Spring 2023

Developing the Housing Attribute and Spatial Index (HASI) Tool to Identify Characteristic Neighborhoods Using Variable Importance Factors Calculated Utilizing Random Forest Regression Modeling in ArcGIS Pro

William A. Wallace
Fort Hays State University, williamwallace625@gmail.com

DOI: 10.58809/HTDM9912

DEVELOPING THE HOUSING ATTRIBUTE AND SPATIAL INDEX (HASI) TOOL TO IDENTIFY CHARACTERISTIC NEIGHBORHOODS USING VARIABLE IMPORTANCE FACTORS CALCULATED UTILIZING RANDOM FOREST REGRESSION MODELING IN ARCGIS PRO

A Thesis Presented to the Graduate Faculty
of Fort Hays State University in
Partial Fulfillment of the Requirements for
the Degree of Master of Science

by

William Ashton Wallace
Bachelor of General Engineering, Ottawa University

Date 04/21/23

Approved ______________________
Major Professor

Approved ______________________
Graduate Dean

GRADUATE COMMITTEE APPROVAL

The Graduate Committee of William Wallace hereby approves his thesis as meeting partial fulfilment of the requirements for the Degree of Master of Science.

Chair, Supervisory Committee

Supervisory Committee

Supervisory Committee

On this day of 04/21/23

ABSTRACT

The purpose of this research is to examine the functionality in utilizing Random Forest Regression (RFR) Variable Importance (VI) values in characterizing neighborhoods based on the attributes of existing housing units by creating an automated GIS tool. An important concept that has been implemented in the past in real-estate valuation is the concept of Hedonic Price Modeling (HPM), which uses regression techniques to identify the impacts that individual attributes have on the cost of a good in a heterogenous market outside of mere utility. The benefit of this research is to produce a tool that automates the RFR process such that city planners and GIS analysis with access to ArcGIS Pro software have the capability of identifying neighborhoods that characterize specific housing value ranges with real-world examples utilizing multiple data types. From this research it was found that VI is a valid method for visualizing characteristic neighborhoods based on the housing attributes for values within a specific range, but in terms of spatial analysis other methods need to be implemented into the analysis other the VI factors.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my academic advisor, Dr. Richard Lisichenko, for believing a lowly engineer can learn spatial analysis and GIS, and for the encouragement throughout my tenure at Fort Hays State University. A thanks to all the faculty and staff within the department of Geosciences with special acknowledgment to my master's thesis committee members, Dr. Keith Bremer and Dr. Tom Schafer.

Recognition to the Ellis County IT and the City of Hays Public Works, specifically Eammon Coveney and Curtis Deines, for the sharing of data and the willingness to sit down and discuss spatial analysis problems within the City of Hays.

Finally, a big thanks to my friends and colleagues within the Department of Geosciences, my family and to my lovely wife, Abigail Wallace, for all the love, encouragement, comradery, and friendship shared throughout this process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICIES

INTRODUCTION

The Harvard University Joint Center for Housing Studies (JCHS) has produced an annual briefing called the *State of the Nation's Housing Report* for over 30 years (Herbert 2022). The purpose of the *State of the Nation's Housing Report* is to present the current housing market conditions, how the market has developed into its current state, the potential consequences for families and individuals, and finally recommendations on how to alleviate the ever-growing American housing crisis. In 2022, the "headline" issue was the record-setting increases in home prices and rental rates, which reached an all-time high of 20.6% in March of 2022, along with a housing shortage characterized by the fewest existing homes on the market since the late 1990's at just 850,000 units available on the market at the beginning of 2022 (JCHS Fact Sheet 2022). Already at-risk populations, such as low-income households and minorities found an even bigger challenge in identifying affordable housing opportunities. These issues point to a broad solution: to expand the supply of quality, modestly priced homes (Herbert 2022).

According to the Stanford Social Innovation Review, there are six pathways to making housing more affordable, and one of those is to preserve and produce affordable housing in neighborhoods (Ivory & Colton 2020). To preserve and produce affordable housing neighborhoods, two crucial questions need to be answered, where are affordable housing units located in neighborhoods currently, and what physical and environmental attributes drive the price of a housing unit. The purpose of this research to create a tool-based Geographic Information System (GIS) method for analyzing these physical and environmental characteristics that drive housing prices in different value ranges in an easily implementable and accessible tool for city planning offices and GIS analysts.

Since Kelvin J. Lancaster produced his paper *"A New Approach to Consumer Theory"* (Lancaster 1966) and introduced a novel method for evaluating consumer choices as no longer the sum of utility of a heterogenous good in a market but as the sum of its attributes, both value adding and removing, there has been a push for utilizing these theories into public planning and development. Lancaster produced the first formal theory of Hedonic Price Modeling (HPM) (*Appendix 1*) that allowed for the quantitative analysis of value in a heterogenous market, such as housing. With early pushes into Hedonic Price Modeling (HPM) occurring in the late 60s and early 70s the first approaches for HPM utilized linear regression models, either singular or multiple, to identify the impact of housing attributes on the sum of the value of a housing unit. With the advent of machine learning and Random Forest Regression (RFR) models in the early 2000's a shift occurred and has been transitioning to utilizing machine learning for HPM.

This research was designed to analyze the viability of utilizing RFR Variable Importance (VI) values in classifying characteristic neighborhoods in an ArcGIS Pro – Python script referred to as the HASI (Housing and Spatial Index) Tool. There is utility in this analysis because of the current methods used for combating the American Housing Crisis, i.e. preserve and produce affordable housing (Herbert 2022) and to implement RFR models into housing analysis. This tool was designed to be an aid for city planning and GIS analysis offices as a python script for automating the regression analysis for the different housing attributes and spatial amenities or disamenities that can drive the value of a housing unit in an isolated market, while also producing intuitive maps that can identify characteristic neighborhoods.

The environmental inputs of the HASI tool include feature classes that represent the locations of either amenities or disamenities within the data set, such as commercial parcels, parks and open spaces, utilities infrastructure, multiple living unit housing, and any other types of locations. The input for the housing attributes is a single feature class that contains all the variables for analysis. These variables can include anything from total square-footage of a housing unit to the number of bedrooms and any other potential housing attributes for analysis. Also included in the inputs for the HASI tool is the range of values put forth for analysis. This intentionally isolates housing units within a particular value range so that the characteristic neighborhoods can be identified for that value range. The outputs include two feature classes that are the copies of the original input feature class that include new distance tables fields that are defined as the distance from the center of the residential parcel to the outer edge of the nearest spatial variable for each spatial variable. The other major output is the validation tables calculated during the random forest regression models that include the coefficient of determination (*Appendix 1*) (R^2) for each run completed in the model.

The HASI tool was developed utilizing the ArcGIS Pro – Python Application Programming Interface (API) (*Appendix 1*). An API is a set of software protocols and rules that enable programming languages to interact with and run data transfers with a parent application. In the case of ArcGIS Pro, the name of the API is ArcPy and contains an extensive library that enables a Python programmer to utilize nearly all geospatial analysis tools available in the ArcGIS Pro suite. At the time of writing, ArcGIS Pro 3.0 was the most up-to-date version of ArcGIS Pro and was compatible with Python 3, and as such all programming and for the HASI tool was completed utilizing Python 3 for the ArcGIS Pro 3.0 ArcPy interface.

LITERATURE REVIEW

Introduction

Urban geographers and urban planning offices have a variety of tools at their disposal to analyze and affect change in the respective urban landscape. One such of those tools has had a long history in the way in which housing costs have been analyzed, and that is the hedonic price model. Hedonic Price Modeling (HPM) is a concept formally proposed by Kelvin J. Lancaster at Johns Hopkins University in a paper entitled "*A New Approach to Consumer Theory*" (Lancaster 1966). In Lancaster's presentation of a new theoretical approach to analyzing consumer choices, Lancaster states that "[t]he chief technical novelty" of his coined new approach "lies in the breaking away from the traditional approach that goods are the direct objects of utility and, instead, supposing that it is the properties or characteristics of the goods form which utility is derived." (Lancaster, 1966). With this statement Lancaster challenged the traditional economic viewpoint of utility and added a slight asterisk to the laws of supply and demand. That even singular items in a seemingly homogenous market can have different values based on consumer preferences for specific attributes and that this relationship can be modeled as a function of the total cost of a good or collection of goods as the sum of its parts.

Hedonic Price Modeling, Random Forest, and Other Regression Methods

What is thought to be the initial research published on "hedonic" price modeling, although the term hedonic was not used at the time of this study (Mo 2014), was published by Haas in the year 1922 at the University of Minnesota. Haas' study included the correlation of the sales prices of 160 farms in Minnesota with factors influencing prices, especially the value of buildings, the type and zoning of land, annual crop yields, the distance from markets and villages and the type or road connecting the farmland to the markets and villages (Haas 1922). Haas performed this correlation utilizing a multiple regression model, in which a correlation coefficient was found for each of the previously described variables.

While Haas' methods may have been the first implementation of what would seemingly be HPM, a formal proposition of the theory would not occur until much later when Lancaster proposed his theory of evaluating a consumer choice as previously discussed. Lancaster's proposition resulted in a strictly linear model that ignored many principles of free market competition. In 1974, Sherwin Rosen took the theory of HDP many steps forward and maintained the relationship between market drivers such as supply and demand and the more consumer-oriented theories of Lancaster by directly incorporating the idea of a "Bid-Function" as the derivative transformation of the original regression model (Rosen 1974). Rosen's equilibrium in market "shows the functions of both supply and demand and assumes a nonlinear relationship between price and inherent characteristics" (Mo 2014).

Rosen's model would lead to the publication of what has coined as the Baseline Hedonic Model (BHM) which has been utilized in a multitude of housing studies (Laszkiewicz et al. 2019, Seo et al. 2014, Yao & Fotheringham 2016, Chung et al. 2018, Aziz et al. 2020, Man et al. 2008). BHM incorporates other indices for home price modeling by including attributes such as the qualities of the neighborhood, and relative location to be locational amenities and disamenities. The BHM has been taken as the standard for hedonic price modeling in many sectors and industries, from real-estate, to development, to urban planning and economics. Hedonic price modeling has had its fair share of criticism in the past about the efficacy of the model.

Authors Stanislaw Belniak and Damian Wieczorek, faculty at the Cracow University of Technology in Poland, Department of Civil Engineering, conducted a study on the local housing market and outlined the benefits and drawbacks of the BHM (Belniak & Wiecsorek 2017). Some of the advantages are the degree of freedom that the model can incorporate based on the available data, the universality and versatility of the model for property evaluation, the possibility of updating and making corrections without compromising the model and finally the overall ease of implementation. Disadvantages listed were that the model requires large amounts of data to be valid, there is no possible way to make allowance for external factors such as interest rates or the current sociopolitical climate, and finally the model requires the use of knowledge within the field of inferential statistics to best interpret the results (Belniak & Wieczorek 2017). Although the BHM has its fair share of drawbacks, it can still be held as reliable for use within planning and economics.

While hedonic price modeling has been dominant in its implementation for the last few decades, recently there has been a push in the literature to consider other forms of modeling the relationship between product cost and attribute. Random Forest Regression Modeling (RFRM) is a form of supervised machine learning that was introduced in 2001 by Leo Breiman, during his tenure as professor at University of California, Berkeley. "Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (Breiman 2001). In recent studies, the implementation of Random Forest Regression into hedonic price modeling has gained a lot of traction for a few distinct reasons. A 2012 study by Sanglim Yoo, Jungho Im, and John E. Wagner analyzed the different techniques for variable selection for hedonic modeling in Onondaga County, NY. This study compared the accuracy of three different types of models, the traditional Spatial Hedonic Model calibrated by OLS regression, the Random Forest technique, and the Cubist methods for variable selection (Yoo et al. 2012). For the purposes of this thesis, Cubist regression was neglected because the cubist function does not exist natively within Arcpy (discussion continued in section 2.5). This study yielded the results that the two machine learning techniques (Random Forest and Cubist) outperformed the classical OLS model. Random Forest consistently produced an R^2 value higher than the other two models while simultaneously producing lower root-mean square and relative root-mean-square to quantify the amount of deviation in the model. Also tested by the research group was the validity of the variables chosen through regression, in which Random Forest also ranked higher than the other two methods, with the other added benefit of reducing Spatial Autocorrelation. Other research also supports the use of random forest regression models in the analysis of housing markets.

Direct Comparison Models

The direct comparison model is quite simply the comparison of previously sold goods or homes to a good or home that is currently on the market (Follain & Jimenez 1985). While the most common mode for estimating the price of a single good, this method is quite often not reproducible as the assumptions made about price are more qualitative in nature as opposed to quantitative and requires the recent sale of multiple goods with very similar characteristics. This one-off comparison does not analyze the extent by which agents in an open market makes decisions based on specific attributes because it does not compare the impact of attributes across a market and therefore has no predicting power outside of estimating the cost based on what nearly identical goods cost at market.

Spatial Hedonic Model

Equation 1 outlines the BHM analyzed through the Ordinary Least Squares (*Appendix 1*) (OLS) regression model is as follows (Yao & Fotheringham 2016):

$$p = f(S, N, L) \qquad \textit{(Equation 1)}$$

Where:

p - represents property price

S - represents a set of variables containing structural attributes of a property

N - represents a set of variables containing neighborhood characteristics

L – represents a set of variables containing location attributes

f – represents a standard linear regression function, commonly calibrated by the Ordinary Least Squares (OLS) technique

Random Forest Regression Models

Random Forest Regression Models (RFRM) are built upon the basis of a concept known as Classification and Regression Tree (CART) (*Appendix 1*). CART works through a process referred to as recursive partitioning of data, which is a stagewise process that systematically breaks the data into smaller and smaller subsets to determine the impact of each potential independent variable onto the dependent variable in question. Unlike many linear regression models (such as OLS) the term stagewise is used to describe the process and not stepwise because in CART models, the earlier subsets of data determined to be inferior to the leading subset is no longer referred to in steps going forward and no backwards comparisons are utilized in the decision tree to continue regression (Burke 2008). The most common method of displaying the CART output is the inverted tree figure. Figure 1 below demonstrates the typical CART output model.

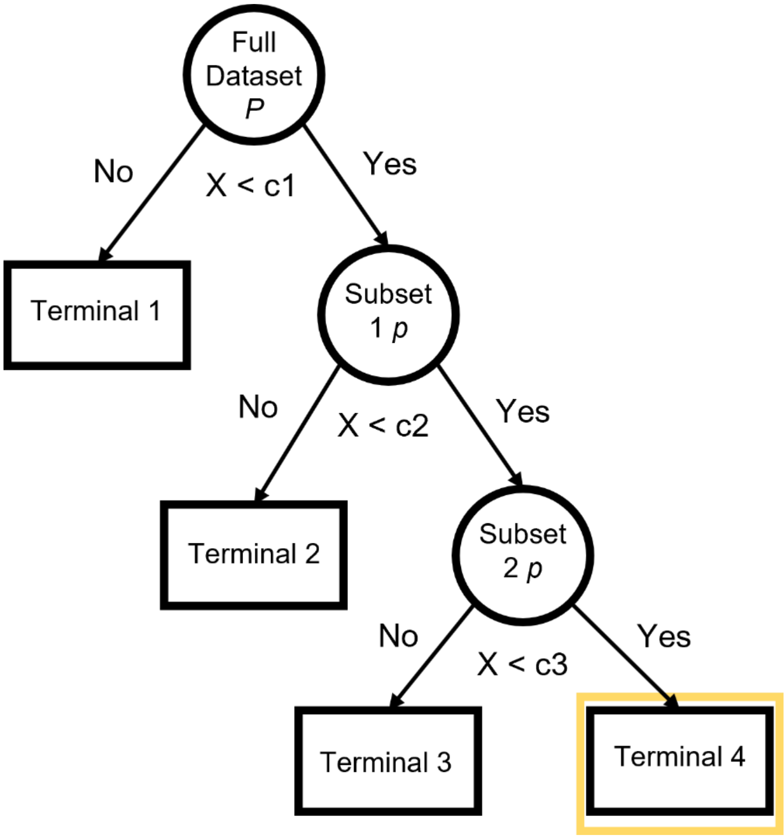


Figure 1: Typical CART Output Model (Burke, 2008)

Where:

P – represents the full dataset being analyzed

p – subset of the previously analyzed set of data

X – computed criteria value for a set

c_i – the computed reduction criteria for each step

Terminals – represent the values where continued regression would not improve on the difference between the criteria computed value x and the criteria cut off value.

Terminal 4 – (given a right-handed regression model) would be considered the highest performing variable for this tree and therefore the vote that this tree would put forth.

For the regression trees developed for this use case the ArcGIS Pro prebuilt RFRM algorithm utilizes the comparison criteria of Residual Sum of Squares (RSS) at each step in the decision tree process (ESRI). The smaller the RSS derived value, the less variance there is between the predicted dependent variable and the observed independent variable, and during the regression process, the split can be made based on the value that will best reduce variance. Each tree involved in the RFRM algorithm is constructed of data that is sampled in a manner that is referred to as "bootstrapping", in which each sample is taken and replaced back into the general population, so while each tree cannot utilize data more than one time, data points can occur multiple times within the forest (Burk 2008).

GIS Implementations of Hedonic Price Modeling/Real Estate Analysis

Hedonic Price Modeling, in both forms, linear regression and random forest regression, has been broadly implemented in GIS. Many times, professionals and researchers who implement these techniques are analyzing in two different spaces, physical attributes and characteristics of a property and the potential amenities or disamenities of the surrounding locations based on spatial variables. For this study both types of data were utilized in comparing how different physical attributes affect the price of a single-family housing unit, while also comparing how different types of land use also effect the price of a property.

In previous studies, there have been a whole host of housing characteristics that have been utilized for hedonic price modeling, across multiple types of housing and cultures. One study produced in 2019 by Edyta Laszkiewics, et al. proposed testing whether or not access to urban greenspaces play a vital role as a price determinant for apartments in Lodz, Poland. The housing attribute variables utilized included the living area normalized by natural log, the age of apartment building, the story in which the apartment is located in a building and finally the area in which the rental transaction took place (Laszkiewics et al. 2019). Other studies look at variables such as the number of bedrooms and bathrooms a property has, the ratio of the size of the house to the size of the plot and others. Table 1 below is a table of housing characteristics utilized and studies that support the use of those variables.

Table 1: Housing Characteristics and Supportive Studies

Housing Characteristic	**Categorical/ Numeric/Dummy**	**Supportive Study(ies**)
Total Appraised Value**	Numeric	(Liao & Wang, 2012),(Seo, et al., 2014)*
Age	Numeric	(Lia & Wang, 2012),(Seo, et al., 2014)*
Air Conditioning	Dummy	(Ottensmann et al., 2008)
Building Style	Categorical	--
Basement	Dummy	(Sander & Polasky, 2009)*
Basement Type	Categorical	(Sander & Polasky, 2009)*
Number of Bedrooms	Numeric	(Liao & Wang, 2012)*
Foundation Type	Dummy	(Ottensmann et al., 2008)
Number of Full Bathrooms	Numeric	(Seo, et al., 2014),(Wilhelmsson, 2014)
Garage Capacity	Numeric	(Ottensmann et al., 2008)
Number of Half Bathrooms	Numeric	(Seo, et al., 2014),(Wilhelmsson, 2014)*
Total Acres**	Numeric	(Sander & Polasky, 2009)
Number of Total Rooms	Numeric	(Wilhelmsson, 2014),(Ceh et al, 2018)
Total Sqft**	Numeric	(Liao & Wang, 2012),(Ceh et al., 2018)
Deck	Dummy	(Ottensmann et al., 2008)

*Studies denoted by * are variables supported by the Journal of European Real Estate Research*
*Characteristic Names Denoted by ** are variables that were normalized using the min–max.*
Appendix 1 contains data for all of the raw and normalized values used in this study.
Tables are not comprehensive, there are other examples of these variables used broadly in the literature.

Spatial variables are often described as either being amenities or disamenities. An amenity would be a local characteristic that has a majority positive impact on the price of a housing unit. A disamenity would be a local characteristic that has a majority negative impact on the price of a housing unit. Table 2 outlines the different spatial variables utilized for this study and the respective supportive study(ies) that demonstrate the use of similar variables. In the instance of this study, spatial variables are centered around the linear distance between spatial variables and each property evaluated in the study, and does not utilize information such as crime rates, local noise measurements, and recorded pollution rates. The purpose for the exclusion of this information is that the purpose of this study is to utilize data that is readily available and abundant in many local and regional government databases.

Table 2: Spatial Variables and Supportive Studies

Spatial Variable (Distance from)	Estimated Impact on Price (+/-)	Supportive Study(ies)
Commercial Buildings	-	(Laszkiewics, et al. 2019)
K-12 Schools	+	(Laszkiewics, et al. 2019),(Che et al, 2018)
Multiple Living Units Housing	-	--
Museum	+	(Laszkiewics, et al. 2019)
Parks and Open Spaces	+	(Liao & Wang, 2012), (Laszkiewics, et al. 2019),(Seo et al., 2014)
University (FHSU)	-	(Laszkiewics, et al. 2019)
Utilities	-	--

Appendix 2 contains data for all of the distances from each housing unit to each of the aforementioned spatial variables.
*Studies denoted by * are variables supported by the Journal of European Real Estate Research*
Tables are not comprehensive, there are other examples of these variables used broadly in the literature.

ArcGIS Pro

Geographic Information Systems (GIS) is a general term that is used to represent any software suite that is designed to create, manage, analyze, and map various types of data. GIS has a wide range of use cases, from education to wildlife management to urban planning, GIS is utilized by hundreds of thousands of companies and professionals around the world (ESRI 2023). From ESRI’s history web page, GIS has its beginnings during the early 1960’s when rudimentary computational programs began being developed specifically for early quantitative and computational geographic analysis. In 1969 ESRI (Environmental Systems Research Institute, Inc.) was founded by Jack Dangermond in Redlands California. ESRI is the company that produces the ArcGIS platform along with several other online mapping software packages and support systems, including the most up-to-date package (at the time of writing), ArcGIS Pro 3.0. (ESRI 2023).

GIS Scripting

There is often confusion between what constitutes scripting and programming. Despite Python being considered a programming language it is often utilized as a scripting language. The difference being that programming involves the development of more complex, multi-use software systems while scripting more often refers to the automating of functionality within another program (Zandbergen 2020). GIS scripting can take place in a few different locations within ArcGIS Pro itself, there exists a built in ArcGIS Pro – Jupyter Notebook interface that is accessible through ArcGIS Pro Analysis pane, and a command-line script testing window available in the same location. Jupyter Notebook is a Python interface that allows users and scriptures to test lines of code one cell at a time and allows for extensive communication in script annotation and other methods of script testing and communication. Another available method for scripting in GIS is the separate ArcPy API interface. This package is native to the ArcGIS download and comes pre-installed with a download of ArcGIS Pro. The API opens a separate scripting window that is linked to an ArcGIS tool – script interface within ArcGIS Pro.

METHODS

Introduction

This research utilized Hedonic Price Modeling as explained by the methods of random forest regression (RFR). In full context of this study, the goal was to create both a tool and a metric for analyzing the cost of a housing unit, the characteristics that establish that cost and to determine if insight into representative housing units and neighborhoods can be gained from the perspective of a city manager or GIS analyst. The study area comprised the incorporated city area of Hays, located in the western portion of Kansas, the city presented itself as an interesting case study, based on the need for developing affordable housing within city limits, while also having access to a reliable source of data analysis of the data was accomplished using a custom programmed GIS tool developed for use in the ArcGIS Pro software environment.

The Housing Attribute and Spatial Index Tool (HASI) was developed utilizing the ArcGIS Pro (3.X) python interface and utilizes the ArcPy libraries produced and maintained by ESRI. HASI will be accessible through ArcGIS Pro's toolbox and script reading interface to allow for simple access and implementation. The model was developed with ESRI's pre-built RFR algorithm at the core of calculation and the specifics for this algorithm will be discussed in further detail in this section. Also included in the methods section of this research report is a brief description of the data utilized in the study, an analysis of whether or not the data is normally distributed and a brief investigation into potential correlations occurring between variables in the data set and how that might affect the performance of the model.

Research Goals and Impact

This research was designed to test the viability of utilizing variable importance (VI) factor, calculated as a byproduct of RFR, as a method for identifying characteristic neighborhoods with housing units that fall within a specified value range within a given geographic location. For this study, a characteristic neighborhood is a cluster of single-family housing units that share characteristics close to the mean for each housing attribute and spatial variable for that range of values. A secondary goal of this research is to develop an ArcGIS Pro tool that can automate this process and compute the HASI or Housing Attribute and Spatial Index. HASI in simple terms can be defined as the normalized distance from the mean for each attribute for each housing attribute within the sample size multiplied by the respective normalized variable importance factor for that particular value range.

Successfully meeting research criteria was determined using two metrics. The first, the coefficient of determination (R^2), is characterized as the quotient between sum of squared differences between the observed and predicted values, and the sum of squared differences between the observed value and the mean of the dataset (equation 2). The second metric for validation will be the visual analysis of the distribution of the HASI values. A cluster of high attribute index scores and or high spatial index scores will indicate a neighborhood that is close to the means for the housing and spatial attributes. The results of the HASI will be aggregated across US Census Bureau Block Group polygons to visualize the ability of the regression model to classify neighborhoods. Figure 2 is a map containing the block group polygons form the 2023 TigerLine Shapefile Repository for Hays, KS.

With the introduction of HPM, a brief overview of the history of the theory and the conversation surrounding the benefits and drawbacks of different types and styles of regression analysis previously discussed, this section is dedicated to introducing the study area, a discussion on the viability of this research, and the potential impact that this study can have. The purpose of this research is to create an easily implementable and accessible tool for planners, researchers, and GIS professionals alike to analyze city-scale datasets in terms of housing characteristics that have the largest impact on housing prices, while also creating a visual tool for ‘characteristic neighborhoods’ based on housing value and what attributes drive value within a specific market. As previously discussed, Belnaik and Wieczorek in their 2017 study discussed the benefits and drawbacks of HPM. The main drawback and concern put forth by the authors relate to the relative complexity of the implementation of HPM and the proper interpretation of results to the fullest degree (Belniak & Wieczorek 2017). This research project is designed to create a tool that allows for the simplified visualization of the results of Hedonic Price Modeling

This project has many potential benefits to the practical applications of geography, GIS and statistical regression in general. As previously quoted and stated, HPM (even in the more simplified form of linear regression) takes knowledge in statistical regression to fully interpreted results. The proposed research will help to bridge this gap by creating a tool that automates the visualization of RFRM into a series of parcels organized by color dependent on the relative impact of a housing feature and how far off the average value within a specified value range a property is. The question to be answered is whether this type of visualization is a viable method for analyzing home prices based on what is affectionately referred to as the basis of Geography, or Tobler’s Law, that “… everything is related to everything else, but near things are more related than distant things” (Tobler 1970). Success will be determined based on the visual comparison between basic price mapping and mapping based on variable variance and impact, and whether there is a spatial clustering of houses with highly similar characteristics.

