spatial Distribution of Recreation Areas in Peri-Urban Beijing. *Journal Geographic Science,* 20(5): 741–56.

Mennis, J. L. 2005. The Distribution and Enforcement of Air Polluting Facilities in New Jersey. *The Professional Geographer,* 57(3):411–22.

O'Kelly, M. E., and M. A. Niedzielski. 2012. Spatial Interaction Models from Irish Commuting Data: Variations in Trip Length by Occupation and Gender. *Journal of Geographical Systems,* 14:357–87.

Noulas, A., S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. 2012. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE,* 7(5): e37027.

Petrakos, G. 2001. Patterns of Regional Inequality in Transition Economies. *European Planning Studies,* 9(3): 359–83.

Spinney, J., P. Kanaroglou, and D. Scott. 2011. Exploring Spatial Dynamics with Land Price Indexes. *Urban Studies*, 48(4): 719–35.

Tsutsumi, M., A. Shimada, and D. Murakami. 2011. Land price maps of Tokyo Metropolitan Area. *Procedia Social and Behavioral Sciences*, 21: 193–202.

Viel J. F., M. Hagi, E. Upegui, and L. Laurian. 2011. Environmental Justice in a French Industrial Region: Are Polluting Industrial Facilities Equally Distributed? *Health & Place*, 17:257–62.

Winson-Geideman K., D. Jourdan, and S. Gao. 2011. The Impact of Age on the Value of Historic Homes in a Nationally Recognized Historic District. *Journal of Real Estate Research*, 33(1): 25–47.

BIOGRAPHY

Reuben Hooley graduated from Corbett High School, Corbett, Oregon in 2000. He received his Bachelor of Science from Oral Roberts University in 2004. He was employed as an All Source Analyst in the US army at Fort Campbell, Kentucky from 2004 to 2009. He was further employed as a Geospatial Analyst contactor from 2009 to 2013 in Washington, DC. with Harding Security Associates, Boeing and eventually Digital Globe.