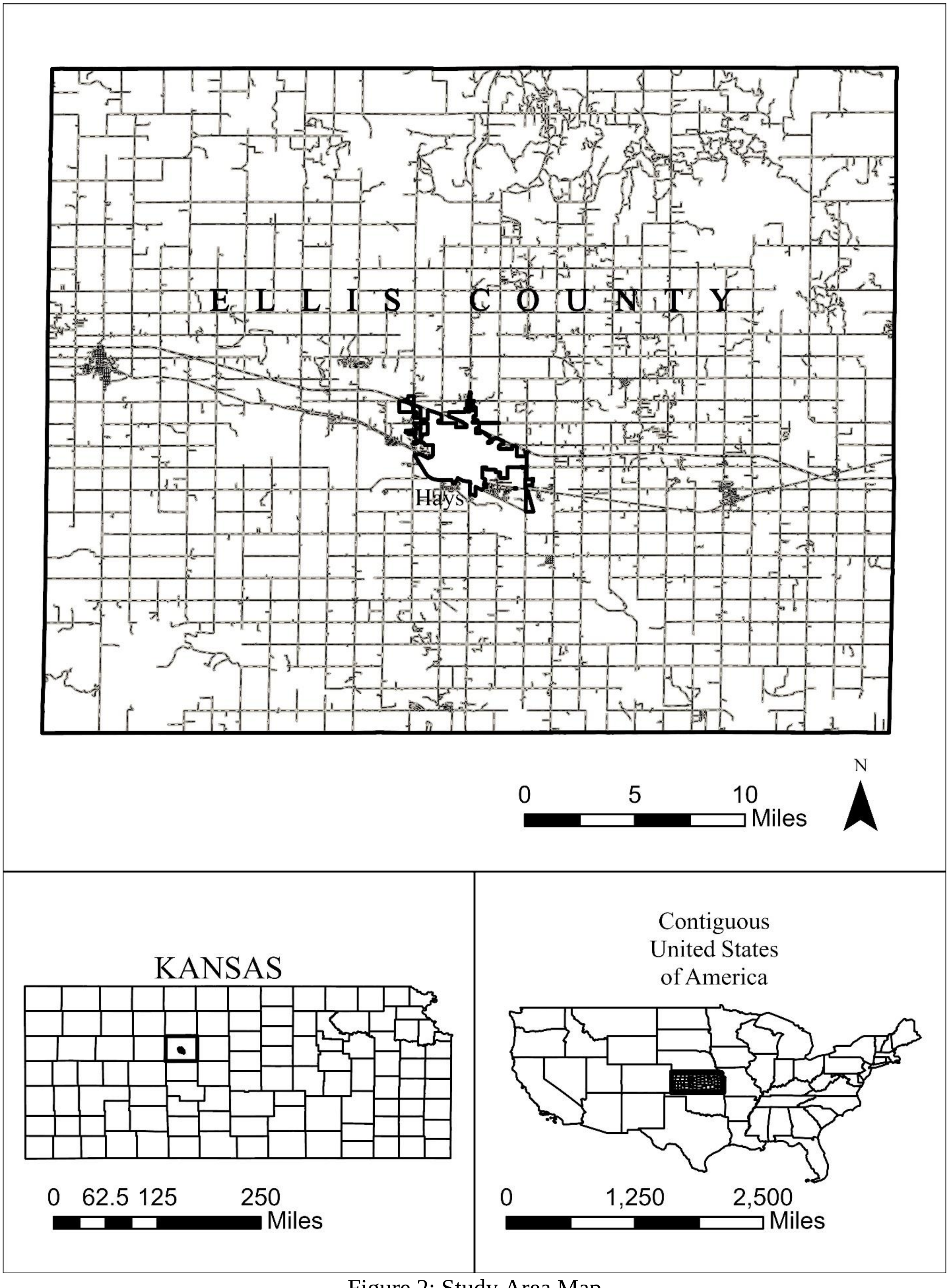


Figure 2: Study Area Map

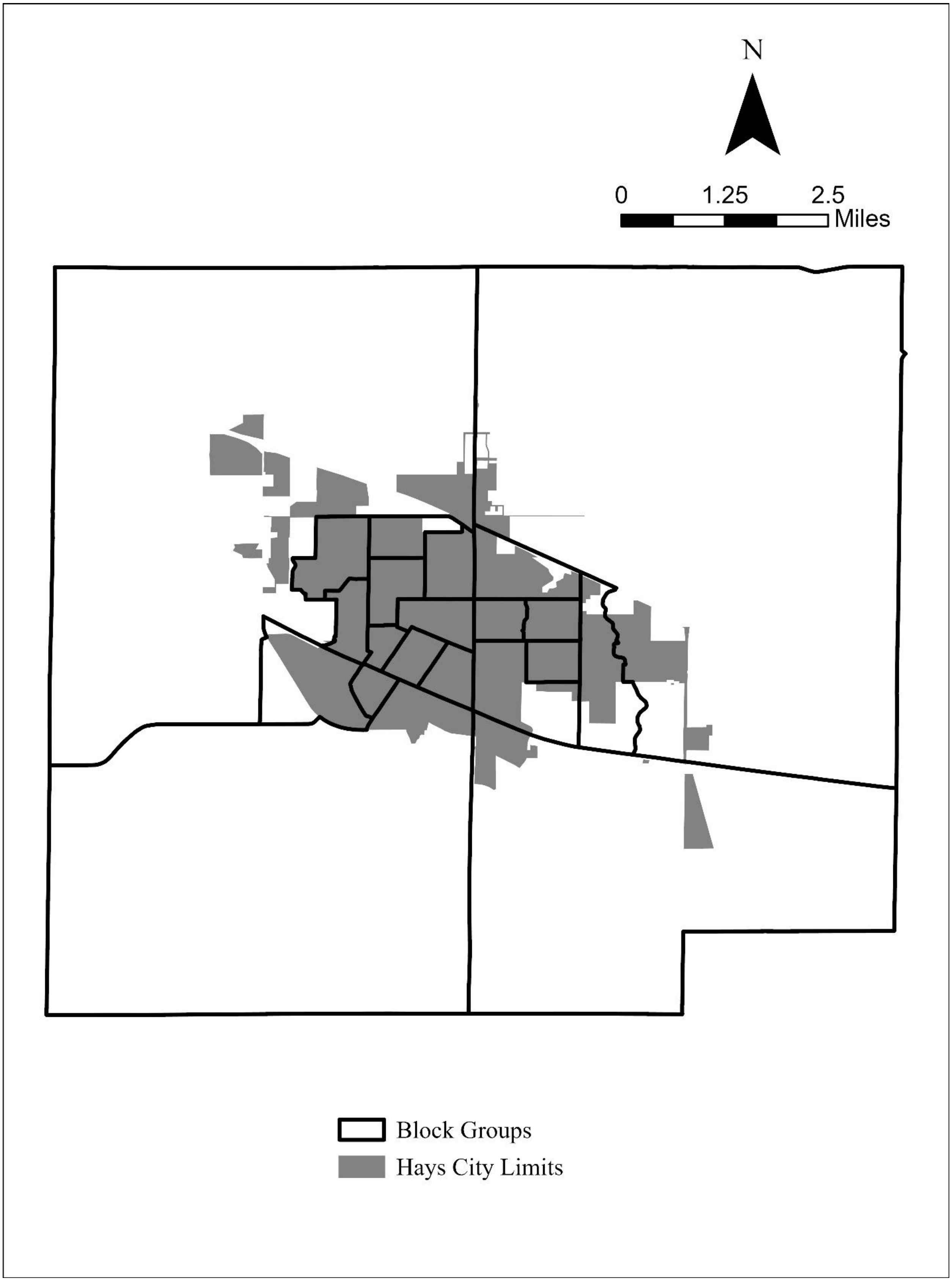


Figure 3: Hays Block Group Study Area

Study Area

The data used for this study occurs within the city of Hays, Kansas (figures 2,3 &4). Information about the population growth and housing units available within the City of Hays comes from the Hays Housing Needs Study 2021 from the Docking Institute of Public Affairs. The city of Hays housing market has a proportionately high rate of older housing units, with 40.9% being built prior to the year 1970, while only 2.1% of units were built between 2014 and 2019 (Sun 2021). Despite this slowly developing housing stock, the U.S. Census reports a 2.95% population increase from 2000 – 2020, and while the city has experienced an overall gain in housing stock, that trend has been declining in general since 2013. Future predictions for population growth show an upwards trend in Ellis County (in which Hays is the county seat) through the year 2060, while the eight surrounding counties are projected to experience a population decrease, implying that much of Ellis Counties population growth will be a result of local migration (Sun 2021). Compared to other cities in Western Kansas, Hays residents pay a higher percentage of their income towards housing and in large part with more recent development trends, that ratio will only increase (Sun 2021).

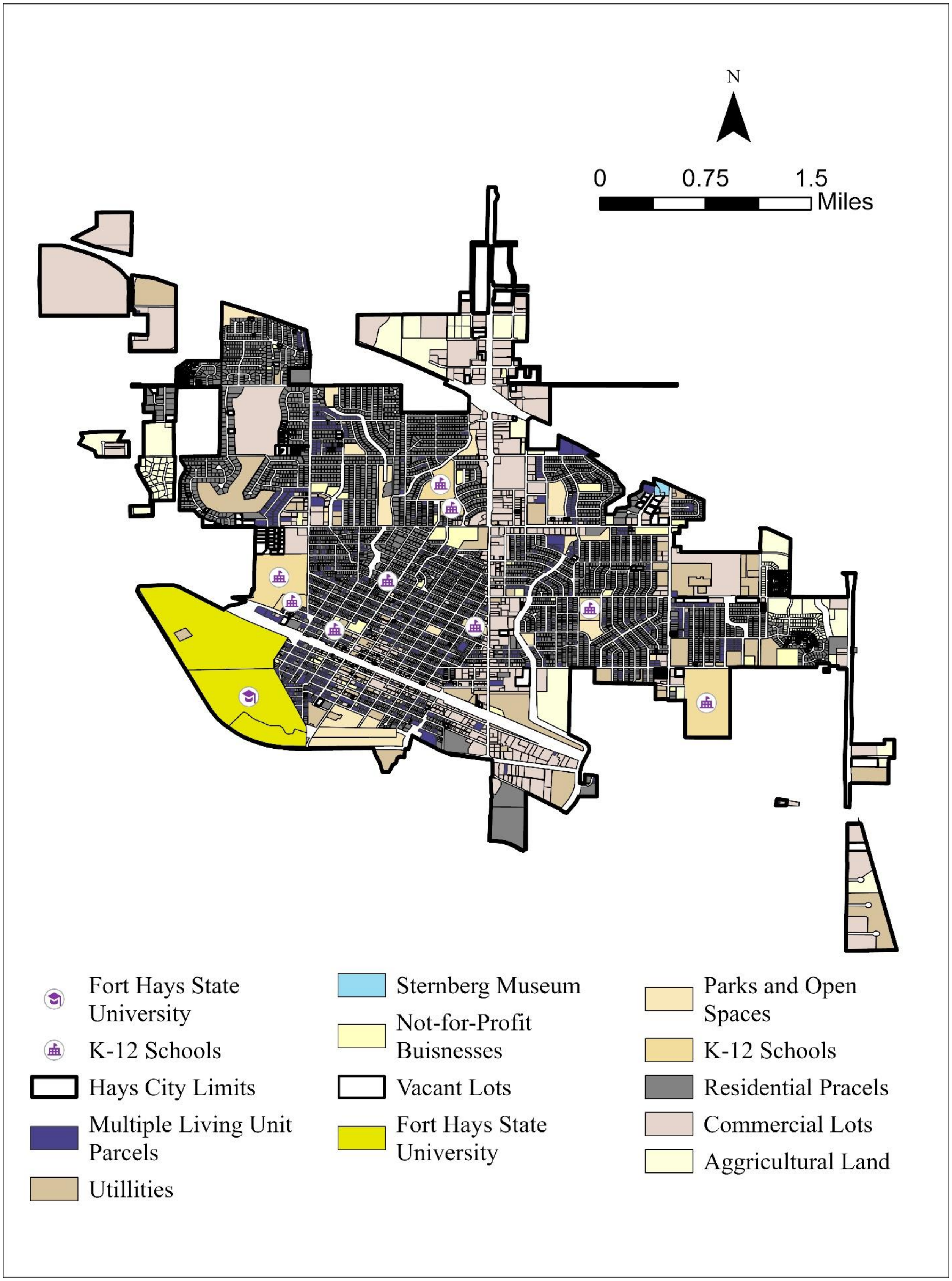


Figure 4: Hays City Map

Hays, KS (99.2786° W, 38.8684° N) is the 21st largest city in the state of Kansas with a population of 21,116 residents (2020 Decennial Census). Although the county containing the city of Hays, Ellis County, is a highly rural area with an approximate population density of 32.1 $\frac{p}{m^2}$, the city of Hays carries the designation as an urban center. The population of Hays lives within an area of approximately 8.64 m^2 with a population density of 2,443.1 $\frac{p}{m^2}$. From 2017 – 2021 the United States Census Bureau reported an owner-occupied housing unit rate of 58.1% and a rate of 77.1% of households retaining their home from the previous year. Compared to a very similarly sized town with several similar characteristics, Pittsburgh, KS, has an owner-occupied housing rate of 44.7% and rate of 67.9% of households retaining their home from the previous year across the same 2017–2021-time frame. Comparatively the housing market in Hays, KS experiences less volatility in terms of housing unit availability, and with a notably higher population growth rate than Pittsburgh, KS, would experience higher competition rates for housing units. The total stock of housing units in Hays is approximately 9,724 units (single family homes and multiple living unit buildings combined).

The economic viability of owning a home within the city of Hays as compared to Pittsburgh, the median value of owner-occupied housing in Hays (2017 – 2021) was approximately $185,700 with a median household income across the same time frame of $50,941, a rate of 3.64. Pittsburgh on the other hand had a median value of owner-occupied housing of $85,600 and a median household income of $36,657, at a housing value to income rate of only 2.33. This same comparison was made in Dr. Sun's housing report of Hays, KS where six other cities (Dodge City KS, Garden City KS, Great Bend KS, Emporia KS, Kearney NE, and Liberal KS) and the state of Kansas itself experienced comparatively lower rates of housing values to median household incomes (Sun 2021)..

Part of this discrepancy between income and housing unit cost can be due in part to the comparatively slow turnover rate in units as previously discussed, but also by the proclivity of developers to build new housing units more recently with a relatively higher value than the previously mentioned city-wide median. Of housing units built between 2000 & 2022, the median appraised value during the 2022 tax year was $347,100 and for units built between 2010 & 2022 the median appraised value was $367,070 with the median values for both timeframes nearly doubling the city-wide median household values. While some of this discrepancy of cost can be attributed to the relative age of the housing units, much of it is also dependent on the attributes of the housing units themselves. Table 3 shows how the attributes of newly built housing units have changed within the city of Hays over the past 60 years. Table 3 shows that while there has been a net increase in housing price across the decades, the most expensive and largest houses were built in the 1990's, and that all factors have had a net positive increase driving housing value up. To further demonstrate the relationship between relative age and cost, figures 5 and 6 are two different maps, figure 5 shows the age of the single-family units in Hays, and figure 6 represents the cost.

Table 3: Average SFU Housing Attributes as per Decade Built (Numerical Variables)

Decade	Value	Tot SQFT	Lot Acers	# Total Rooms	# Bath (half+full)	Garage Cap	# Bed	# Units	% Units
+1950 - 60	$143,741.7	8,144.33	0.19	6.76	1.89	0.49	3.46	2,500	%37.74
1961 -70	$195,452.5	10,066.56	0.23	7.48	2.32	1.15	3.99	976	%14.73
1971 - 80	$228,008.9	13,108.67	0.30	7.58	2.58	1.53	3.96	1,453	%21.93
1981 - 90	$259,568.2	11,843.83	0.26	7.65	2.70	1.73	4.05	679	%10.25
1991 - 00	$361,927.4	17,418.28	0.40	8.14	3.27	2.12	4.34	381	%5.75
2001 - 10	$355,698.2	15,981.65	0.37	8.37	3.12	2.13	4.54	331	%5.00
2011 - 20+	$378,285.7	13,295.08	0.31	8.46	3.09	2.35	4.74	305	%4.60
% Change	%62.00	%38.74	%38.74	$20.08	%38.93	%79.25	%27.03	TOTAL	6625

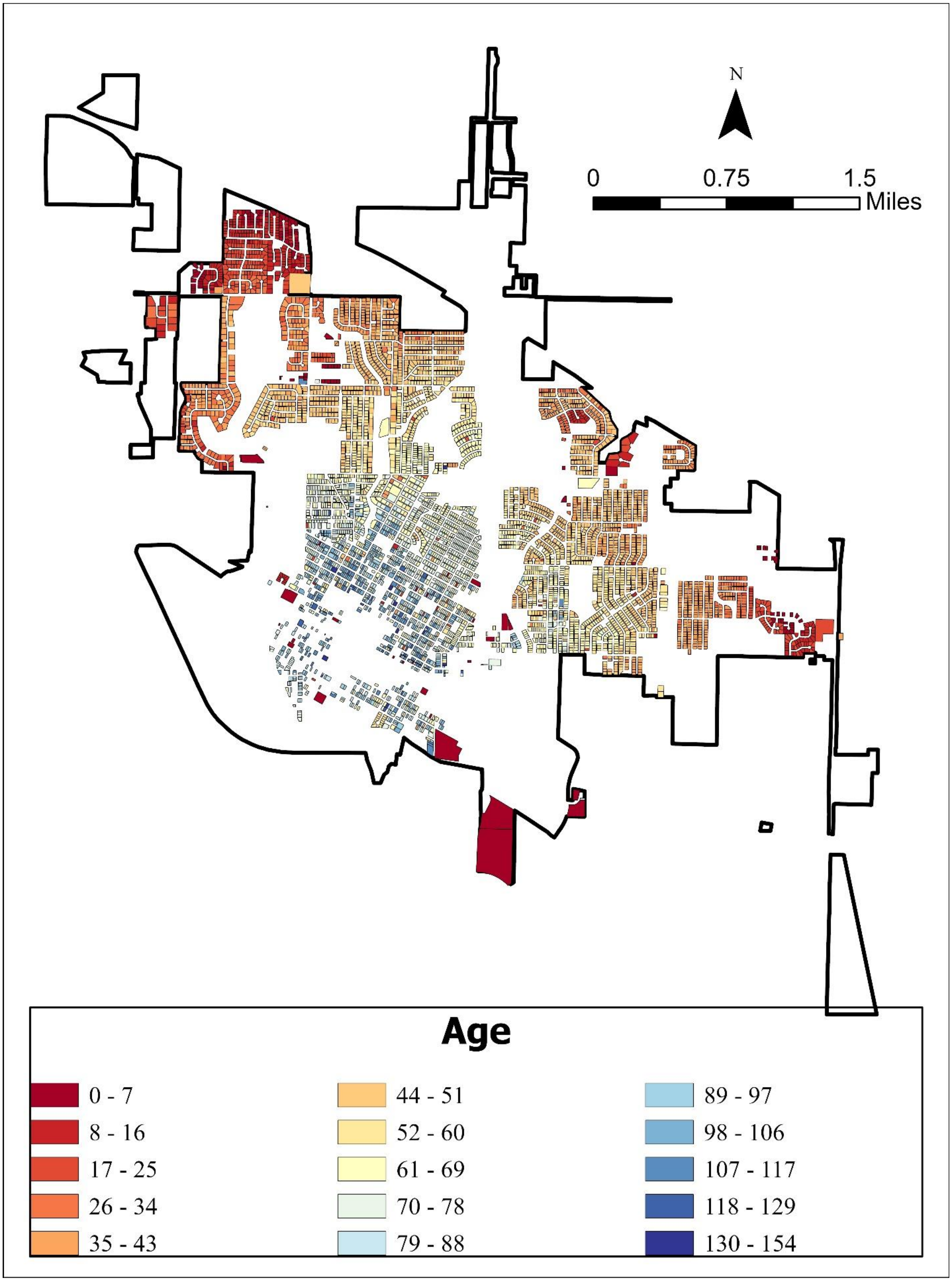


Figure 5: Hays Residential Parcels by Age

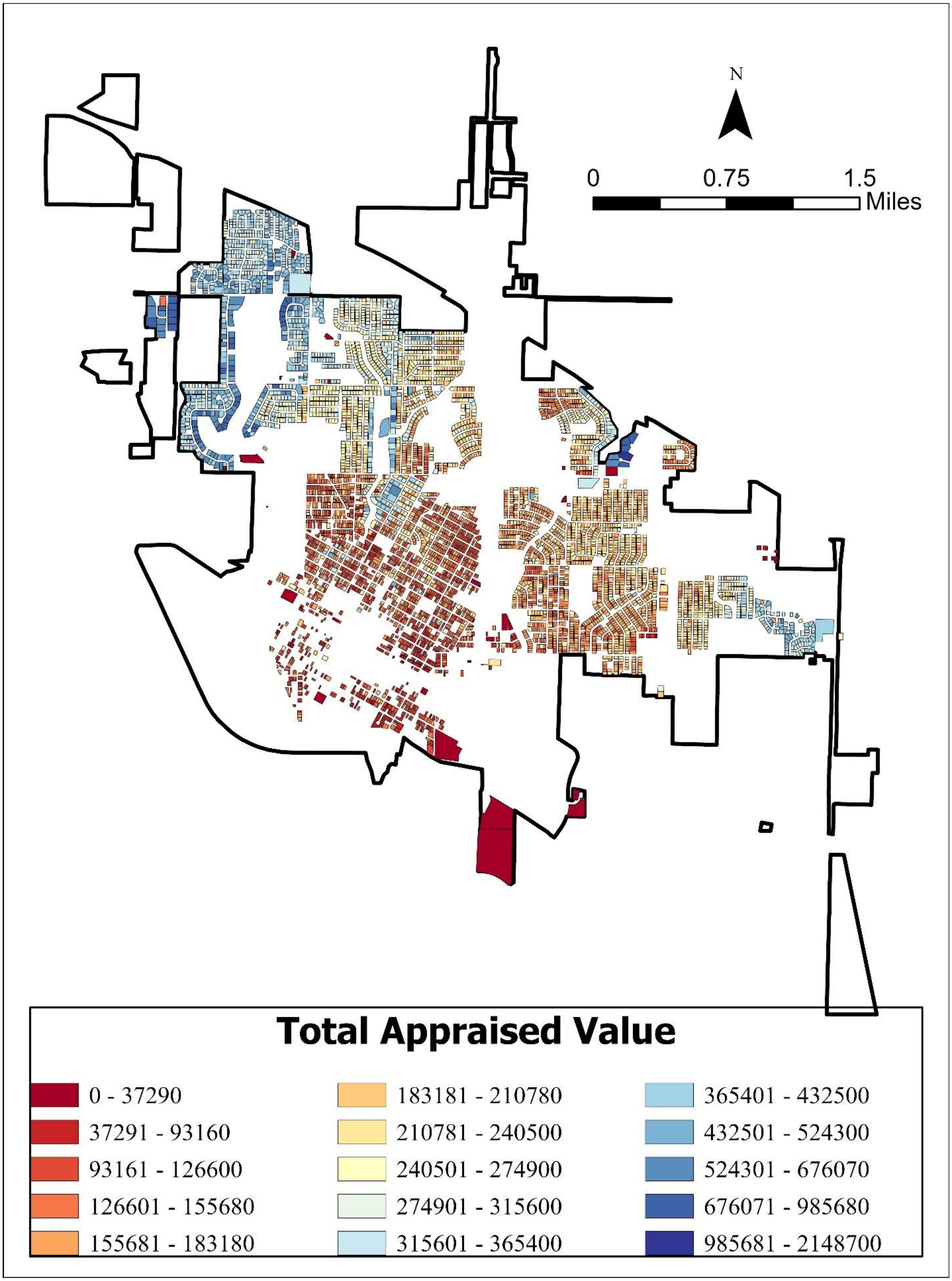


Figure 6: Hays Residential Parcels by Total Appraised Value

Data

The Hays city tax data for the 2022 tax year was utilized for analysis. As with all data types, tax data has both advantages and disadvantages. Some of the more common disadvantages related to utilizing tax data for regression analysis is determining the accuracy and timeliness of the respective data and it is sometimes difficult to tell. Depending on the size of a city and the scope of the county or city appraisers' office, it can be between five and ten years before an appraiser is able to do a site visit to a property to ensure that all tax data is up to date. While the dataset did not provide specified previous sight visit data, Ellis Counties appraisal policy is to perform a house visit at minimum every six years, if not more often. Ideally during a home visit, the inspector will perform an outside survey of the property and interview the owner about any upgrades or changes to the property in general (ellisco.net). Another disadvantage of using tax data is that it is often taken separate to local real estate trends and the nominal appraisal value does not necessarily match up in all ways to what a house could be capable of bringing on the open market.

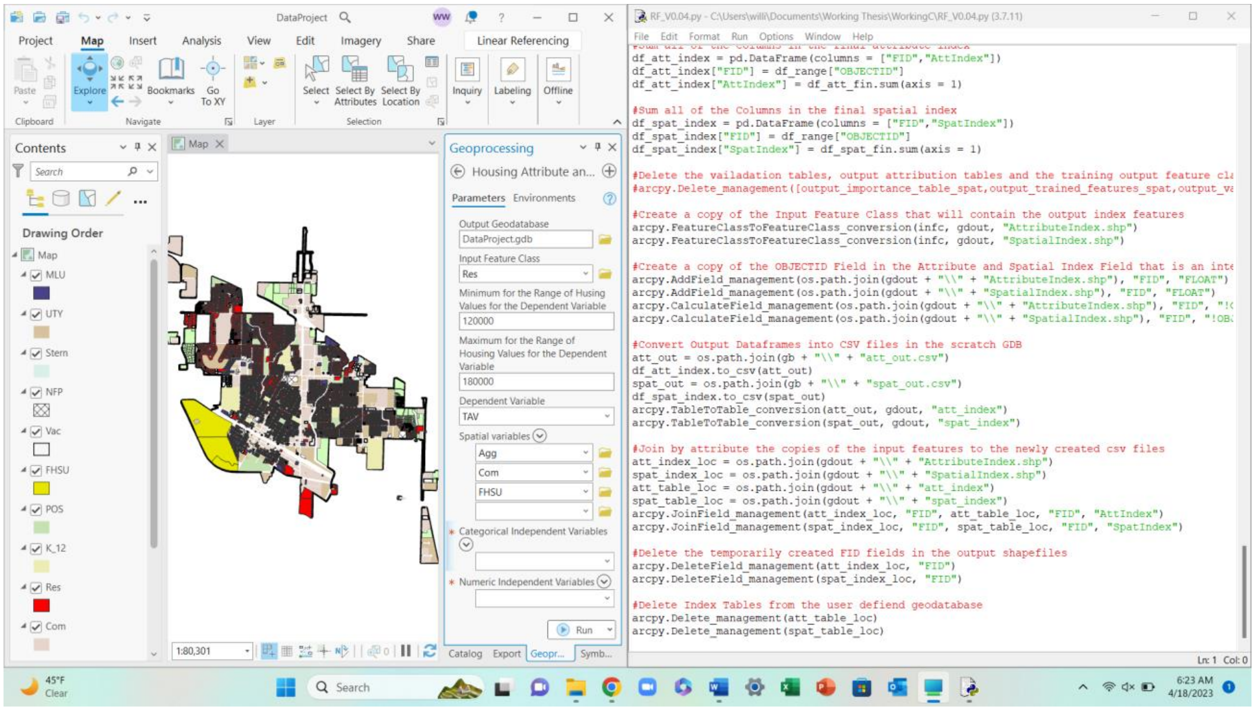
There are several advantages to using tax data for regression modeling. From a spatial perspective, tax data will often have full geographic coverage and on a very real level cover all aspects of a community, including residential parcels, multiple living units, non-profit and for-profit businesses, schools, parks, open spaces, utilities, and other types of land usage. The main reason for using tax data for this study is the fact that this GIS – Python script has been developed to be easily implemented by cities and local governments, and tax data is often the cheapest and most readily available resource for those types of entities. The following section of this research review will discuss the data, and potential corelative interactions between variables.

GIS

With a major portion of this project being devoted to developing an accompanying Python tool for analysis, there was an extensive list of software and Python packages utilized for this project. The tool for this research was developed as a Python script that can run in the background of ArcGIS Pro (3.0). Python was first selected as the scripting language for ArcMap 9.0 and an application program interface (API) was developed to implement Python 2.0. Since that first implementation of Python into ArcMap, ArcGIS Desktop also implemented Python 2.0 as the main API scripting language. That all changed with the release of ArcGIS Pro, which uses the most up-to-date Python version, Python 3.0 (Zandbergen 2020). The HASI tool was developed to work with ArcGIS Pro and therefore is structured using Python 3.0 and the most up-to-date version of the ArcGIS – Python API available.

This script was implemented using four different python libraries, ArcPy, pandas, numpy and os. ArcPy, or the ArcGIS Pro – Python API was used in this script as the basis for all spatial and statistical methods for this analysis. The ArcPy packages used in this research include the management package, the search and update cursor data interfaces, the describe packages which can be used to define datatypes and attributes about shapefiles and other forms of data, the analysis package, and the stats package. ArcPy is in version 3.1.

As for the other packages, pandas, numpy and os are all different housekeeping and data management packages. Pandas allows Python users to implement data frame storage systems and provides a wide variety of analytical and statistical tools quickly and easily. At the time of this research, pandas 1.3.5 is the package compatible with ArcGIS Pro, and with respect to this project, allowed for data management and ultimately the calculation of the final index outputs. Numpy is Python package that allows for numeric processing to be run in the background of the script and is ubiquitous to almost all Python scripting. For this script numpy 1.20.1 was used. Os or the operating systems interface is another necessary Python package widely implemented in scripting and is used to ensure that all file references and output locations follow the proper formatting for the windows operating system and file explorer. All the packages used in this study are native to the ArcGIS Pro – Python API and do not require the ArcGIS Python environment to be altered in any way, leading to even easier implementation of the HASI tool. Development of the HASI tool was done utilizing the Python scripting window for the ArcGIS Pro – Python API and a screen snip is included in figure 7.



Output Geodatabase: DataProject.gdb
Input Feature Class: Res
Minimum for the Range of Husing Values for the Dependent Variable: 120000
Maximum for the Range of Housing Values for the Dependent Variable: 180000
Dependent Variable: TAV
Spatial variables: Agg, Com, FHSU
* Categorical Independent Variables
* Numeric Independent Variables

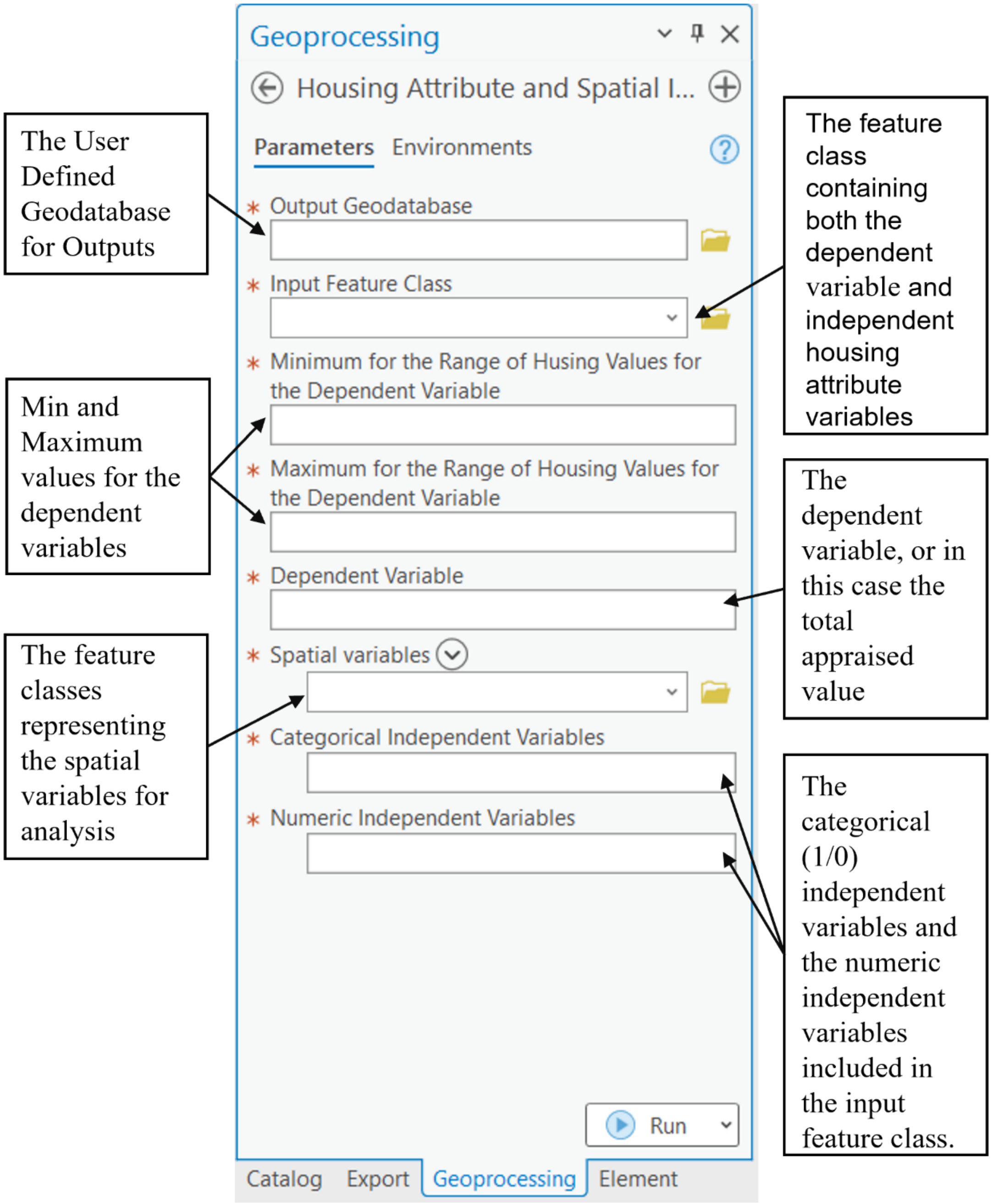
```
df_att_index = pd.DataFrame(columns = ["FID","AttIndex"])
df_att_index["FID"] = df_range["OBJECTID"]
df_att_index["AttIndex"] = df_att_fin.sum(axis = 1)

#Sum all of the Columns in the final spatial index
df_spat_index = pd.DataFrame(columns = ["FID","SpatIndex"])
df_spat_index["FID"] = df_range["OBJECTID"]
df_spat_index["SpatIndex"] = df_spat_fin.sum(axis = 1)

#Delete the vailadation tables, output attribution tables and the training output feature cl
#arcpy.Delete_management([output_importance_table_spat,output_trained_features_spat,output_v

#Create a copy of the Input Feature Class that will contain the output index features
arcpy.FeatureClassToFeatureClass_conversion(infc, gdout, "AttributeIndex.shp")
arcpy.FeatureClassToFeatureClass_conversion(infc, gdout, "SpatialIndex.shp")

#Create a copy of the OBJECTID Field in the Attribute and Spatial Index Field that is an int
arcpy.AddField_management(os.path.join(gdout + "\\" + "AttributeIndex.shp"), "FID", "FLOAT")
arcpy.AddField_management(os.path.join(gdout + "\\" + "SpatialIndex.shp"), "FID", "FLOAT")
arcpy.CalculateField_management(os.path.join(gdout + "\\" + "AttributeIndex.shp"), "FID", "!
arcpy.CalculateField_management(os.path.join(gdout + "\\" + "SpatialIndex.shp"), "FID", "!OB

#Convert Output Dataframes into CSV files in the scratch GDB
att_out = os.path.join(gb + "\\" + "att_out.csv")
df_att_index.to_csv(att_out)
spat_out = os.path.join(gb + "\\" + "spat_out.csv")
df_spat_index.to_csv(spat_out)
arcpy.TableToTable_conversion(att_out, gdout, "att_index")
arcpy.TableToTable_conversion(spat_out, gdout, "spat_index")

#Join by attribute the copies of the input features to the newly created csv files
att_index_loc = os.path.join(gdout + "\\" + "AttributeIndex.shp")
spat_index_loc = os.path.join(gdout + "\\" + "SpatialIndex.shp")
att_table_loc = os.path.join(gdout + "\\" + "att_index")
spat_table_loc = os.path.join(gdout + "\\" + "spat_index")
arcpy.JoinField_management(att_index_loc, "FID", att_table_loc, "FID", "AttIndex")
arcpy.JoinField_management(spat_index_loc, "FID", spat_table_loc, "FID", "SpatIndex")

#Delete the temporarily created FID fields in the output shapefiles
arcpy.DeleteField_management(att_index_loc, "FID")
arcpy.DeleteField_management(spat_index_loc, "FID")

#Delete Index Tables from the user defiend geodatabase
arcpy.Delete_management(att_table_loc)
arcpy.Delete_management(spat_table_loc)
```

Figure 7: Screen Snip for the open GIS window and Adjacent HASI Script

HASI Tool Performance and the User Interface

One of the major goals of this research was to generate a Python script that was capable of automating this analysis through the ArcGIS Pro – Python API interface. The HASI tool can be imported into an ArcGIS Pro project (.aprx) file by linking the downloaded tool package into the project through the Toolboxes category in the catalog pane of the ArcGIS Pro user interface (UI). Once the code package is loaded into the project, it can be accessed through the catalog pane at anytime and run just as any other form of geoprocessing tool would be. The first input in the UI is the output geodatabase, this is the location that will house the three output feature classes: AttributeIndex, SpatialIndex, and the working point shapefile upon completion of the tool. The attribute index shape file is a copy of the original input feature class with a single added field. SpatialIndex is also a copy of the input feature class with a few extra fields, the data tables utilized for running the RFR models for the spatial index. These fields are what the script refers to as distance tables (hence the subscript Dt in the field names) and were calculated in the tool as the distance from the center point of a residential parcel in the value range to the nearest edge of occurrence for each spatial variable. Also included in the SpatailIndex feature class is the field SpatIndex, which contains the spatial index values calculated using equation 3 with the distance table fields being utilized in the RFR model. Figure 8 below shows the UI for the HASI tool.

Other outputs also include four diagnostic tables related to the performance of the RFR model. The first two are the raw VI tables before any type of normalization or averaging was completed on the dataset for both the attribute and spatial variables. The other two tables contain the R^2 values for all ten trials for both attribute and spatial variables. These tables are included in the final outputs so that the user can easily analyze the effectiveness of the model.


Geoprocessing
Housing Attribute and Spatial I...
Parameters
Environments
* Output Geodatabase
* Input Feature Class
* Minimum for the Range of Husing Values for the Dependent Variable
* Maximum for the Range of Housing Values for the Dependent Variable
* Dependent Variable
* Spatial variables
* Categorical Independent Variables
* Numeric Independent Variables
Run
Catalog
Export
Geoprocessing
Element

The User Defined Geodatabase for Outputs

The feature class containing both the dependent variable and independent housing attribute variables

Min and Maximum values for the dependent variables

The dependent variable, or in this case the total appraised value

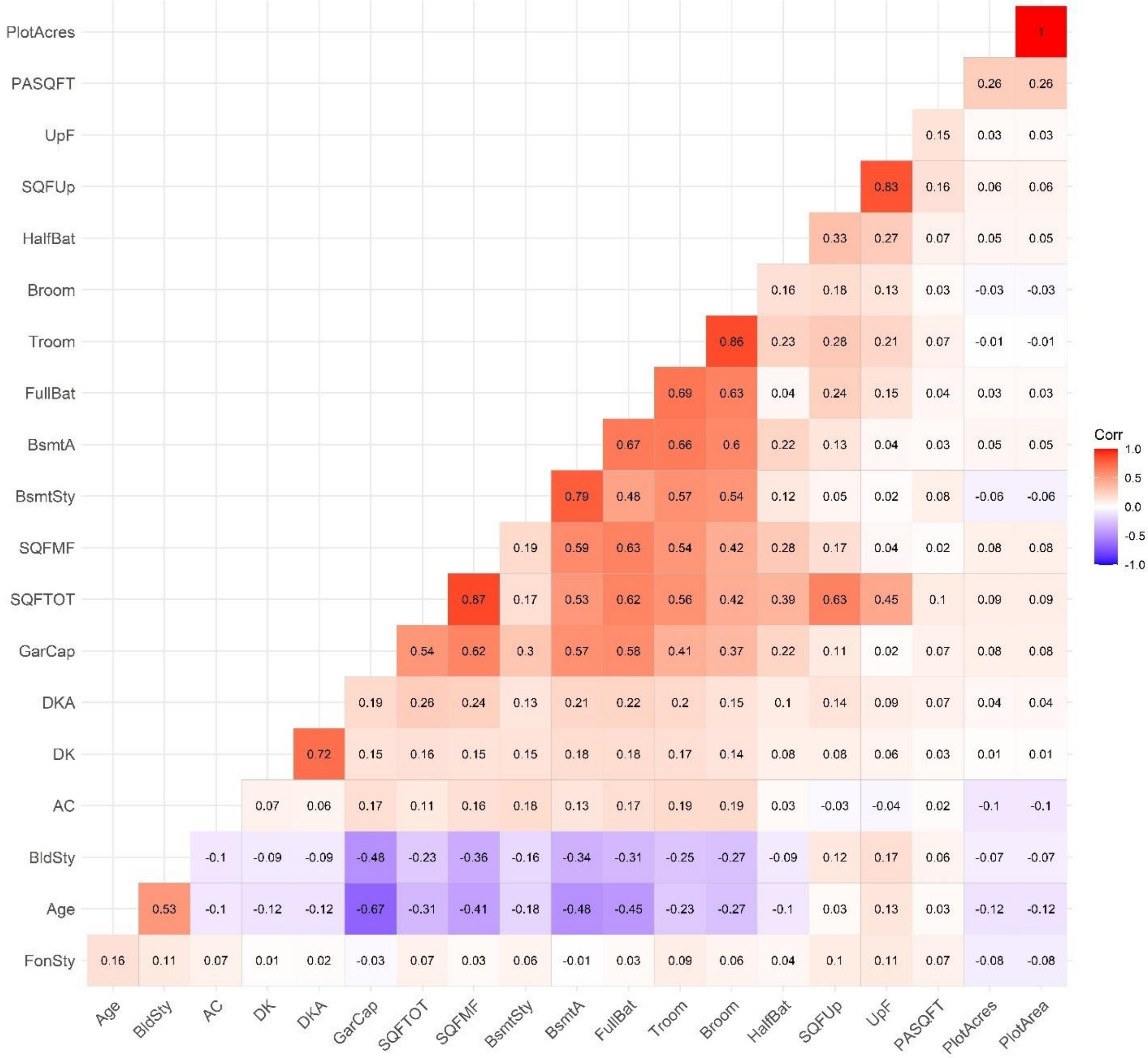
The feature classes representing the spatial variables for analysis

The categorical (1/0) independent variables and the numeric independent variables included in the input feature class.

Figure 8: HASI Tool User Interface

Data Correlation and Potential Concerns

When developing regression models, it is important to understand how variables are correlated to one another. The method utilized for correlation analysis for this research was Pearson’s Correlation Coefficient (PCC). PCC in many cases helps to detect multicollinearity of independent variables. The most common indicator of potential multicollinearity utilizing PCC is that if the absolute value of PCC is greater than or equal to 0.70, then there potentially exists high levels of multicollinearity within the dataset (Shrestha, 2020). Figure 9 displays the PCC plot for the housing attribute variables for the dataset, while Figure 10 displays the PCC plot for the spatial variable’s correlation matrix using the same method.



	Age	BldSty	AC	DK	DKA	GarCap	SQFTOT	SQFMF	BsmtSty	BsmtA	FullBat	Troom	Broom	HalfBat	SQFUp	UpF	PASQFT	PlotAcres	PlotArea
PlotAcres																			1
PASQFT																		0.26	0.26
UpF																	0.15	0.03	0.03
SQFUp																0.83	0.16	0.06	0.06
HalfBat															0.33	0.27	0.07	0.05	0.05
Broom														0.16	0.18	0.13	0.03	-0.03	-0.03
Troom													0.86	0.23	0.28	0.21	0.07	-0.01	-0.01
FullBat												0.69	0.63	0.04	0.24	0.15	0.04	0.03	0.03
BsmtA											0.67	0.66	0.6	0.22	0.13	0.04	0.03	0.05	0.05
BsmtSty										0.79	0.48	0.57	0.54	0.12	0.05	0.02	0.08	-0.06	-0.06
SQFMF									0.19	0.59	0.63	0.54	0.42	0.28	0.17	0.04	0.02	0.08	0.08
SQFTOT								0.87	0.17	0.53	0.62	0.56	0.42	0.39	0.63	0.45	0.1	0.09	0.09
GarCap							0.54	0.62	0.3	0.57	0.58	0.41	0.37	0.22	0.11	0.02	0.07	0.08	0.08
DKA						0.19	0.26	0.24	0.13	0.21	0.22	0.2	0.15	0.1	0.14	0.09	0.07	0.04	0.04
DK					0.72	0.15	0.16	0.15	0.15	0.18	0.18	0.17	0.14	0.08	0.08	0.06	0.03	0.01	0.01
AC				0.07	0.06	0.17	0.11	0.16	0.18	0.13	0.17	0.19	0.19	0.03	-0.03	-0.04	0.02	-0.1	-0.1
BldSty			-0.1	-0.09	-0.09	-0.48	-0.23	-0.36	-0.16	-0.34	-0.31	-0.25	-0.27	-0.09	0.12	0.17	0.06	-0.07	-0.07
Age		0.53	-0.1	-0.12	-0.12	-0.67	-0.31	-0.41	-0.18	-0.48	-0.45	-0.23	-0.27	-0.1	0.03	0.13	0.03	-0.12	-0.12
FonSty	0.16	0.11	0.07	0.01	0.02	-0.03	0.07	0.03	0.06	-0.01	0.03	0.09	0.06	0.04	0.1	0.11	0.07	-0.08	-0.08

Figure 9: Correlation Plot for Housing Attributes Using Pearson’s Correlation Coefficient

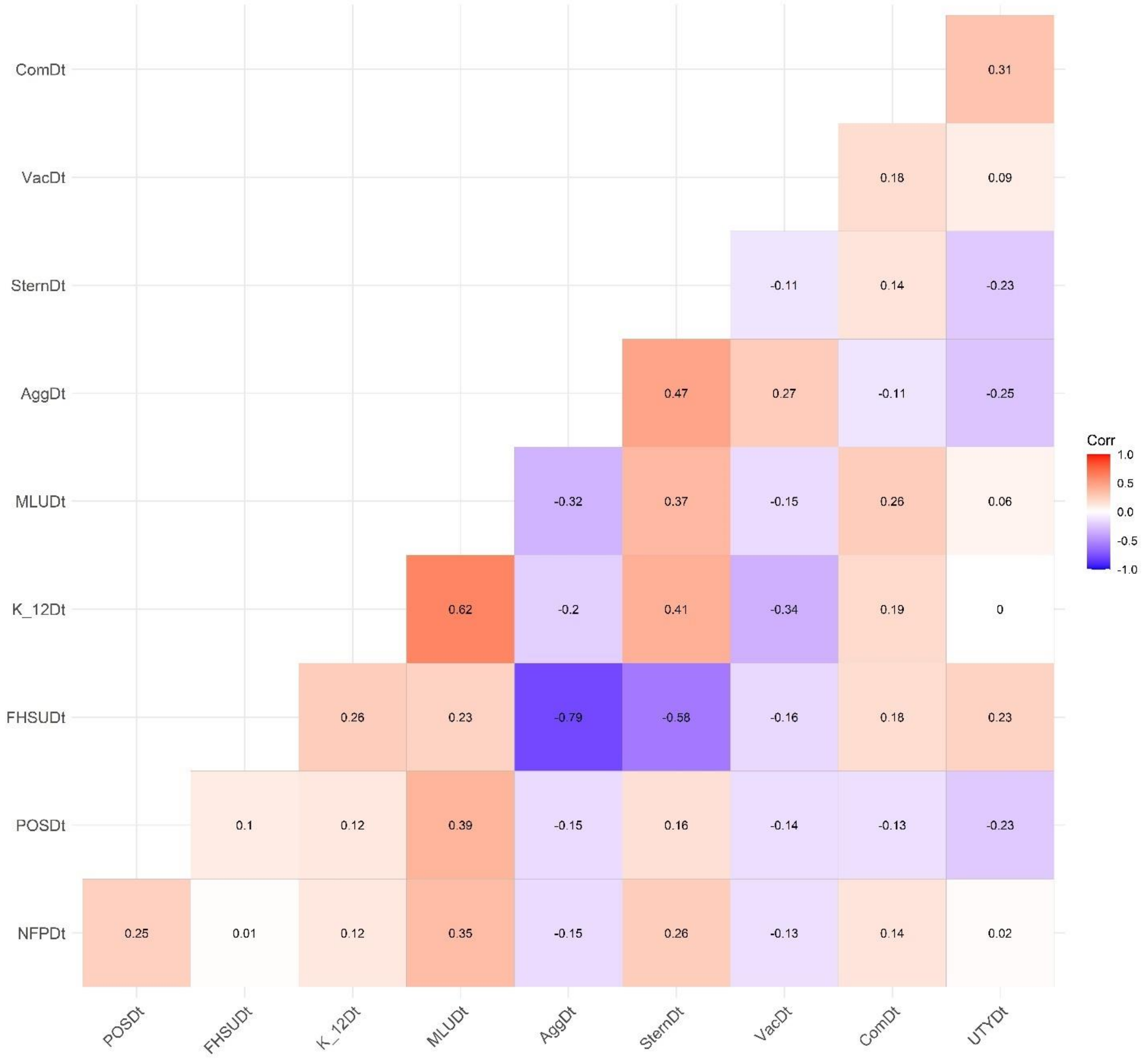


Figure 10: Correlation Plot for Spatial Variables Using Pearson's Correlation Coefficient

Figure 9 shows five PCC relationships that might indicate the presence of high multicollinearity in housing attributes: square footage of the upper floor and whether or not an upper floor exists (0.83), the total number of rooms and the number of bedrooms (0.86), the basement style and basement size (0.79), the total square footage and the square footage of the main floor (0.87), and existence of a deck to the deck area (0.72). The high level of correlation exists in large part because many of the variables represent different aspects of the same attributes. In Figure 10, only one pair of variables shares relatively high values for PCC, a highly negative correlation between the university (FHSU) and agricultural parcels (-0.79). Adjusting a dataset to address multicollinearity problems in RFR models, studies such as the one performed by Chowdhury et al, in 2020 indicate that it is not necessarily important to address multicollinear relationships in the dataset for RFR models (Chowdhury et al 2021). The reason that multicollinearity does not impact the accuracy of predictions by RFR models is twofold. First, the bootstrapping method for data partitioning introduces a high level of variance by randomly sampling the dataset dozens if not hundreds of times for each regression tree. Second, the method for which RFR models are developed utilizes a process referred to as stagewise regression, in which the relative impact of other attributes is not re-evaluated within the next split in a decision tree unlike the stepwise regression techniques utilized in multiple linear regression models.

To test the effect of relatively high levels of correlation on the output of this model, a comparative analysis of multiple combinations of attributes was performed with the results being compared. Keeping all other, non-correlated datasets the same, the rank-order of variable importance was compared across all potential combinations of correlated variables. The purpose of this analysis is to test whether correlated variables, vying for the potential to cause a split at each node through the reduction of error analysis could in turn reduce the overall score of the variable importance in each correlated variable by dividing the number of splits potential caused by each variable.

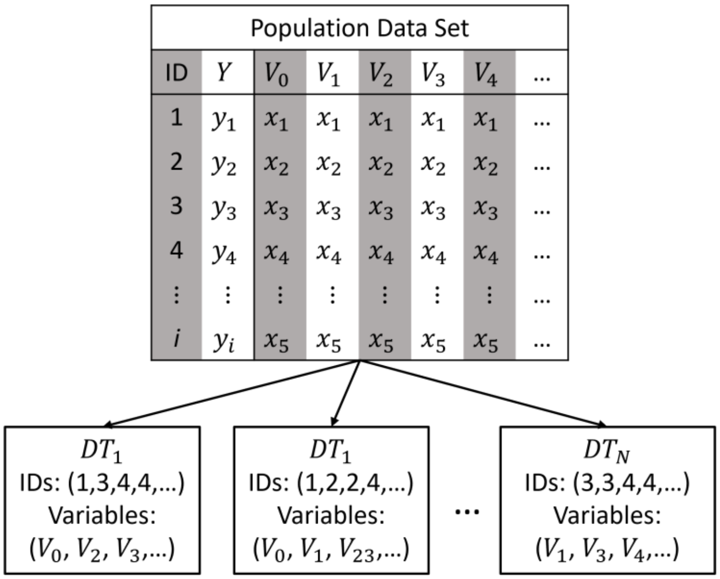
RFR Model Description and Building

Random Forest Regression (RFR) is a technique that utilizes decision tree models to make predictions based on attributes and how they relate and interact with one another in a full set of data. In the field of data science, there are a multitude of different forms and styles for decision trees, and the methods in which these trees make decisions and the parameters used for analysis vary from model to model. The topics that will be covered include the basis and underlying process of the ESRI random forest regression model, the method in which the model will be validated, how the model predicts relative variable importance coefficients, and finally the method for which the visualization process will occur and how the field values are calculated.

ESRI Random Forest Regression Model

Random Forest Regression is a mathematical and statistical process in which a multitude of decision trees individually draw conclusions about the relationship between sets of data and how the independent variables combine to act as predictors for the dependent variable in question. In RFR models, data is selected at random from the population dataset (P) into subsets (p) that is typically 2/3rds the size of the population samples with replacement. This means that as data is selected into the subset it is replaced into the population dataset and potentially selected multiple times to be fed into the same decision tree. The process of data selection is repeated for each tree to be set as the parent node for each decision tree. Along with a random set of observations (p) polled from the population dataset (P), the explanatory variables that are included for analysis within each decision tree are also subseted into random selections from the population dataset and placed into the decision tree. The number of explanatory variables utilized for each decision tree is the square root of the total number of explanatory variables available to the population dataset (P), this time without replacement so an explanatory variable can only be implemented into each decision tree one time. The observations and explanatory variable sets are then established as the parent node of a decision tree. This process is repeated across all decision trees and is referred to as bootstrapping data selection (Burke 2007). The result of bootstrapping is often referred to as low biased with high variance, as the randomness of the sampled data reduces sampling bias while introducing high levels of variance to the dataset. Figure 11 is a visual representation of the bootstrapping process.

Figure 11: Bootstrapping Diagram



Population Data Set

ID	Y	V_0	V_1	V_2	V_3	V_4	...
1	y_1	x_1	x_1	x_1	x_1	x_1	...
2	y_2	x_2	x_2	x_2	x_2	x_2	...
3	y_3	x_3	x_3	x_3	x_3	x_3	...
4	y_4	x_4	x_4	x_4	x_4	x_4	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
i	y_i	x_5	x_5	x_5	x_5	x_5	...

With the parent node for RFR established and data for modeling selected, data is then partitioned from the parent node into the respective child nodes. In instances of Random Forest where categorical classification is performed the methods for node splitting varies on the type and purpose of the model, but in terms of RFR, the splitting of nodes is based on the overall variance of the subsets of explanatory variables being compared. In the process of multiple regression (where there is more than one explanatory variable) the first step is to identify the threshold (or splitting value) for each candidate explanatory variable using the sum of squared residuals (RSS). Finding the threshold candidate that minimizes RSS is described in equation 1.

$$\tau = RSS_T = RSS_L + RSS_U \qquad \textit{Equation (1)}$$

Where:

$$RSS_L = \sum_{x<\tau}^{n} (y_i - \bar{y})^2$$
$$RSS_U = \sum_{x>\tau}^{n} (y_i - \bar{y})^2$$

Where in all cases:

τ – candidate threshold

y_i – the observed dependent variable value at observation i

$\bar{y}$ – the mean of observed dependent variables above or below the threshold

The threshold value (τ) is moved throughout the data present in the parent node until RSS is minimized. This process of RSS reduction is completed for each candidate explanatory variable until the explanatory variable threshold with the smallest RSS value is determined within the parent node. The set of data that is contained within the parent node is then split along this specified threshold value. This results in two child nodes containing subsets of data. The same process is repeated for each child node without replacement to the candidate explanatory variable that caused the split in the dataset. This process of node splitting is continued until one of two conditions is met, the first being that the maximum tree depth (or the number of child nodes) is reached, or the minimum leaf size is reached. Leaf size refers to the number of observations that are included in the split. As described in the literature review section, this process is often referred to as a form of stagewise regression, as opposed to stepwise regression, because after a split in a node is completed, the decision tree will not return to that value of variance when comparing the subsequent sets of explanatory variables (Burke, 2008). At each terminal leaf, the average value of the set of dependent variable values is averaged, resulting in the predicted value made by that tree.

This exact same process is repeated across all decision trees that encompass the entirety of the forest, with each tree putting forth its “vote” for the value of the dependent variable for an input observation based on the previously described splits in the data. The final step in the regression model making its predicted value is to average the predicted values for each prediction across all decision trees to reach an aggregate prediction. This combination of bootstrapping datasets and aggregation of predicted values can also be referred to as bagging or (bootstrapping + aggregating) (Burke 2008). As previously discussed, bootstrapping leads to datasets that are low in bias but high in variance, but the benefit of aggregating the value of the data across multiple predictions does reduce the variance in the predictions.

Model Validation

The method a random forest model is validated is dependent on the purpose and structure of the model (Burke 2007). In regression models the most used method for model validation is the R^2 validation technique or the coefficient of determination. The coefficient of determination is a measure that provides a value that describes the "goodness of fit" of a model. In terms of regression, it is a statistical measure that demonstrates how well a regression line approximates the actual data. Equation 2 represents the standard formula for calculating the coefficient of determination. The coefficient of determination will fall between 0 and 1, where the closer the value of the coefficient of determination falls to 1, the better the model fits the dataset.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \qquad \textit{Equation (2)}$$

Where:

$\hat{y}$ – predicted value for the dependent variable based on the model

$\bar{y}$ – sample means for the dependent variable

Variable Importance

Variable importance or "Gini Importance" is calculated as a byproduct from the greater process of random forest regression analysis (Mense et al 2009). Variable importance is a metric that can be used in large part as the general indicator of feature relevance and is to provide a relative ranking of the impact any given explanatory variable has on the outcome and prediction power of the regression model. Similar to the process of RFR, variable importance is a process deeply rooted in the way the regression analysis is performed and is described as “indicat[ing] how often a particular feature was selected for a split, and how large its overall discriminative value was for the classification problem under study” (Mense et al 2009).

An important factor about variable importance is that it is not particularly scaled as it is the weighted average of the effect a variable has on the purity of a regression model. Values can range from incredibly small (if the random forest and population dataset is small) to increasingly large, dependent on the scale of the forest created. To standardize this for analysis, min-max normalization was performed on the average prediction for variable importance. Another important factor is the concept of directionality in terms of the effect that a potential variable might have on the value of a housing unit. While in linear regression models, a negative correlation coefficient would indicate a negative relationship between dependent and independent variables that is not the case in RFR. This is because of the recursive partitioning nature of RFR, in that the model functions by creating a split that minimizes the residual sum of squares based on threshold values and tracks variable importance in that manner, and makes its respective predictions based on those splits not necessarily on the correlation coefficients produced by regression. Even though a spatial variable or attribute of the housing unit could potentially have a negative impact on the overall value of the unit, variable importance is directionless.

Data Ordering & Multiplication for Visualization Model

A research objective was to analyze if the variable importance factor generated by the process of random forest regression is a viable metric that can be utilized to visualize characteristic neighborhoods based on housing value and attributes. To achieve this, an index was developed that represents a numeric value for each property within a given value range based on how far any given housing unit attribute varies from the mean value of an explanatory variable using normalization, multiplied by the normalized importance factor for each analyzed explanatory variable for each value range. The min-max normalization function used in this research is presented in equation 3.

$$f(x) = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \qquad \textit{Equation (3)}$$

For the actual index value, equation 4 represents the process for calculation.

$$I_{HA} = \sum_{i=0}^{HA}((1 - f(|x_i - \bar{x}|)) * f(VI_{HA})) \qquad \textit{Equation (4)}$$

Where:

I_{HA} – the relative housing attribute index value

$f(x)$ – the min-max normalization function presented in equation 3

x_i – the value of an attribute of any housing unit

$\bar{x}$ – mean value of a housing attribute

VI_{HA} – represents the mean variable importance averaged across all tests

Equation 4 was developed for this research with two specific goals in mind, the first being to compare each housing variable not for its numeric value, but to normalize the attributes in such a way that the housing unit whose attributes lie closest to the mean be represented as the highest value for that attribute. For example, the mean square footage for a single-family unit home between $150,000 and $250,000 is approximately 12,000 square feet. Property A has a square footage of 11,000 square feet and property B has a square footage of 18,000 square feet.

The property that would more closely represent the mean for that value range for the variable total square footage would be property A. Therefore, the absolute value of the difference between the observed value and the mean would be smaller and the normalized value when compared to the entire dataset closer to zero. Subtracting the normalized value for that unit from a constant, one, results in a value that is closer to 1 and would therefore rank higher for that attribute as the characteristic for that housing unit for that attribute than say for property B where the initial difference, despite its positive or negative direction was further from the mean.

The second reason for equation 4 to take on the characteristics that it does is to normalize the variable importance in relation to one another. Variable importance is a number relative to how often a particular variable creates a split in the decision trees that compose the respective model and therefore can drastically range in value dependent on both the size and range of the parent dataset. By normalizing the dataset, the weight of each variable importance factor is retained while minimizing the effect of magnitude. The index is then calculated as the sum of values across each variable for that unit and represented cartographically.

The relative attribute housing index value can fall anywhere along the real number line between zero and HI_{Max} where HI_{Max} is represented by equation 5. Both values are highly unlikely as a value of zero would require the housing unit to occur at either the positive or negative extreme relative to the mean for each variable category, while a value of HA_{Max} would require a single unit to possess the value of the mean for each attribute within the dataset and would be valued as the sum of the normalized variable importance scores.

$$HA_{Max} = \sum_{i=1}^{VI} f(VI_i) \qquad \textit{Equation (5)}$$

RESULTS

Effects of Highly Correlated Variables on the RFR Model

For addressing potential sources of multicollinearity in this project, the Pearson's Correlation Coefficient (PCC) for each attribute and spatial variable across the entire dataset was calculated as a proxy for multicollinearity. Figure 7 identified five pairs of attributes that had a relatively high correlation coefficient, basement type to basement area (0.79), total unit square footage and the square footage of the main floor (0.87), the total number of rooms to the total number of bedrooms (0.86), the presence of an upper floor and the square footage of the upper floor (0.83), and the presence of a deck and the deck area (0.72) . The method for testing the effect that these highly correlated variables have on the model output is the comparison of the results after removing one of the highly correlated variables to a control group and the results are presented in tables 4-8. Figure 8 also shows that there exists a high PCC value existing between two spatial variables, FHSU and agricultural units (-0.79), and that relationship is further examined in table 9.

Tables 4-8 reveal a pattern in how highly correlated variables affect the final variable importance ranking used for the model. In the instance that the initially higher ranked highly correlated variable is removed from the model, then the lower ranked highly correlated variable moves up between two and three ranking places. Conversely, when the initially lower ranked highly correlated variable is removed from the model, there was no effect on the ranking of the higher ranked variable, and all lower ranked variables retained their relative ranking. This relationship held true for all pairs of variables except for the relationship between deck and deck area, where the removal of the lower ranked variable, deck, had increased the relative ranking of the deck area variable.

In terms of spatial variables, the high negative correlation between FHSU and agricultural parcels within city limits represents a large distance between the two types of land use. It should be noted, however, that the FHSU campus does contain agricultural zone lots, those lots fall outside of the City Limits of Hays and therefore outside of the study area. In table 8, there are no discernable patterns with the removal of either variable. This is in large part because the RFR model has poor predictive ability when it came to the distance tables calculated by this tool and will be discussed further in this section. With the scope of this study being to test the viability of utilizing RFR for neighborhood classification, no variables were removed due to high correlation.

Table 4: Model Performance and Variable Ranking with Different Combinations for Basement Type and Basement Area

Base		BsmtSty		BsmtA	
SQFTOT	1	SQFTOT	1	SQFTOT	1
GARCAP	2	SQFMF	2	SQFMF	2
SQFMF	3	GARCAP	3	GARCAP	3
BSMTA	4	BSMTA	4	AGE	4
AGE	5	AGE	5	BLDSTY	5
BLDSTY	6	BLDSTY	6	FULLBAT	6
FULLBAT	7	FULLBAT	7	TROOM	7
TROOM	8	TROOM	8	BSMTSTY	8
SQFUP	9	SQFUP	9	SQFUP	9
BSMTSTY	10	DKA	10	DKA	10
DKA	11	HALFBAT	11	BROOM	11
HALFBAT	12	BROOM	12	HALFBAT	12
BROOM	13	FONSTY	13	FONSTY	13
FONSTY	14	AC	14	AC	14
AC	15	UPF	15	UPF	15
UPF	16	DK	16	DK	16
DK	17				
		PCC		0.79	

Table 5: Variable Ranking with Different Combinations for Upper Floor and SQFT of Upper Floor

Base		NO UpF		No SQFTUp	
SQFTOT	1	SQFTOT	1	SQFTOT	1
GARCAP	2	SQFMF	2	SQFMF	2
SQFMF	3	GARCAP	3	GARCAP	3
BSMTA	4	BSMTA	4	BSMTA	4
AGE	5	AGE	5	AGE	5
BLDSTY	6	BLDSTY	6	BLDSTY	6
FULLBAT	7	FULLBAT	7	FULLBAT	7
TROOM	8	TROOM	8	TROOM	8
SQFUP	9	SQFUP	9	DKA	9
BSMTSTY	10	DKA	10	BSMTSTY	10
DKA	11	BSMTSTY	11	HALFBAT	11
HALFBAT	12	HALFBAT	12	BROOM	12
BROOM	13	BROOM	13	FONSTY	13
FONSTY	14	FONSTY	14	UPF	14
AC	15	AC	15	AC	15
UPF	16	DK	16	DK	16
DK	17				
		PCC		0.83	

Table 6: Variable Ranking with Different Combinations for Bedrooms and Total Rooms

Base		No Broom		No Troom	
SQFTOT	1	SQFTOT	1	SQFTOT	1
GARCAP	2	GARCAP	2	GARCAP	2
SQFMF	3	SQFMF	3	SQFMF	3
BSMTA	4	BSMTA	4	BSMTA	4
AGE	5	AGE	5	AGE	5
BLDSTY	6	BLDSTY	6	BLDSTY	6
FULLBAT	7	FULLBAT	7	FULLBAT	7
TROOM	8	TROOM	8	SQFUP	8
SQFUP	9	SQFUP	9	DKA	9
BSMTSTY	10	DKA	10	BROOM	10
DKA	11	BSMTSTY	11	BSMTSTY	11
HALFBAT	12	HALFBAT	12	HALFBAT	12
BROOM	13	FONSTY	13	FONSTY	13
FONSTY	14	AC	14	AC	14
AC	15	UPF	15	UPF	15
UPF	16	DK	16	DK	16
DK	17				
		PCC		0.86	

Table 7: Variable Ranking with Different Combinations for Square Foot of the Main Floor and Total Square Footage

Base		No SQFTMF		No SQFTTOT	
SQFTOT	1	SQFTOT	1	SQFMF	1
GARCAP	2	GARCAP	2	GARCAP	2
SQFMF	3	BSMTA	3	BSMTA	3
BSMTA	4	AGE	4	AGE	4
AGE	5	BLDSTY	5	BLDSTY	5
BLDSTY	6	FULLBAT	6	SQFUP	6
FULLBAT	7	TROOM	7	FULLBAT	7
TROOM	8	SQFUP	8	TROOM	8
SQFUP	9	DKA	9	UPF	9
BSMTSTY	10	HALFBAT	10	DKA	10
DKA	11	BSMTSTY	11	HALFBAT	11
HALFBAT	12	BROOM	12	BSMTSTY	12
BROOM	13	FONSTY	13	BROOM	13
FONSTY	14	AC	14	FONSTY	14
AC	15	UPF	15	AC	15
UPF	16	DK	16	DK	16
DK	17				
		PCC		0.87	

Table 8: Variable Ranking with Different Combinations for Deck and Deck Area

Base		No DK		No DKA	
SQFTOT	1	SQFTOT	1	SQFTOT	1
GARCAP	2	SQFMF	2	GARCAP	2
SQFMF	3	GARCAP	3	SQFMF	3
BSMTA	4	BSMTA	4	BSMTA	4
AGE	5	AGE	5	AGE	5
BLDSTY	6	BLDSTY	6	BLDSTY	6
FULLBAT	7	FULLBAT	7	TROOM	7
TROOM	8	TROOM	8	FULLBAT	8
SQFUP	9	SQFUP	9	SQFUP	9
BSMTSTY	10	DKA	10	BSMTSTY	10
DKA	11	BSMTSTY	11	HALFBAT	11
HALFBAT	12	HALFBAT	12	BROOM	12
BROOM	13	BROOM	13	FONSTY	13
FONSTY	14	FONSTY	14	AC	14
AC	15	AC	15	DK	15
UPF	16	UPF	16	UPF	16
DK	17				
		PCC		0.87	

Table 9: Spatial Variable Ranking with Different Combinations for Distance from Agricultural units and Fort Hays State University

Base		No Agg		No FHSU	
MLUDT	1	FHSUDT	1	MLUDT	1
NFPDT	2	NFPDT	2	COMDT	2
COMDT	3	K_12DT	3	NFPDT	3
K_12DT	4	MLUDT	4	AGGDT	4
FHSUDT	5	COMDT	5	STERNDT	5
UTYDT	6	STERNDT	6	K_12DT	6
AGGDT	7	POSDT	7	POSDT	7
VACDT	8	UTYDT	8	UTYDT	8
STERNDT	9	VACDT	9	VACDT	9
POSDT	10				
		PCC		-0.79	

Variable Importance and Attribute Ranking

The overall variable importance ranking was highly variable dependent on the range of housing units isolated for RFR model building. Based on the average ranking for the variable importance factor across all the datasets, the variable that on average had the best ranking was the age of the housing unit, followed by the total square foot of the unit, the basement area and the square foot of the main floor (table 10). The variable with the lowest impact across all value ranges was the presence of an upper floor and whether or not the unit had AC. In the case of Hays, KS almost all the housing units within the study area are equipped with AC and as a result consistently had very low variability, which reduced the overall chance of the reduction of error through data partitioning.

For the spatial variables analyzed, the highest performing variable was the relative distance to residential parcels designated as multiple living unit parcels. These include anything from duplexes to large apartment complexes. The spatial variables with the two lowest average ranking were the distance between the university (FHSU) and the museum (The Sternberg).

Neither of these spatial variables had a large impact on the overall attribute or spatial indices. (table 11)

Table 10: Variable Importance Ranking by Sample for Housing Attribute

Sample	Data Set	100 - 150	150 - 200	200 - 250	250 - 300	150 - 250	Average Rank
Attribute	Rank						
SQFTOT	1	3	4	2	2	5	2.83
GarCap	2	7	5	5	12	1	5.33
SQFMF	3	1	3	4	4	4	3.16
BsmtA	4	5	2	3	3	2	3.16
Age	5	4	1	1	1	3	2.5
BldSty	6	2	9	11	10	8	7.66
FullBat	7	9	8	6	7	6	7.16
Troom	8	6	6	8	6	7	6.83
SQFUp	9	13	15	15	14	14	13.33
DKA	10	10	10	7	5	9	8.5
BsmtSty	11	12	7	13	11	10	10.66
HalfBat	12	14	13	12	13	13	12.83
Broom	13	8	12	9	9	11	10.33
FonSty	14	11	11	10	8	12	11
AC	15	15	16	16	16	16	15.66
UpF	16	17	17	17	17	17	16.83

Table 11: Variable Importance Ranking by sample for Spatial Variable

Sample	Full Set	100 - 150	150 - 200	200 - 250	250 - 300	150 - 250	Average Rank
Attribute	Rank						
MLU	1	4	1	3	1	2	2
NFP	2	3	4	2	4	6	3.5
Com	3	5	2	4	6	4	4
K_12	4	7	5	9	7	1	5.5
UTY	5	1	3	1	5	3	3
Vac	6	2	7	6	3	5	4.83
FHSU	7	9	9	8	8	9	8.33
Agg	8	8	8	7	9	8	8
Stern	9	10	10	10	10	10	9.83
POS	10	6	6	5	2	7	6

Coefficient of Determination (R^2)

The coefficient of determination, or R^2, is a value that compares the slope of the trend line for the predicted dependent variable against the slope of the trend line for the sample mean for the dependent variable. In this RFR model, 10% of the population was set aside for sampling purposes. Table 12 shows the average coefficient of determination for the housing attributes for each of the value ranges analyzed. Table 13 shows the average coefficient of determination values for the value ranges for the spatial variables.

Table 12 shows a distinct pattern in terms of the accuracy of the prediction model for housing attributes. That pattern is relative to the sample size for each value range. As sample size decreased, the predictive power of the RFR model also decreased. For example, when the full data set was fed into the RFR model, an average R^2 of 0.92 was achieved, which is indicative of a model with good predictive power. The full data set incorporated 5,706 observations and provided the model with ample data. Conversely, the sample with the lowest average coefficient of determination was the value range between \$250,000 and \$300,000. This sample range only had 524 observations and reached an average R^2 of only 0.14 which indicates a model with low predictive power. The pattern is that the model loses some of its predictive power when fed smaller and smaller datasets.

Table 12: R^2 Score for Housing Attribute Runs

Sample	Full Data	100 - 150	150 - 200	200 - 250	250 - 300	150 - 250
Trial	R^2					
1	0.878698	0.338783	0.49379	0.210905	0.063458	0.71381
2	0.928562	0.320648	0.370543	0.315108	0.116572	0.731633
3	0.911804	0.408481	0.467687	0.244533	0.10069	0.705301
4	0.918105	0.447661	0.53946	0.060576	0.151049	0.67191
5	0.914549	0.586214	0.43055	0.212428	0.317778	0.738743
6	0.926142	0.591312	0.377624	0.207376	0.057378	0.69299
7	0.933147	0.444979	0.55397	0.361513	0.060764	0.715113
8	0.927077	0.421861	0.445618	0.192197	0.207571	0.699267
9	0.925481	0.384362	0.253625	0.244694	0.260767	0.706906
10	0.890153	0.308576	0.471611	0.316276	0.078124	0.726075
Average	0.9153718	0.4252877	0.4404478	0.2365606	0.1414151	0.7101748
n of units	5,706	1,139	1,506	1,106	524	2,610

When it comes to the coefficients of determination calculated for the spatial variables, even in instances of the full dataset being utilized, to have very low predictive power. While the full dataset did achieve the highest average R^2 that value was only 0.06, which indicates a model that does not have much predictive power table (13).

Table 13: R^2 Score for Spatial Variable Runs

Sample	Full Data	100 - 150	150 - 200	200 - 250	250 - 300	150 - 250
Trial	R^2					
1	0.039957	0.000067	0.000016	0.006606	0.025552	0.021528
2	0.061737	0.000336	0.022121	0.00076	0.003754	0.0196
3	0.045329	0.002106	0.000231	0.010371	0.001903	0.020829
4	0.06359	0.000001	0.006403	0.00006	0.046767	0.028737
5	0.078872	0.035577	0.005731	0.004524	0.025008	0.060546
6	0.079593	0.000097	0.000979	0.010774	0.004751	0.024378
7	0.060121	0.009977	0.00017	0.006046	0.032702	0.131084
8	0.074357	0.000046	0.003247	0.000413	0.004918	0.04486
9	0.071652	0.001436	0.005531	0.007262	0.006879	0.037262
10	0.107895	0.016062	0.000013	0.000626	0.122562	0.025951
Average	0.0683103	0.0065705	0.0044442	0.0047442	0.0274796	0.0414775
n of units	5,706	1,139	1,506	1,106	524	2,610

Parcel Level Result Maps

Figures 12-23 are the result maps for the attribute and spatial indexes created utilizing the methods presented in equation 3. Figures 12 and 13 represent the attribute and spatial index result maps for the RFR model developed utilizing the full data set. With the model being developed utilizing the entire for result maps 12 and 13, little can be said about the overall distribution of the range of values being identifiers for characteristic neighborhoods, but there do exists hotspot neighborhoods with a relatively high attribute and spatial index scores. Figures 14 and 15 show the result maps for the model that was developed utilizing housing units between \$100,000 and \$150,000. Nearly all the housing units that fell between this value range are located in the southwest corner of the city, in the distinctly older portion of Hays. In terms of the attribute and spatial indices, the attribute index demonstrates slight clustering near the center of these housing values, while there is no clustering in the spatial index scores.

Figures 16 and 17 show the result maps for the model that was developed utilizing housing units between $150,000 and $200,000. Isolating this value range shows a general clustering for this value range on the east – southeast portions of the city and demonstrates moderate clustering in both the attribute and spatial index maps. Figures 18 and 19 show the result maps for the model that isolate housing units between $200,000 and $250,000. Housing units within this value range tend to exist in the east and north within city limits and the model shows relatively good clustering in specific neighborhoods. Figures 20 and 21 represent the value range between $250,000 and $300,000 dollars, the value range with the fewest total observations. Most housing units within this value range are in the north and northwest sections of the study area. There is a clustering of high attribute index housing units to the north and little to no clustering within the spatial index values. The final two figures, 22 and 23 are the result maps for the model developed with the value range of housing units between $150,000 and $250,000 dollars. This value range is distributed evenly across the study area, but there does exist a gradient for both the housing attribute and spatial indices that begin low towards the southwest and increase to the northeast with a cluster on the northeast side of the study area.

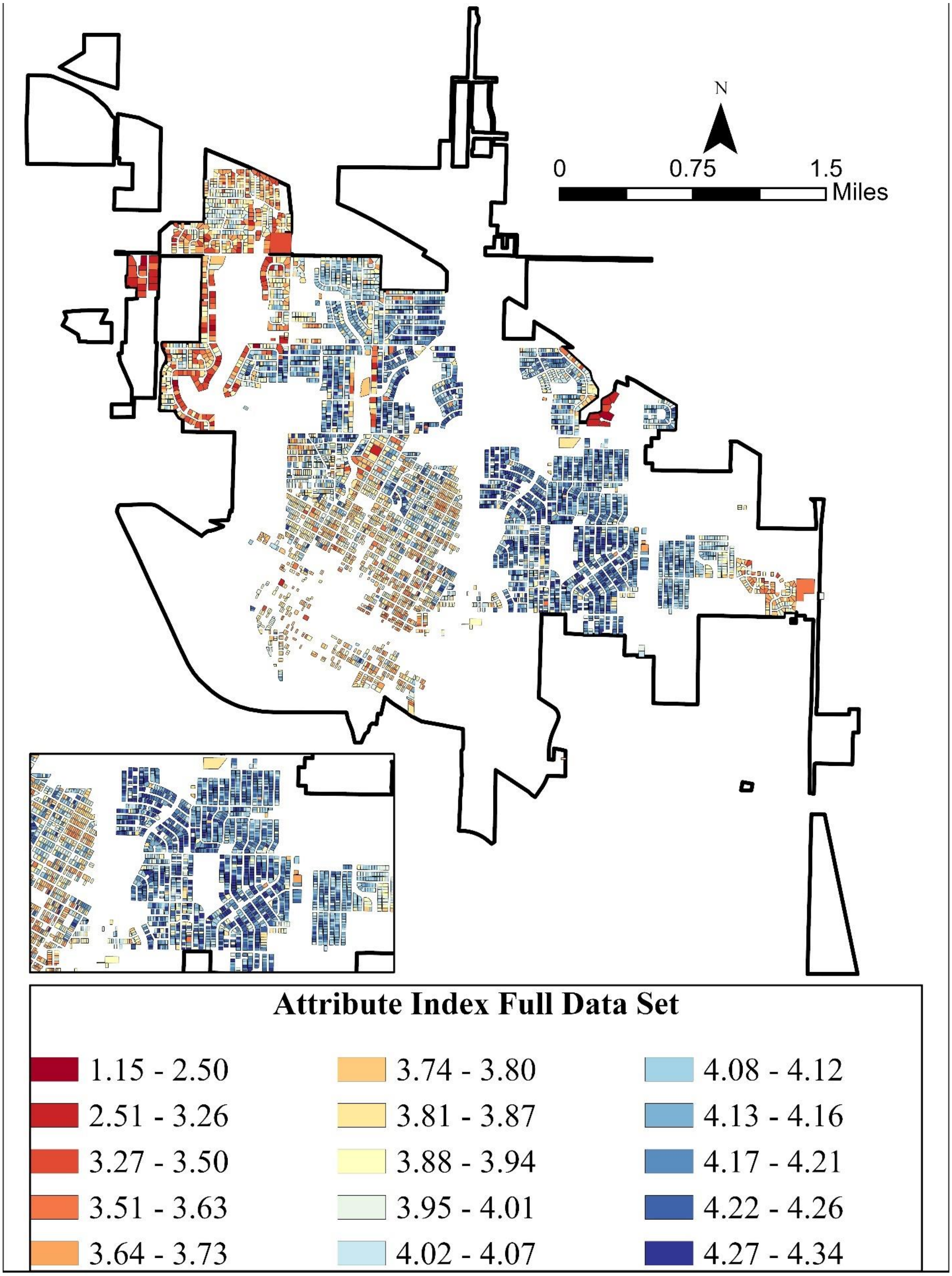


Figure 12: Attribute Index for Full Data Set

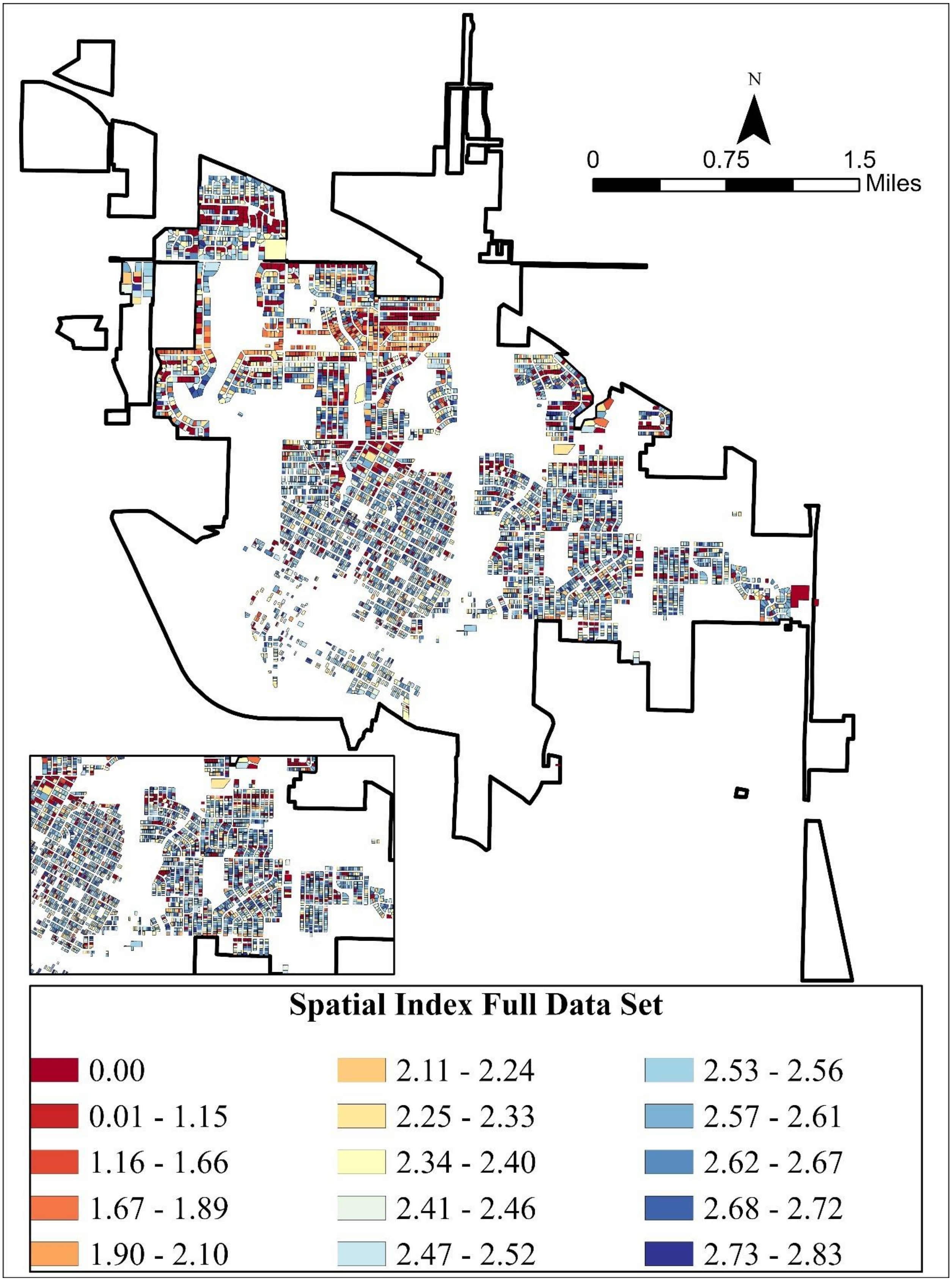


Figure 13: Spatial Index for Full Data Set

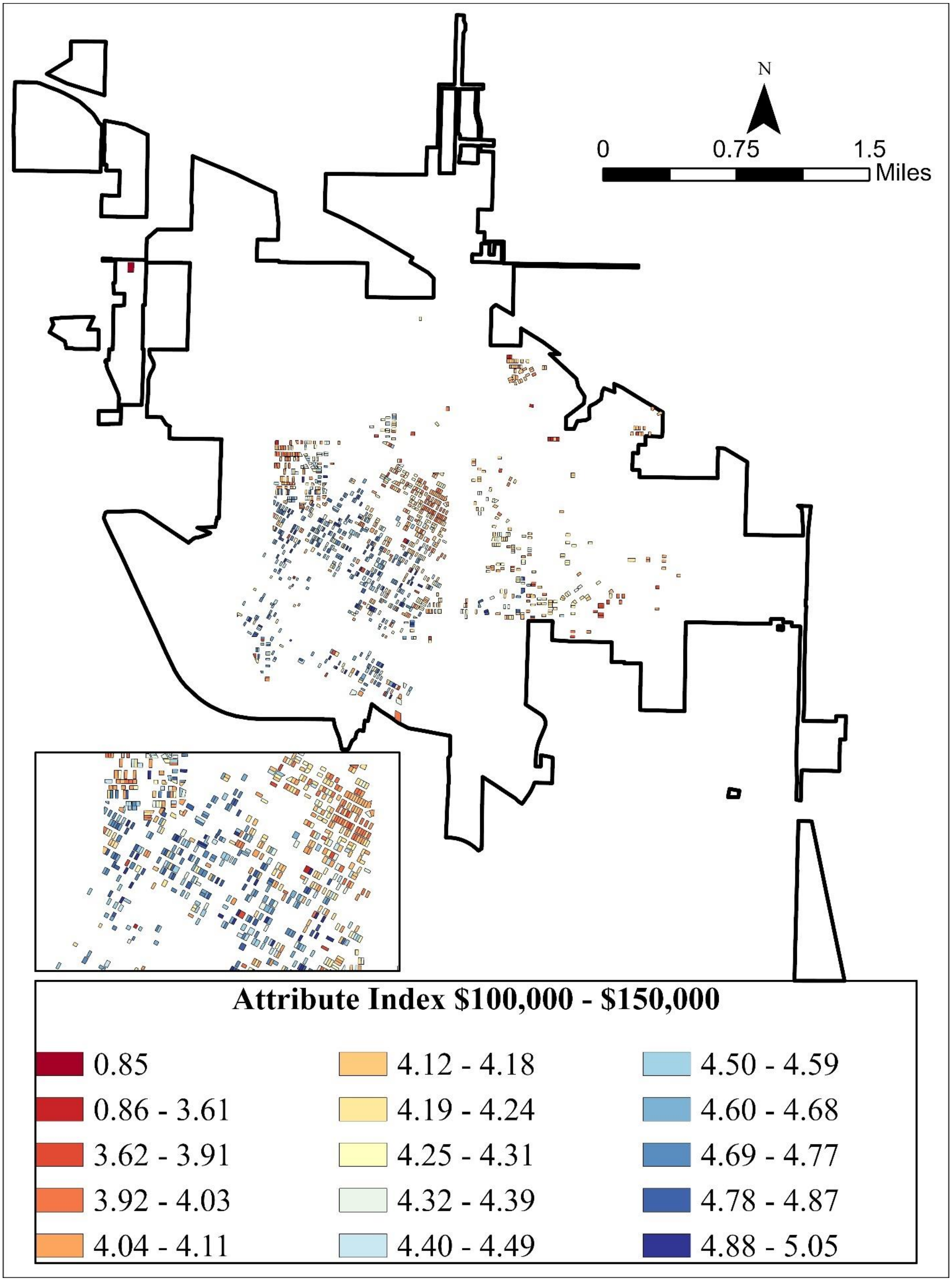


Figure 14: Attribute Index $100,000 - $150,000

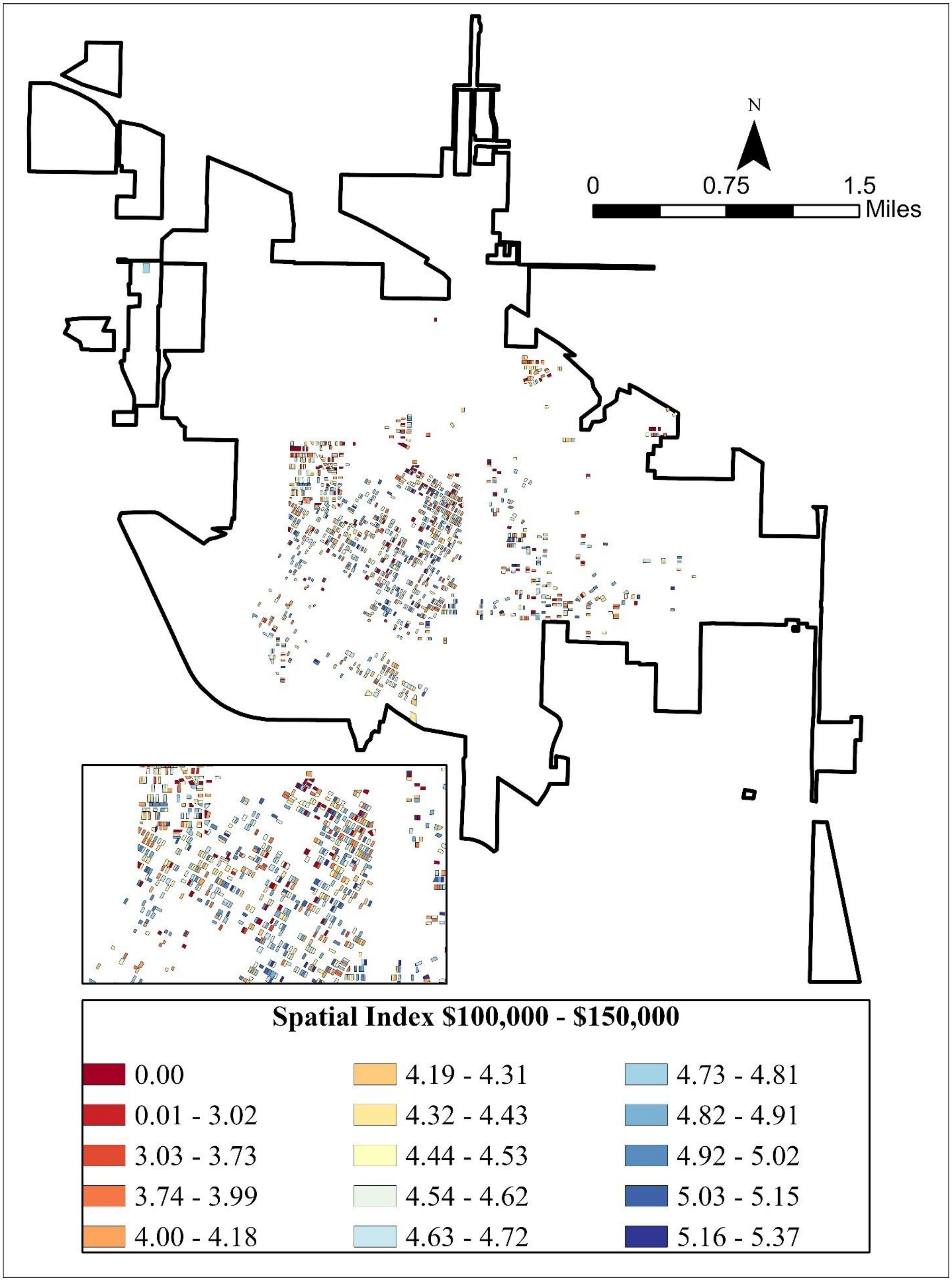


Figure 15: Spatial Index $100,000 - $150,000

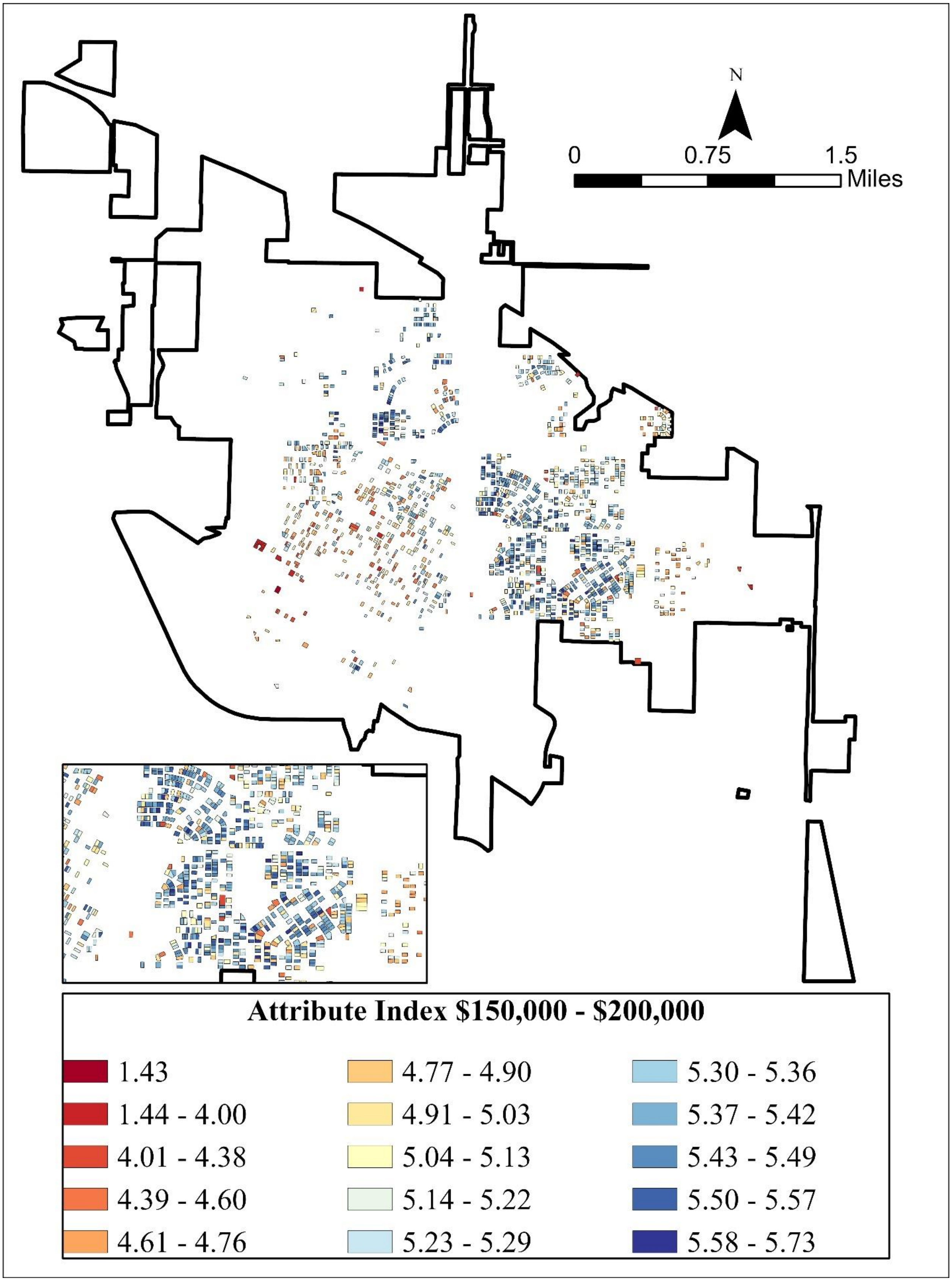


Figure 16: Attribute Index $150,000 - $200,000

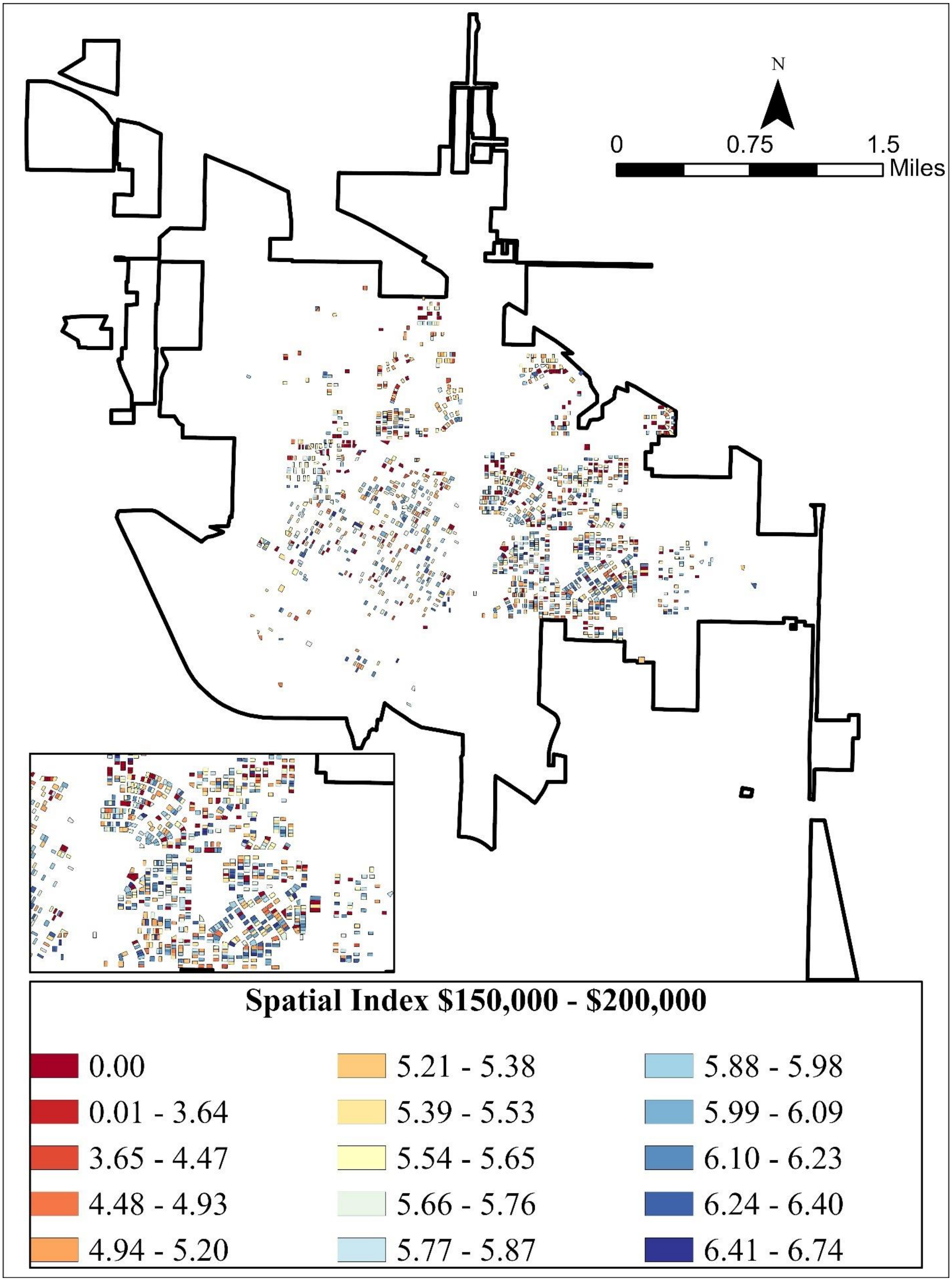


Figure 17: Spatial Index $150,000 - $200,000

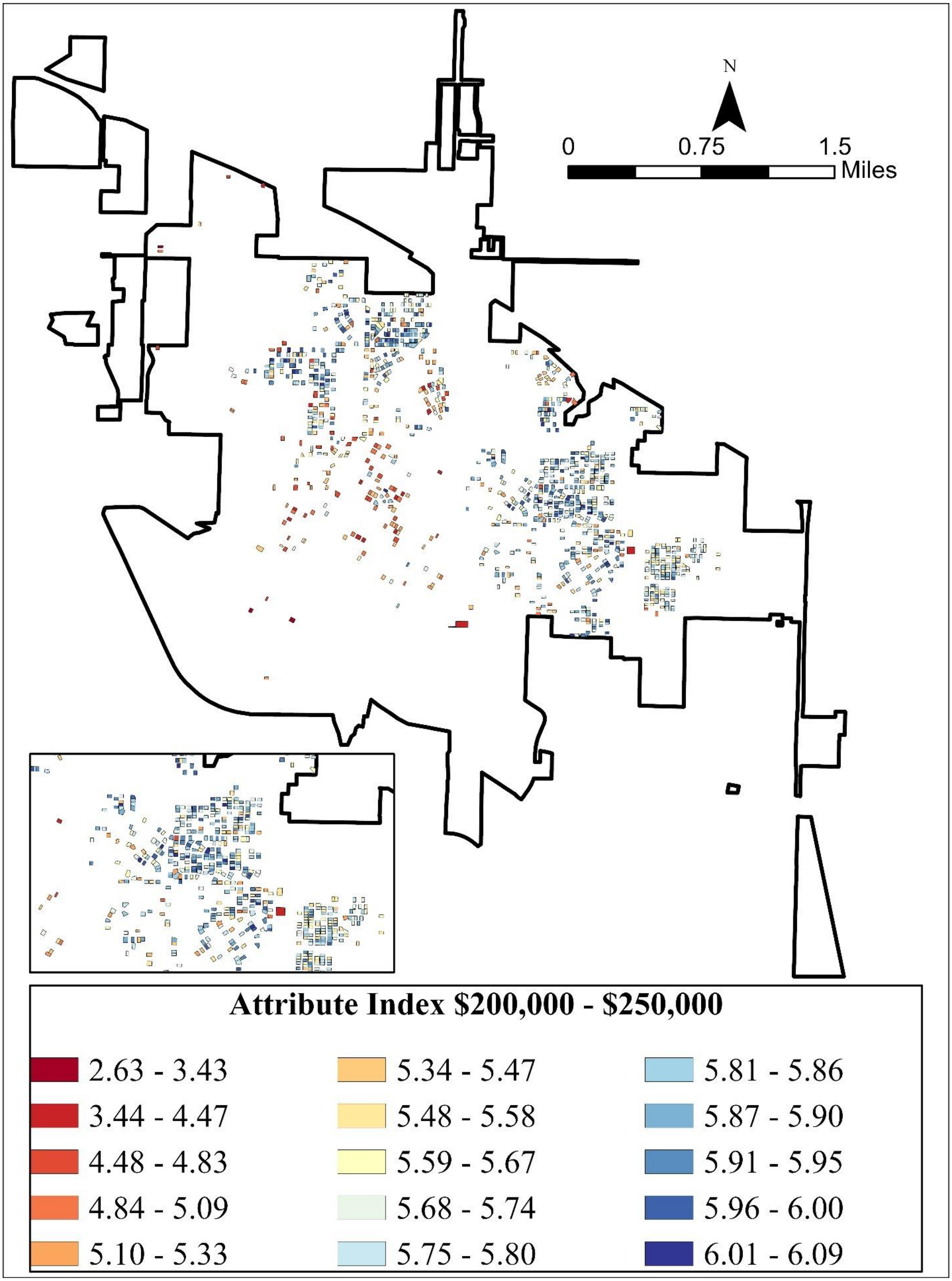


Figure 18: Attribute Index $200,000 - $250,000

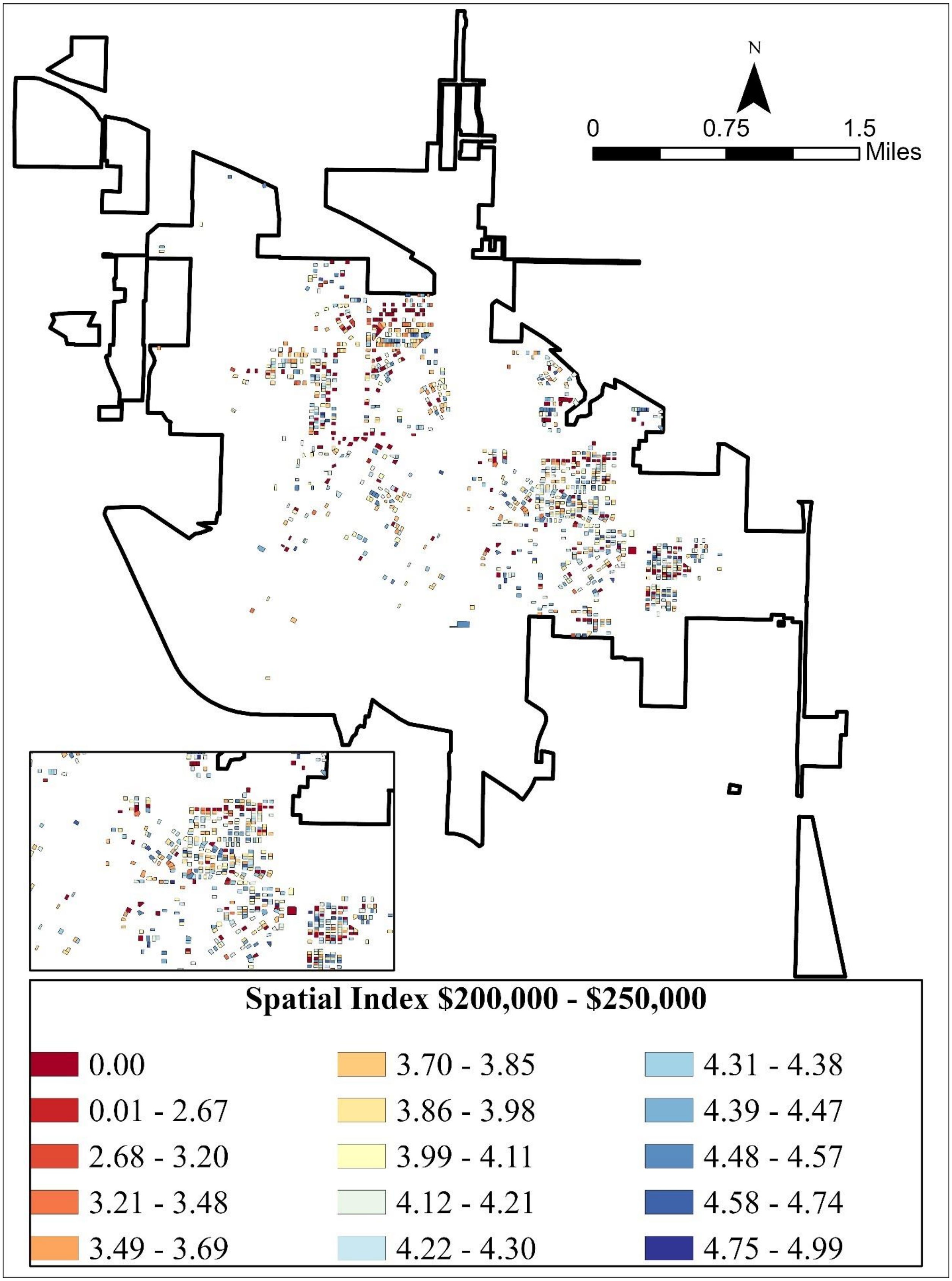


Figure 19: Spatial Index $200,000 - $250,000

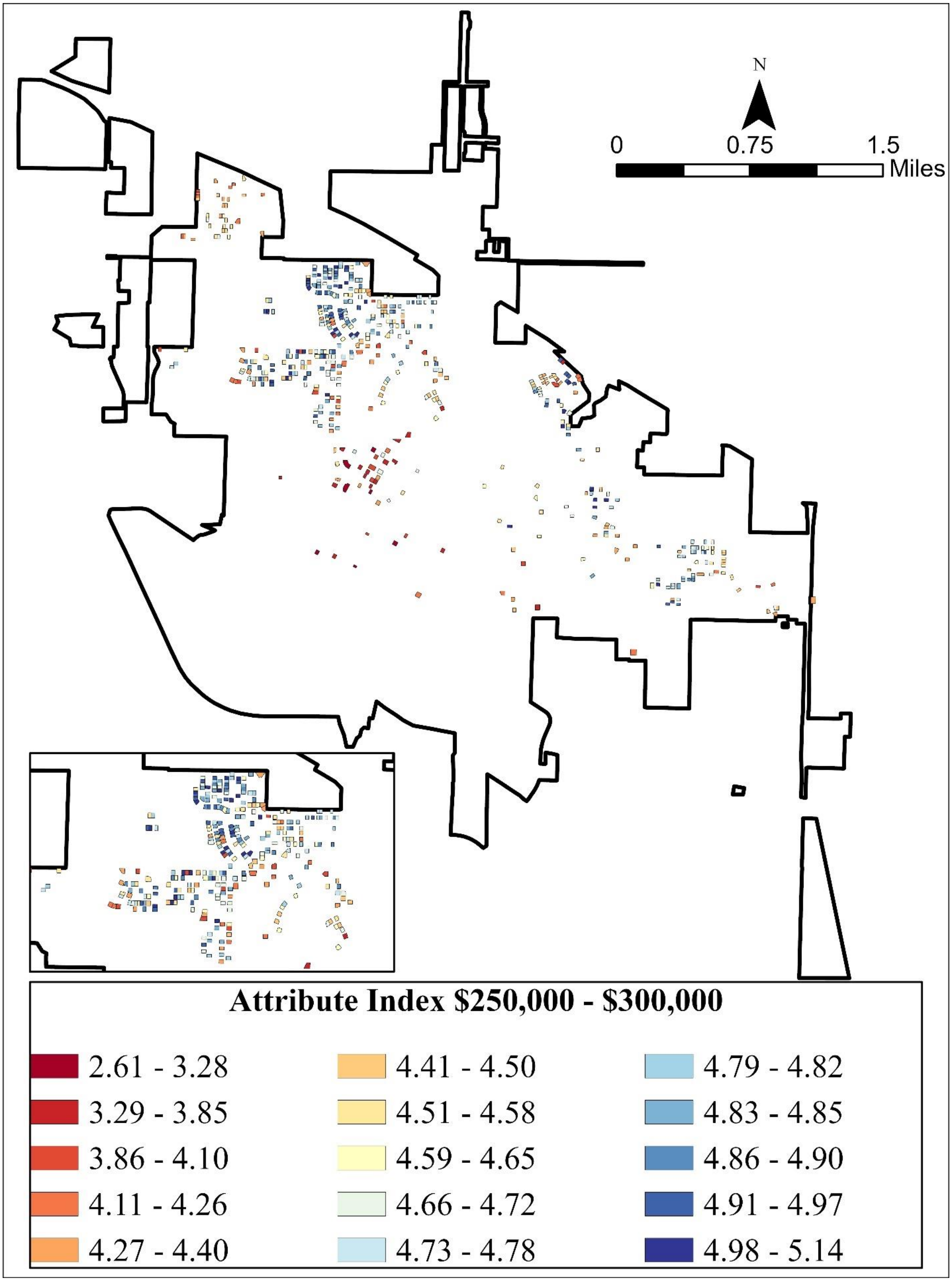


Figure 20: Attribute Index $250,000 - $300,000

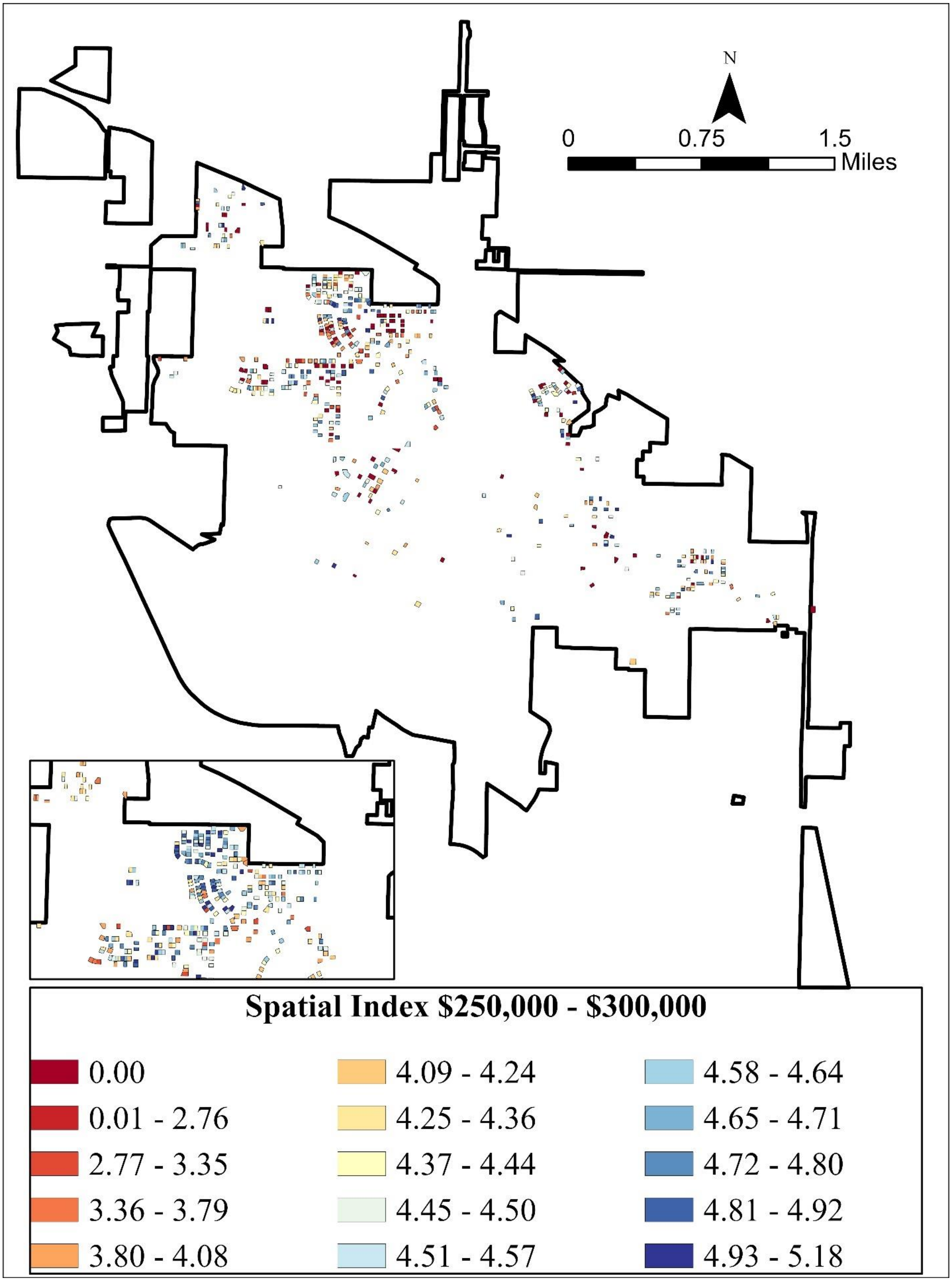


Figure 21: Spatial Index $250,000 - $300,000

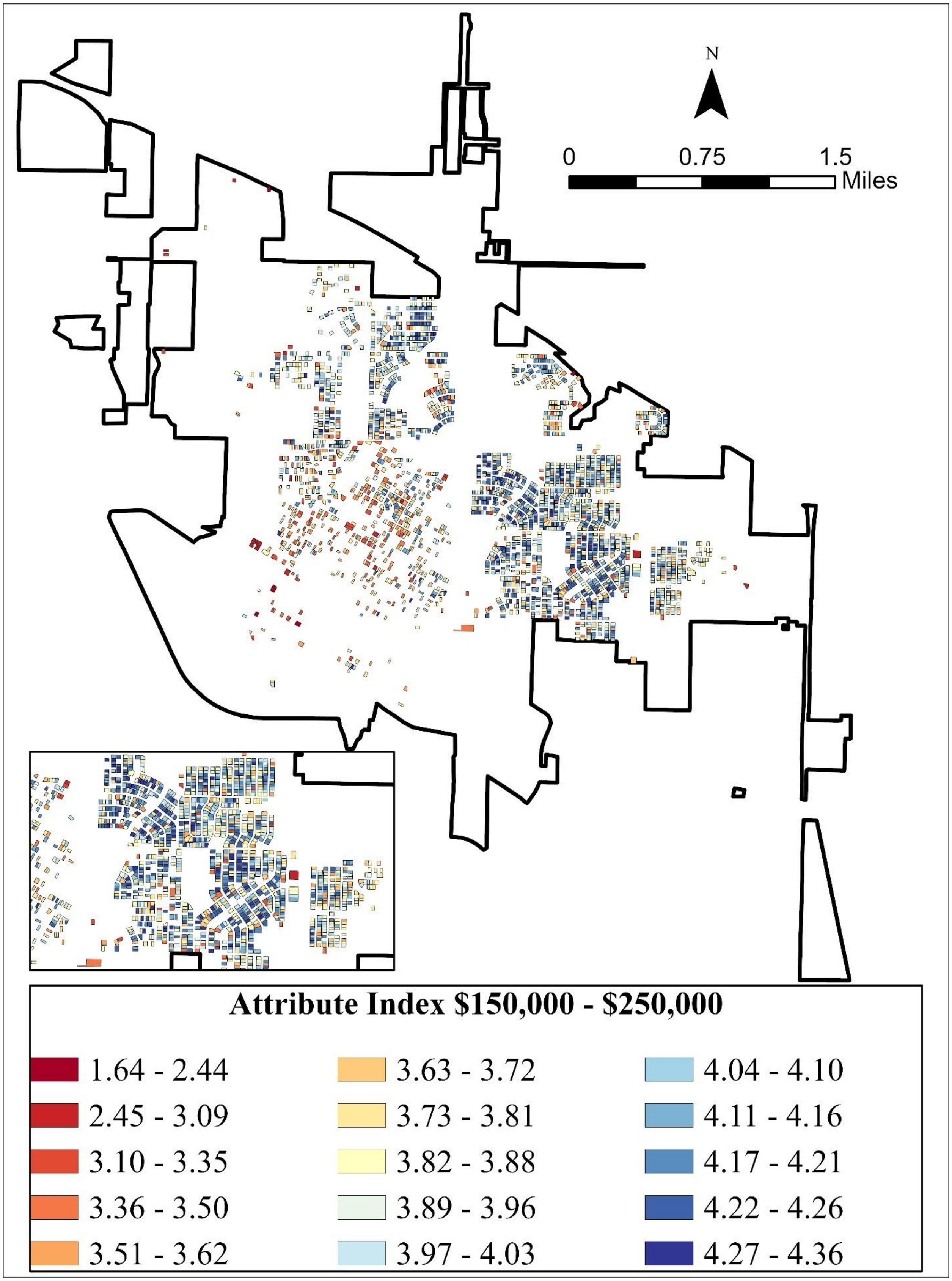


Figure 22: Attribute Index $150,000 - $250,000

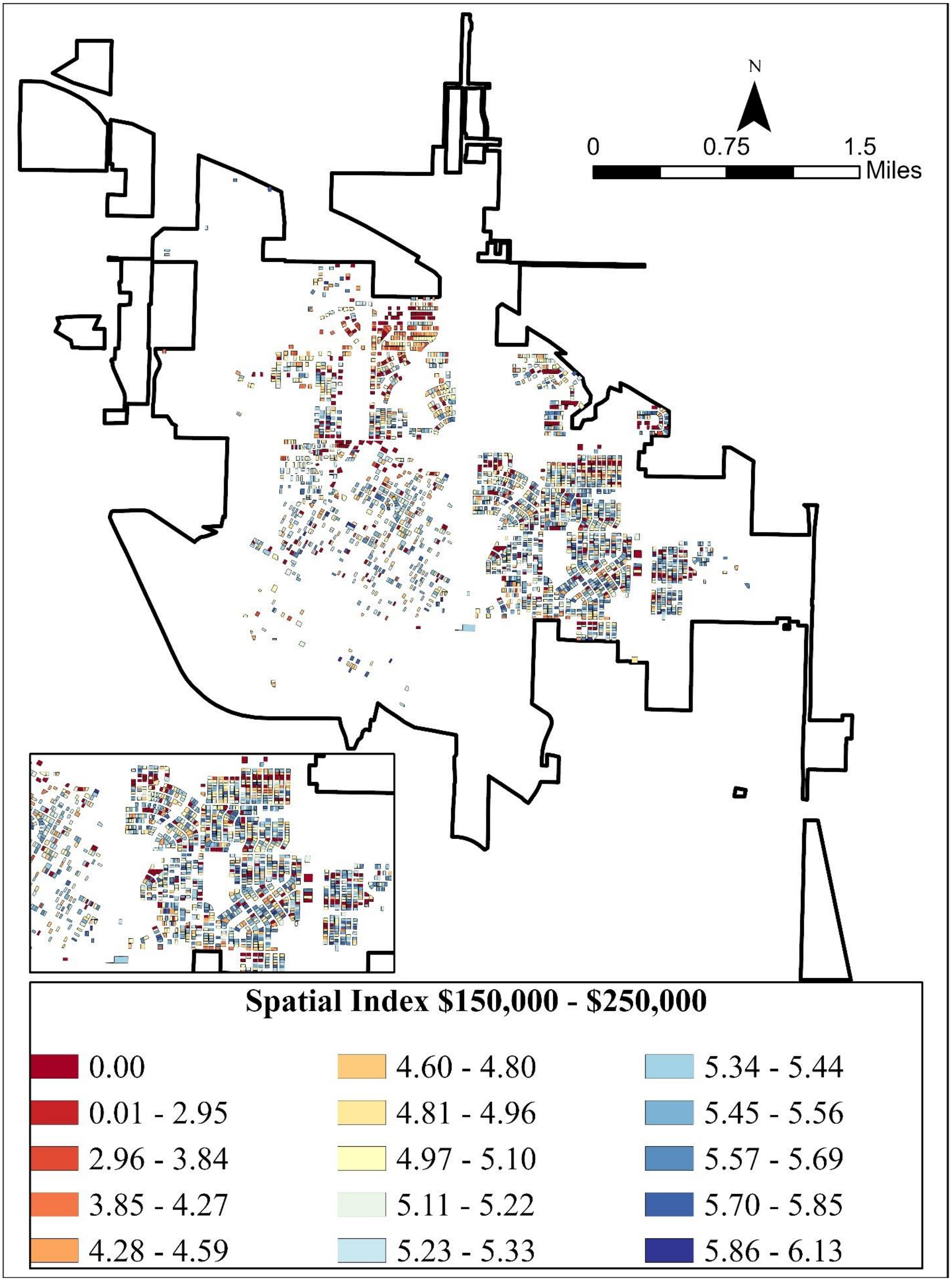


Figure 23: Spatial Index $150,000 - $250,000

Block Group Aggregate Results

The aggregate result maps presented in figures 24 – 35 represent two different metrics, the first being the mean index value for either the attribute or spatial values for units within that value range, and the second being the count of housing units within that particular value range. This way, if only one or two units, with relatively high index values exist within a block group, the quantity is considered when identifying characteristic neighborhoods. So, the block groups that contain both high mean values for either the spatial index value or attribute index value and a high count of units will represent a characteristic neighborhood for housing units within a particular data range. Figures 24 and 25 represent the result maps for the model that was developed utilizing the entire dataset and represent as clustering of both mean values and total count to the east of center and north of the block group map. Figures 26 and 27 are the result maps for the model generated using the $100,000 to $150,000 dollar range and show many of the same characteristics of the initial parcel maps, that there is clustering in the southwest portion of the city both in terms of high mean attribute index values and count of housing units, with a less defined pattern for the spatial index value. Figures 28 and 29 are the aggregate result maps at the block group level for the model generated utilizing the $150,000 - $200,0000 dollar range. In figure 28 there is a distinctive pattern for both count and mean attribute index to the southeast quadrant of the study area, once again with little to no similar pattern upholding for the spatial attribute index score.

Figures 30 and 31 show the aggregate result maps for the value range $200,000 - $250,000 dollars. There is very little neighborhood clustering visible in these representations. The lack of neighborhood cluster can either be contributed to the inaccuracy of the model created utilizing this value range and could be also caused by the lack of clustering in the parcel map itself, that housing values within this range are relatively scattered across the study area to begin with. For the value range $250,000 to $300,00 dollars, figures 32 and 33 were created. With this value range having the fewest number of sample units, it would be assumed that there would be no pattern for either the attribute or spatial maps for this value range, but in figure 32 and 33 there is cluster to the northwest in the study area, and there is a pattern that would suggest clustering occurring in that neighborhood in terms of both attribute and spatial index scores. Figures 34 and 35 show the aggregate value ranges for the $150,000 - $250,000 dollar range. These maps represent a similar pattern to figures 22 and 23 for the parcel maps with a general clustering to the east and southeast. There is a relatively low amount of clustering in this model and that may be because of the large range of values brought into consideration, despite this model having the second-best coefficient of determination behind only the full dataset.

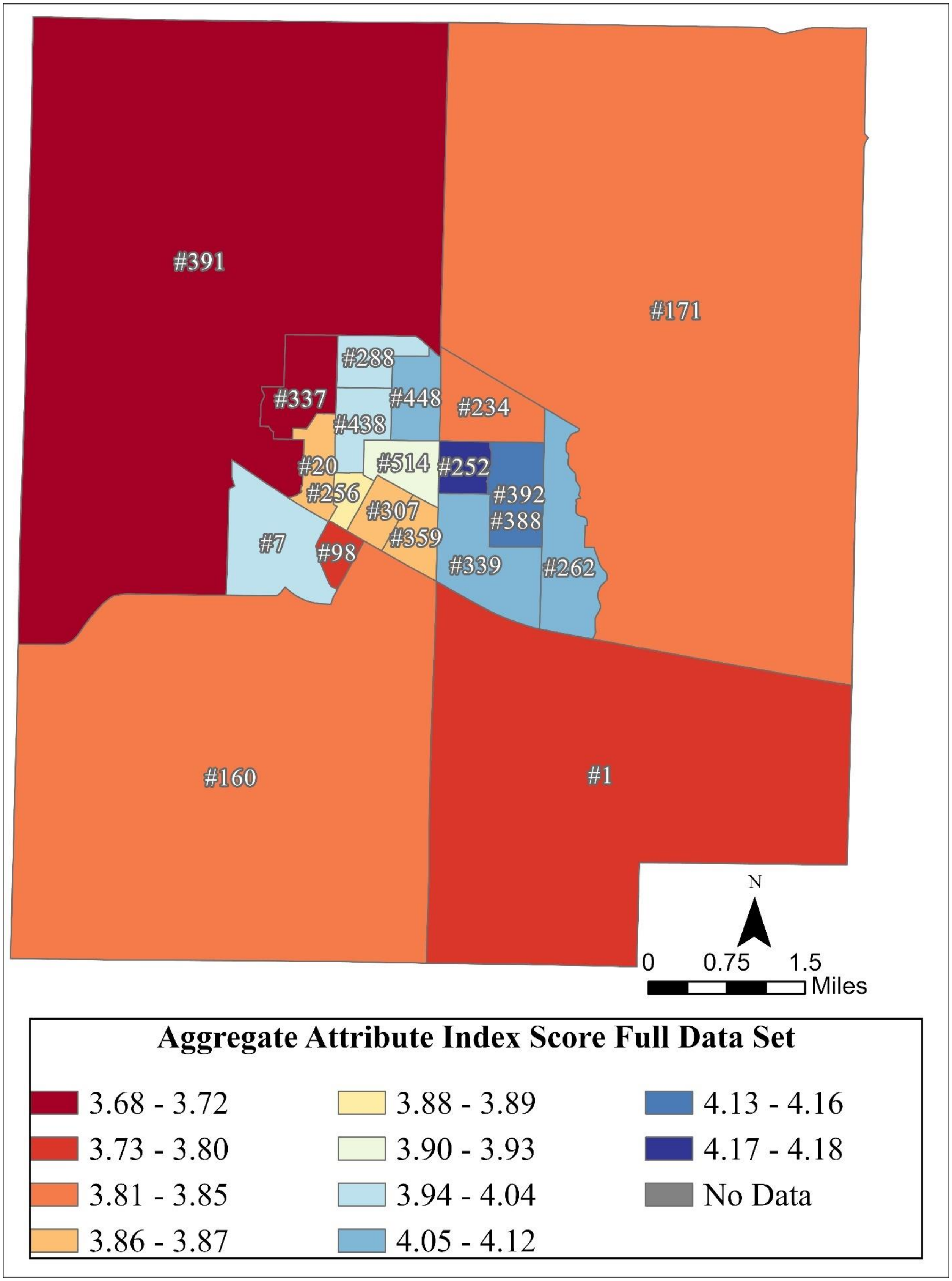


Figure 24: Aggregate Attribute Index Score Full Data Set

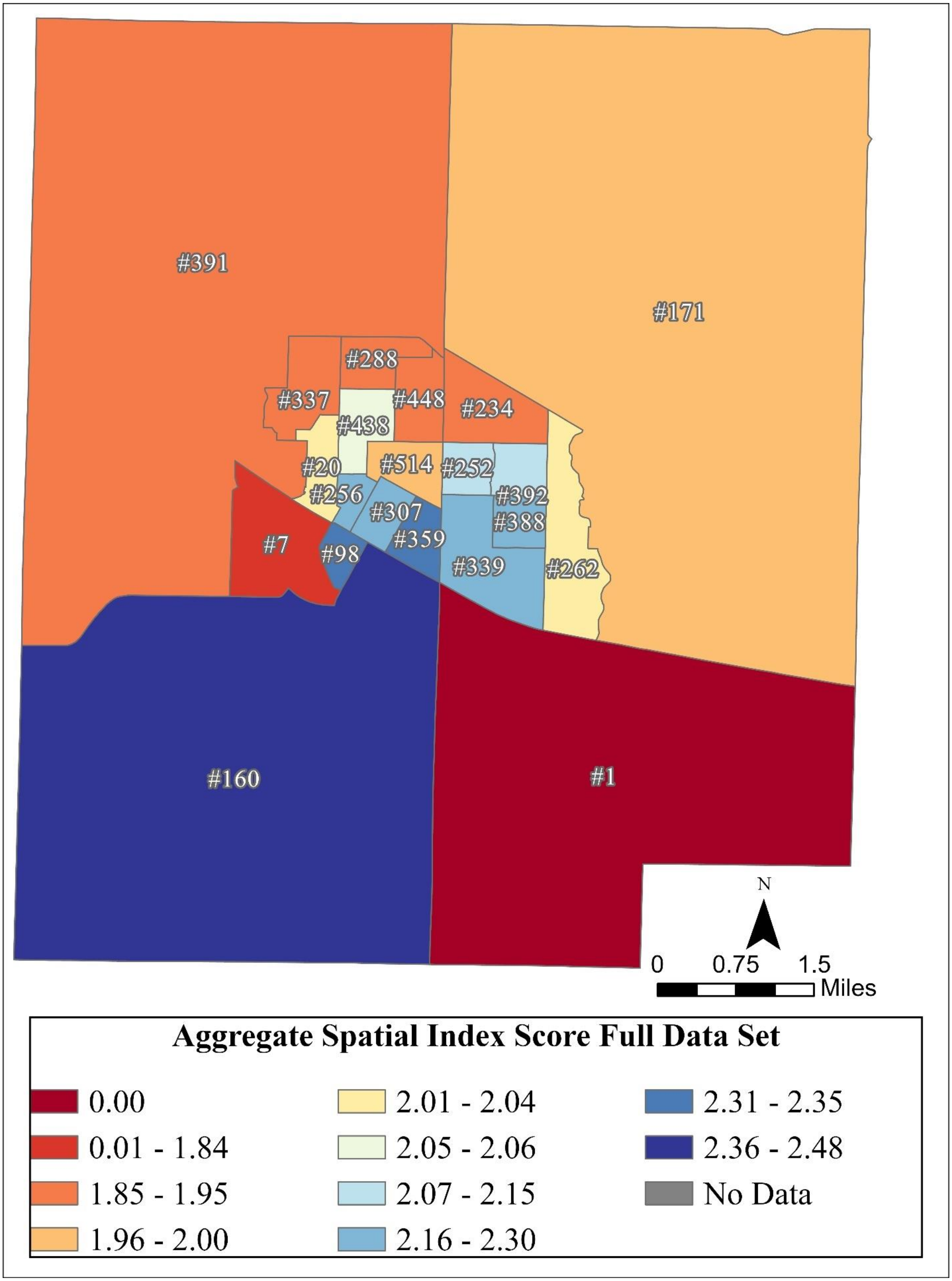


Figure 25: Aggregate Spatial Index Score Full Data Set

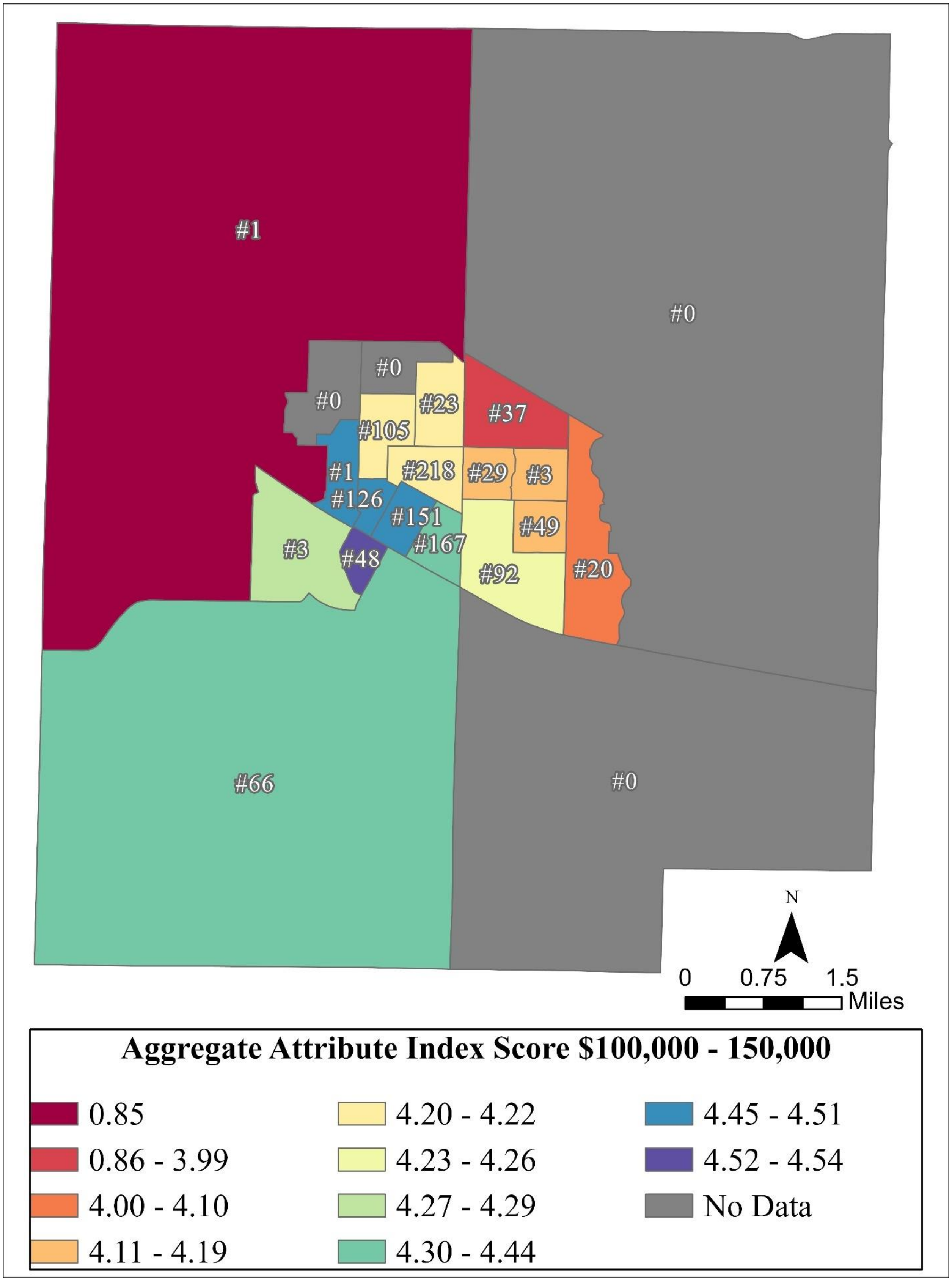


Figure 26: Aggregate Attribute Index Score $100,000 - $150,000

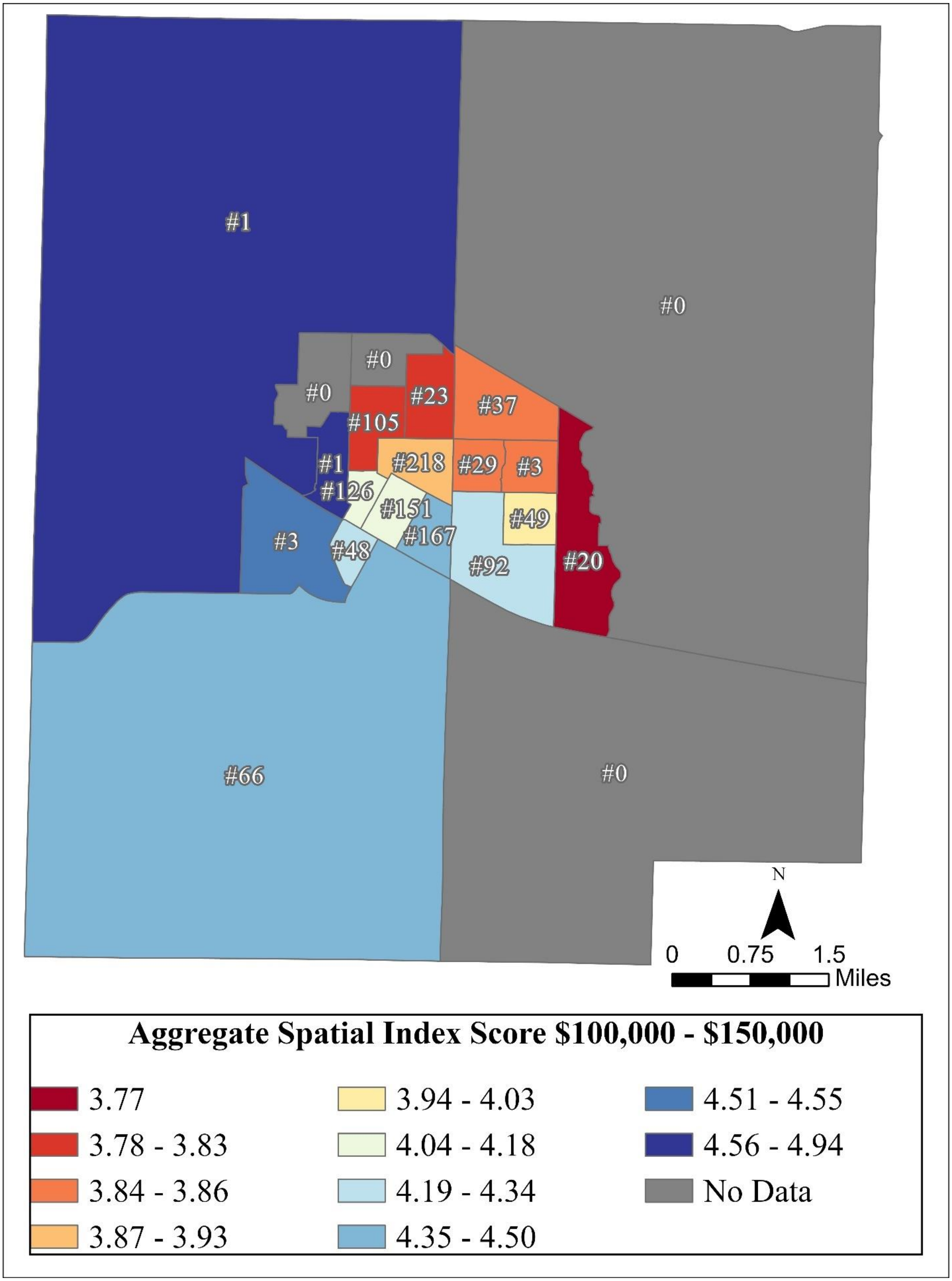


Figure 27: Aggregate Spatial Index Score $100,000 - $150,000

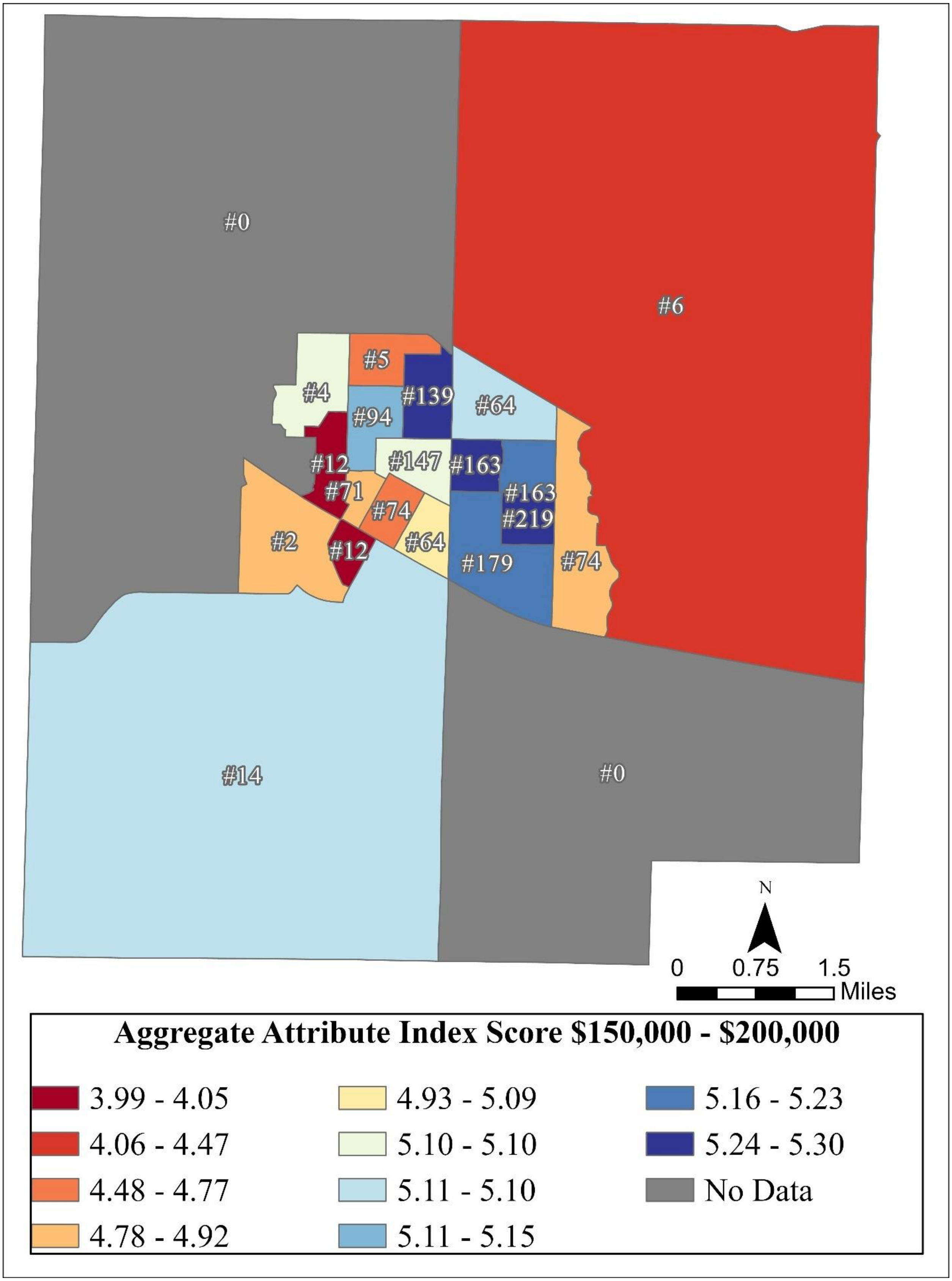


Figure 28: Aggregate Attribute Index Score $150,000 - $200,000

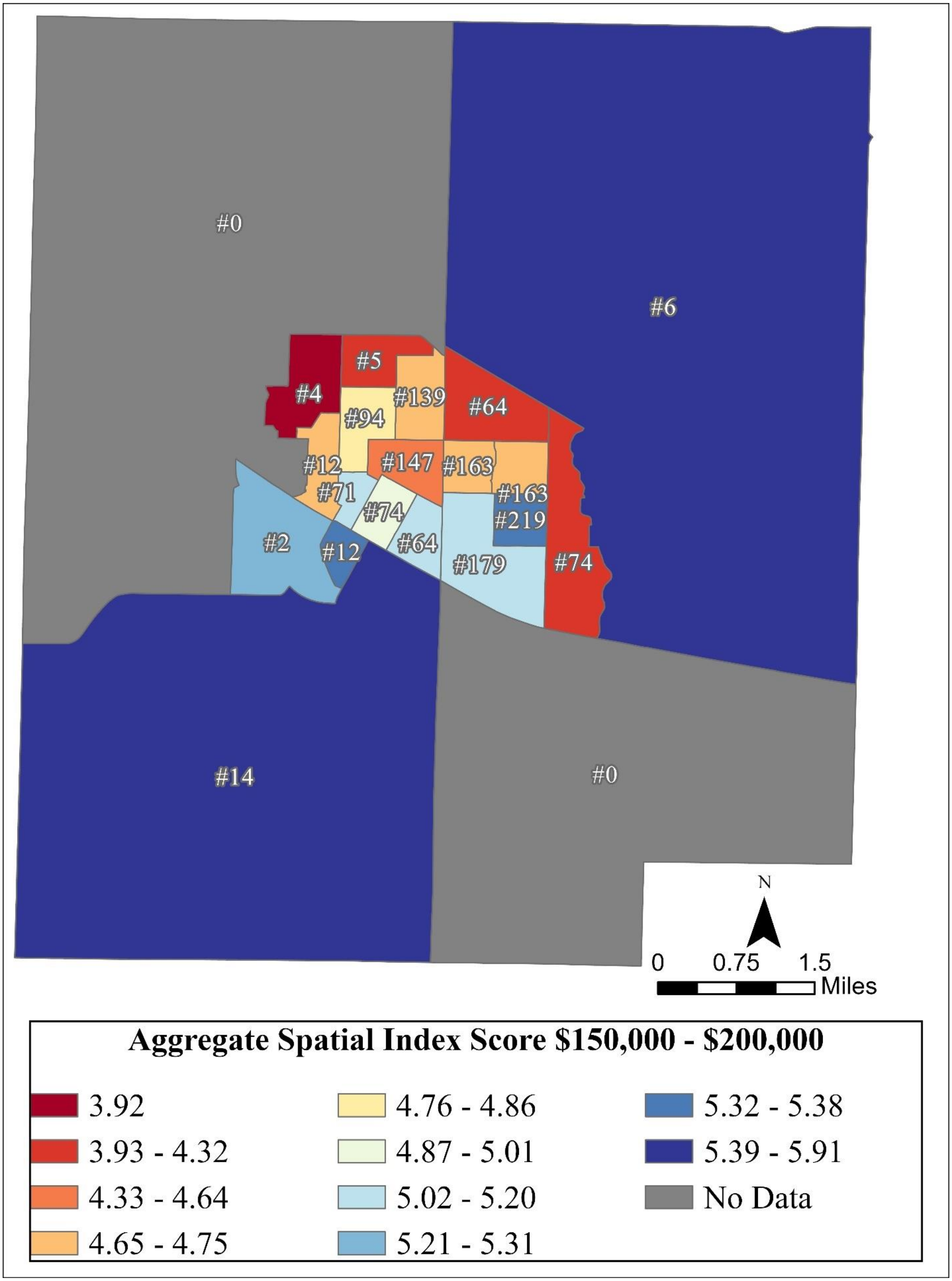


Figure 29: Aggregate Spatial Index Score $150,000 - $200,000

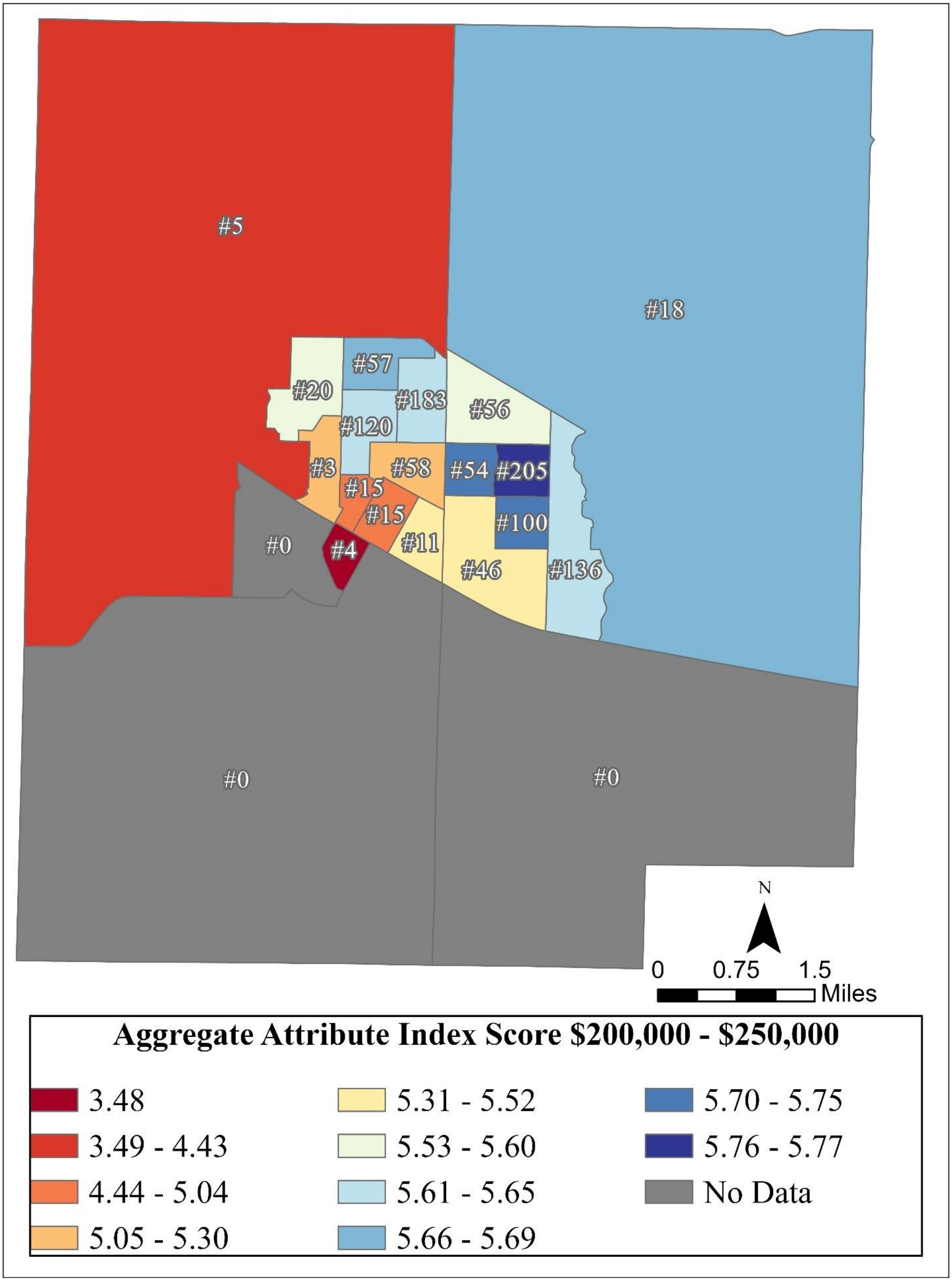


Figure 30: Aggregate Attribute Index Sore $200,000 - $250,000

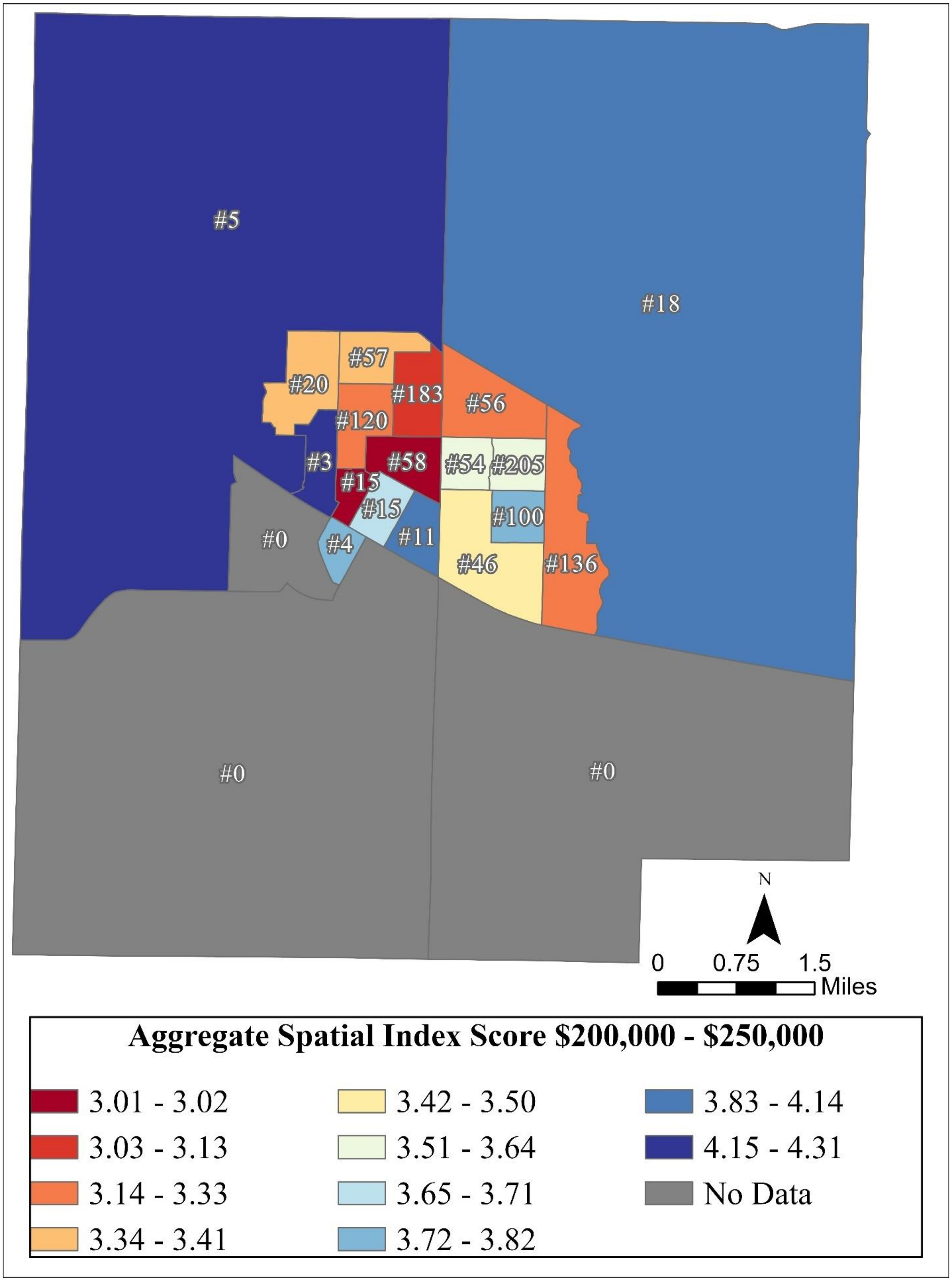


Figure 31: Aggregate Spatial Index Score $200,000 - $250,000

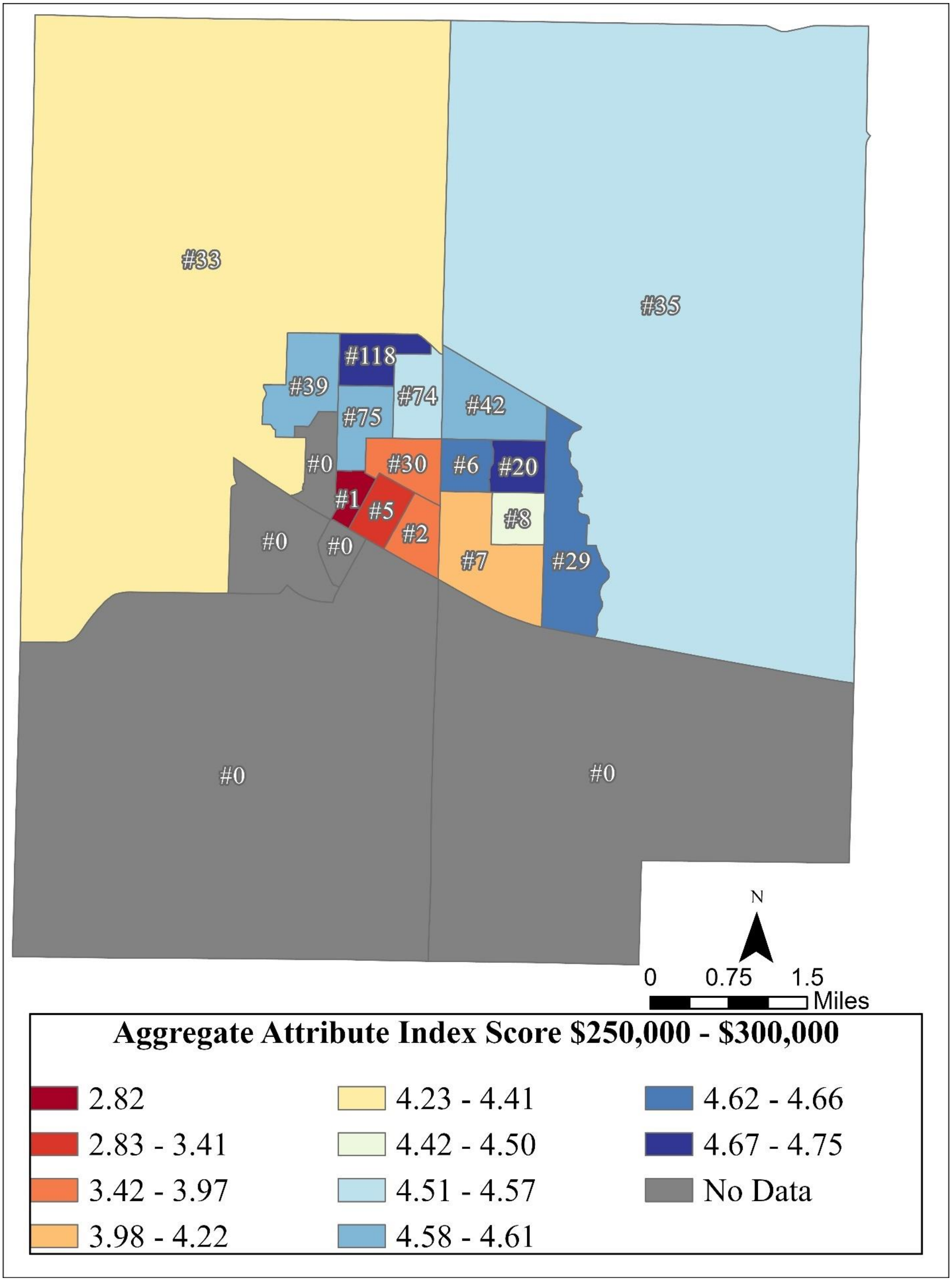


Figure 32: Aggregate Attribute Index Score $250,000 - $300,000

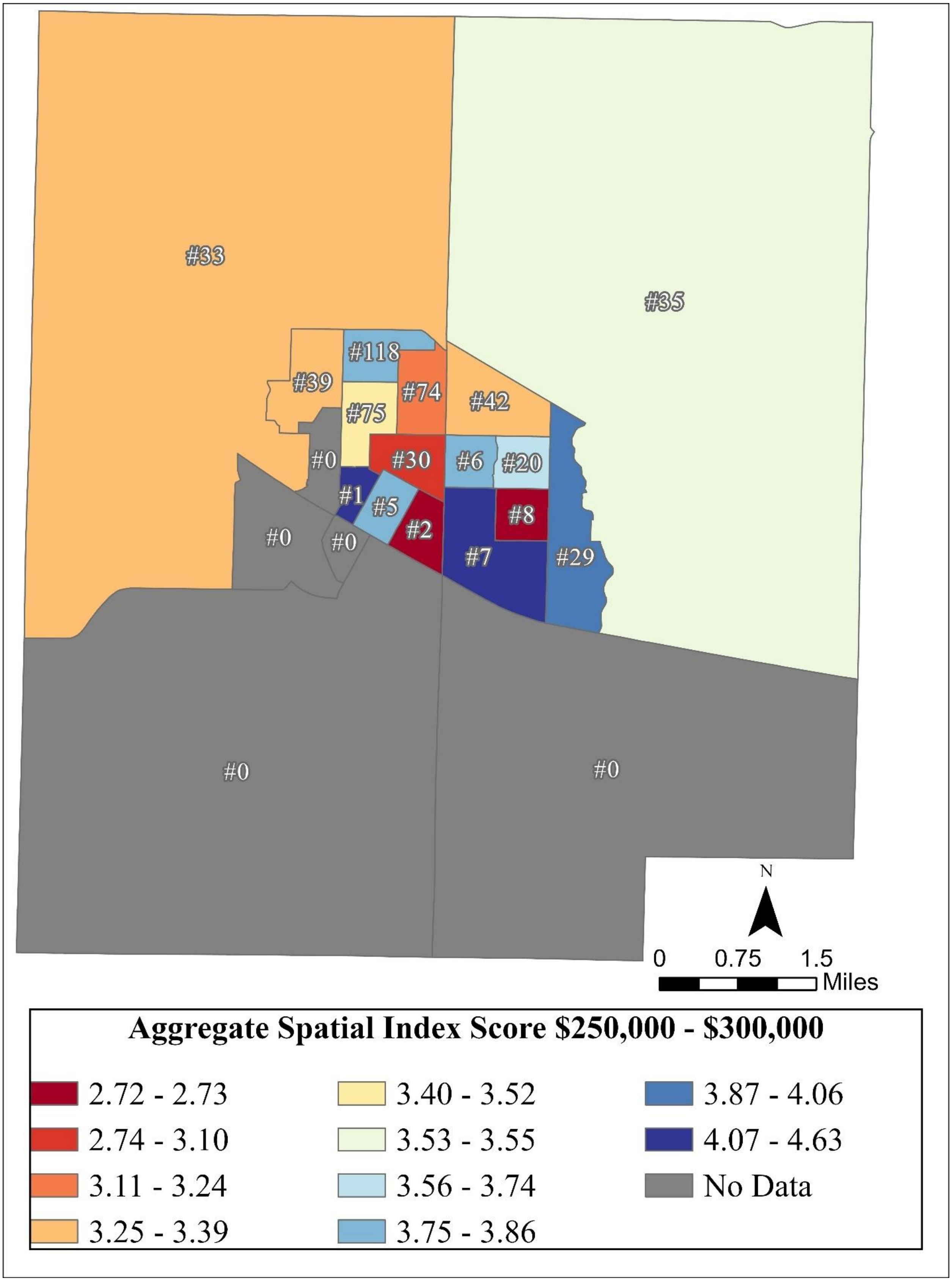


Figure 33: Aggregate Spatial Index Score $250,000 - $300,000

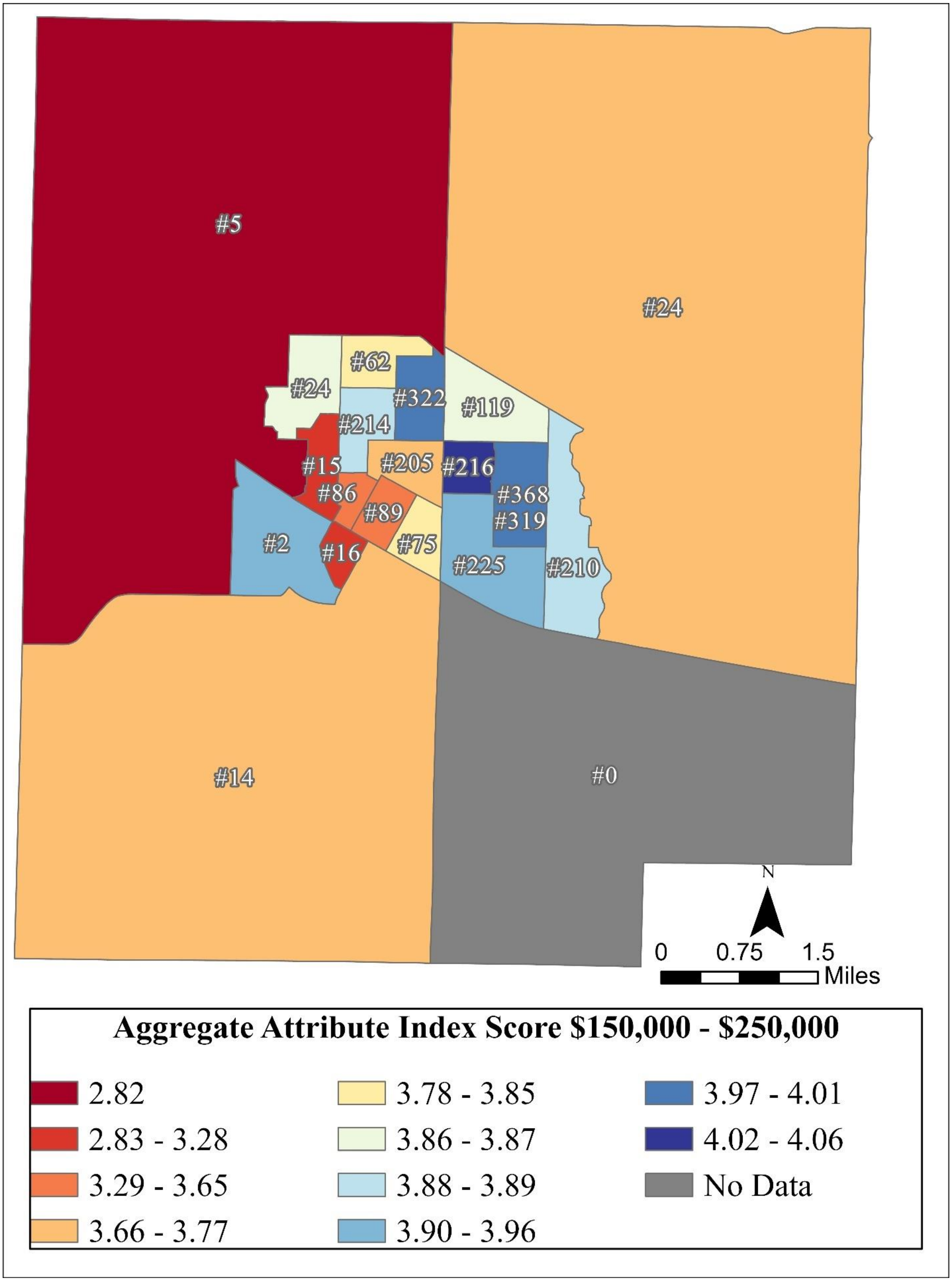


Figure 34: Aggregate Attribute Index Score $150,000 - $250,000

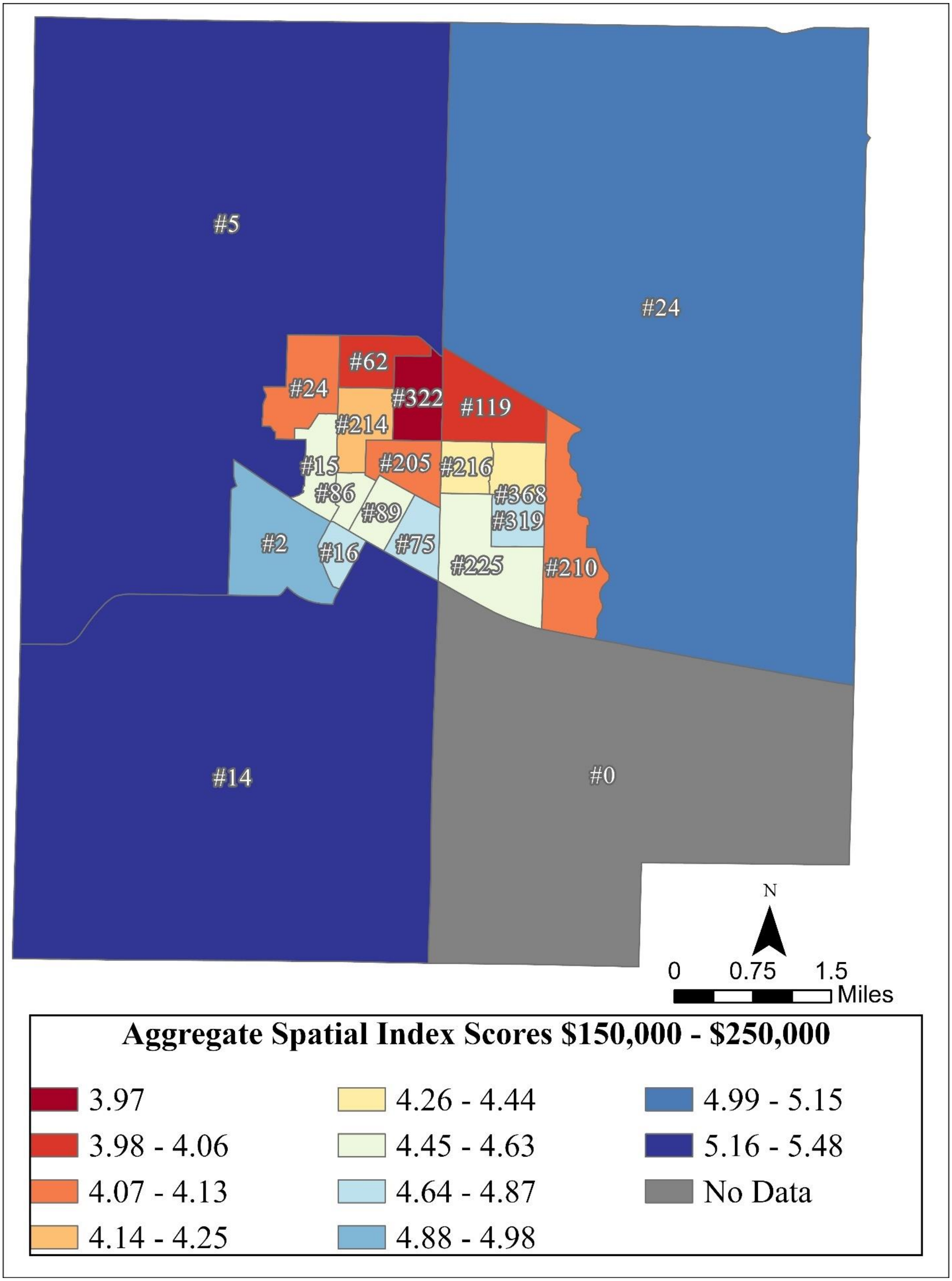


Figure 35: Aggregate Spatial Index Score $150,000 - $250,000

DISCUSSION/CONCLUSION

This research set out to analyze the viability of utilizing Variable Importance (VI) calculated as a by-product of Random Forrest Regression (RFR) modeling to classify characteristic neighborhoods based off of attribute and spatial data. This research was completed considering the current state and aspects of what is referred to as the American housing crisis and is an attempt at developing a GIS based method for analyzing the characteristics of existing housing supply to inform policy and decision makers on the attributes that have the greatest influence on price for a housing unit. With one of the main avenues for relieving the housing crisis being the protection and production of affordable housing units, it is more important than ever for municipalities and local interest groups to be informed on the attributes that have the greatest impact on housing price, and the distribution of characteristic neighborhoods within a city, both in terms of space and characteristics.

The key findings from this project is that there exists relationships between highly correlated variables in terms of the output of the RFR model, and that any individual interested in utilizing the HASI tool needs to be aware of the potential interactions between highly correlated variables before selecting input data for analysis, and that any user needs to determine the scope and resolution of variables that they would like to analyze and be aware of potential sources of autocorrelation to ensure the accuracy of the model. In terms of model accuracy and the coefficients of determination (R^2), the RFR model has a strong correlation between model performance and the number of housing units available for analysis. In order to optimize the utilization of this tool, this tool is better optimized for larger study areas, in which each $50,000 dollar range has at minimum 2,000 observations, as there seemed to be a large drop off in performance between 2,600 (value range $150,000 - $250,000) and 1,500 (value range $150,000

- $200,000). In terms of the spatial variables, the RFR model did not perform well in making predictions for the predicted value of a housing unit and would in term have little success in reducing the error associated in the model and produce uncharacteristic and inconsistent VI values for the spatial variables. For the spatial variables other statistical models may be implemented to greater effect.

The rank-ordering of the variable importance for housing attributes, each model does produce a different ranking for the attributes analyzed. This implies that different attributes had effects on the RFR model at different scales and supports the initial hypothesis of this research, that different housing unit value ranges will rank attributes differently in term of importance to the model. The result maps generated by analysis (figures 12 – 35) show that there is clustering in the different analysis ranges and that there is a balance to be had in terms of model performance and the clustering of characteristic neighborhoods. By optimizing model performance by opening up the value range, such as the $150,000 - $250,000 dollar range introduces more housing units for modeling and does increase model performance, but in turn reduce the ability to identify key characteristics neighborhoods by reducing the clustering of characteristic neighborhoods. In terms of housing attributes, selecting a range that has too few observations, such as the $250,000 - $300,000 dollar range, does improve clustering as can be seen in figures 19 and 32.

By analyzing the results of this research there are apparent limitations to this study, the main one being that the RFR models need large amounts of data within each value range to increase accuracy. This can be accomplished by utilizing data from larger cities that have more observations per value range than were available within the city of Hays. Also prevalent in the results are the low coefficient of determination scores for spatial variables. With the inputs in the

RFR model being the distance from every residential parcel within value range to the nearest spatial variable it is apparent that other, more tried and true methodologies for spatial correlation need to be utilized, such as the Global Moran’s I spatial autocorrelation analysis or other methods.

In summary, the HASI tool does show potential in identifying the physical characteristics that give housing units their price by utilizing the variable importance score generated through Random Forest Regression models, and an implementation of this type of variable analysis can be visualized in ArcGIS Pro. While the housing market is a highly fluid and complex system of push-pull factors in terms of value, the first step in addressing many of the housing issues facing America at the time of writing is the identification of currently affordable housing neighborhoods and the development of more affordable housing units. This research takes its place in the literature as a method for identifying the neighborhoods that characterize a particular value range and can provide GIS analysist and city planners with visual, tangible, and visitable locations for examples when setting policy and approving development projects.

REFERENCES

Appraiser. *Appraiser | Ellis County, KS - Official Website*. https://www.ellisco.net/91/Appraiser (last accessed 19 April 2023).

ArcGIS pro python reference. *ArcGIS Pro Python reference-ArcGIS Pro | Documentation*. https://pro.arcgis.com/en/pro-app/latest/arcpy/main/arcgis-pro-arcpy-reference.htm (last accessed 30 March 2023).

Aziz, A., M. M. Anwar, and M. Dawood. 2020. The impact of neighborhood services on land values: An estimation through the hedonic pricing model. *GeoJournal* 86 (4):1915–1925. doi:10.1007/s10708-019-10127-w

Belniak, S., and D. Wieczorek. 2017. Property valuation using hedonic price method – procedure and its application. *Czasopismo Techniczne* 6. doi:10.4467/2353737XCT.17.087.6563

Berk, R. A. 2010. *Statistical learning form a regression perspective*. New York, NY: Springer-Verlag.

Breiman, L. 2001. *Machine Learning* 45 (1):5–32.

Chen, S., D. Zhuang, and H. Zhang. 2020. GIS-based spatial autocorrelation analysis of housing prices oriented towards a view of spatiotemporal homogeneity and nonstationarity: A case study of guangzhou, China. *Complexity* 2020:1–16. doi:10.1155/2020/1079024

Chowdhury, S., Y. Lin, B. Liaw, and L. Kerby. 2022. Evaluation of tree based regression over multiple linear regression for non-normally distributed data in Battery Performance. *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. doi:10.1109/IDSTA55301.2022.9923169

Chung, Y., D. Seo, and J. Kim. 2018. Price determinants and GIS analysis of the housing market in Vietnam: The cases of Ho Chi Minh City and Hanoi. *Sustainability* 10 (12):4720. doi:10.3390/su10124720

Ellis County, KS - official website: Official Website. *Ellis County, KS - Official Website | Official Website*. https://www.ellisco.net/ (last accessed 19 April 2023).

Follain, J. R., and E. Jimenez. 1985. Estimating the demand for housing characteristics: A survey and critique. *Regional Science and Urban Economics* 15 (1):77–107. doi:10.1016/0166-0462(85)90033-X

Forest-based classification and regression (spatial statistics). *Forest-based Classification and Regression (Spatial Statistics)-ArcGIS Pro | Documentation*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/forestbasedclassificationregression.htm (last accessed 30 March 2023).

Hong, J., H. Choi, and W.-sung Kim. 2020. A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management* 24 (3):140–152. doi:10.3846/ijspm.2020.11544

How spatial autocorrelation (Global Moran's I) works. *How Spatial Autocorrelation (Global Moran's I) works-ArcGIS Pro | Documentation*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm (last accessed 30 March 2023).

Hu, L., S. He, Z. Han, H. Xiao, S. Su, M. Weng, and Z. Cai. 2019. Monitoring housing rental prices based on social media: an integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy* 82:657–673. doi:10.1016/j.landusepol.2018.10.030

Ivory, A., and K. W. Colton. Innovative Solutions for the housing crisis (SSIR). *Stanford Social Innovation Review: Informing and Inspiring Leaders of Social Change*. https://ssir.org/articles/entry/innovative_solutions_for_the_housing_crisis# (last accessed 30 March 2023).

Key facts - joint center for housing studies. https://www.jchs.harvard.edu/sites/default/files/interactive-item/files/Harvard_JCHS_State_Nations_Housing_2022_Key_Facts.pdf (last accessed 30 March 2023).

Kong, F., H. Yin, and N. Nakagoshi. 2007. Using GIS and landscape metrics in the hedonic price modeling of the amenity value of Urban Green Space: A case study in Jinan City, China. *Landscape and Urban Planning* 79 (3-4):240–252. doi:10.1016/j.landurbplan.2006.02.013

Ks292.cichosting.com. https://ks292.cichosting.com/webportal/appraiser/Default.aspx (last accessed 19 April 2023).

Lancaster, K. J. 1966. A new approach to consumer theory. *Journal of Political Economy* 74 (2):132–157. doi:10.1086/259131

Liao, W.-C., and X. Wang. 2012. Hedonic house prices and spatial quantile regression. *Journal of Housing Economics* 21 (1):16–27. doi:10.1016/j.jhe.2011.11.001

Lisi, G. 2019. Property valuation: The hedonic pricing model – location and housing submarkets. *Journal of Property Investment & Finance* 37 (6):589–596. doi:10.1108/JPIF-07-2019-0093

Mayer, M., S. C. Bourassa, M. Hoesli, and D. Scognamiglio. 2019. Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research* 12 (1):134–150. doi:10.1108/JERER-08-2018-0034

Mendonça, R., P. Roebeling, F. Martins, T. Fidélis, C. Teotónio, H. Alves, and J. Rocha. 2019. Assessing economic instruments to steer urban residential sprawl, using a hedonic pricing simulation modelling approach. *Land Use Policy* 92:104458. doi:10.1016/j.landusepol.2019.104458

Menze, B. H., B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of Spectral Data. *BMC Bioinformatics* 10 (1). doi:10.1186/1471-2105-10-213

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82 (1):34–55. doi:10.1086/260169

Seo, K., A. Golub, and M. Kuby. 2014. Combined impacts of highways and light rail transit on residential property values: A spatial hedonic price model for Phoenix, Arizona. *Journal of Transport Geography* 41:53–62. doi:10.1016/j.trangeo.2014.08.003

Shrestha, N. 2020. Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics* 8 (2):39–42. doi:10.12691/ajams-8-2-1.

Yao, J., and A. Stewart Fotheringham. 2015. Local spatiotemporal modeling of House prices: A mixed model approach. *The Professional Geographer* 68 (2):189–201. doi:10.1080/00330124.20151033671

Yoo, S., J. Im., and J. E. Wagner. 2012. Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning* 107 (3):293–306. doi:10.1016/j.landurbplan.2012.06.009

Čeh, M., M. Kilibarda, A. Lisec, and B. Bajat. 2018. Estimating the performance of Random Forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information* 7 (5):168. doi:10.3390/ijgi7050168

Łaszkiewicz, E., A. Heyman, X. Chen, Z. Cimburova, M. Nowell, and D. N. Barton. 2021. Valuing access to urban greenspace using non-linear distance decay in hedonic property pricing. *Ecosystem Services* 53:101394. doi:10.1016/j.ecoser.2021.101394

Łaszkiewicz, E., P. Czembrowski, and J. Kronenberg. 2019. Can proximity to urban green spaces be considered a luxury? classifying a non-tradable good with the use of hedonic pricing method. *Ecological Economics* 161:237–247. doi:10.1016/j.ecolecon.2019.03.025

APPENDIX

Appendix 1: Key Terms

Term	Page	Definition
Characteristic Neighborhood	2	Neighborhoods identified that demonstrate values closest to the mean for a particular value range with the normalized variable importance considered as a multiplier.
Bootstrapping	2	The process of data partitioning into decision trees based on the size of the input dataset.
Sum of Squared Residuals	2	The sum of the differences between an observation and the mean of observations squared. In this case, the amount of error in a sample of data on either side of a split within a decision tree.
Leaf Size	2	Minimum number of observations allowed in a node in a Random Forest Model
Gini Equation	3	The sum of the magnitude of splits caused by a variable in a RFR model
Coefficient of Determination	4	The quotient of the difference between the predicted value and observed value squared and the difference between the observed value and the means squared
Application Program Interface	4	A software package that exists between an application and a script that allows for the seamless transfer of data and user interface

Appendix 2: Data Inputs and Types

Housing Attributes		
Variable Abbreviation	Variable	Data Type
SQFTOT	Total Square Footage	Continuous
GARCAP	Garage Capacity	Continuous
SQFMF	Square Footage Main Floor	Continuous
BSMTA	Basement Area	Continuous
AGE	Age	Continuous
BLDSTY	Building Style	Nominal
FULLBAT	Full Bathrooms	Continuous
TROOM	Total Rooms	Continuous
SQFUP	Square Footage Upper Floor	Continuous
BSMTSTY	Basement Style	Nominal
DKA	Deck Area	Continuous
HALFBAT	Half Bathroom	Continuous
BROOM	Bedrooms	Continuous
FONSTY	Foundation Style	Nominal
AC	Air Conditioning	Nominal
UPF	Upper Floor	Continuous
DK	Deck	Nominal

Spatial Variables		
Abbreviation	Variable	Data Type
MLU	Multiple Living Unit	Nearest Distance
NFP	Not For Profit	Nearest Distance
COM	Commercial	Nearest Distance
K-12	K-12 Schools	Nearest Distance
UTY	Utility Infrastructure	Nearest Distance
VAC	Vacant Lots	Nearest Distance
FHSU	Fort Hays State University	Nearest Distance
AGG	Aggricultural Plots	Nearest Distance
STERN	Sternberg Musume	Nearest Distance
POS	Parks and Open Spaces	Nearest Distance

Appendix 3: Raw Variable Importance Table – Housing Attributes

Sample	Data Set	100 – 150	150 – 200	200 – 250	250-300	150 - 250
Attribute	Variable Importance					
Age	6423888906457	25392431884	48772143261	30691233888	17939162495	225776461656
SQFTOT	17651011721017	27823147791	35274599664	29466180961	15247527778	185555335176
SQFMF	11745935717144	40701606083	36294354076	23026785762	10153198679	186132132704
SQFUp	639754957866	2628916513	1801634013	1347050868	1324914845	10316796857
UpF	75272937346	398821720.9	270531489.7	295565225.7	182840464.2	1664054571
BldSty	2382441690146	31670468629	8773484043	4757758454	2803964646	38698488273
Troom	1219230737923	12815700539	13307234715	9043544346	4235669143	65070779587
Broom	227293203861	7161970971	7923625652	4927176616	3293390824	25510353858
FullBat	1327553850436	6292609397	10832097614	14286119765	3927680743	133837495983
HalfBat	252298996634	1697900877	3444532246	3185176223	1335306777	15674024080
GarCap	12916803251025	8271768687	17366290773	16242462312	2086810992	391064908134
FonSty	183880745055	5761302356	8068369824	4874550326	3900339767	22061576622
BsmtSty	309951779941	4351176673	11017624020	2289642562	2220184180	32108637831
BsmtA	9644093946740	17831675276	48263222470	25632727828	12551480567	312348487792

Appendix 4: Raw Variable Importance Table – Spatial Variables

Sample	Data Set	100 – 150	150 – 200	200 – 250	250-300	150 - 250
Attribute	Variable Importance					
MLUDt	6474944134436	20222938069	26141292954	16899258360	9040151238	165532290747
NFPDt	5955392606423	20982430711	25360501014	20311284786	8171780085	152119547810
ComDt	5611860899538	19828145405	25955249803	16847163361	7357446767	158861438419
K_12Dt	5250728919816	18515903561	25126810334	13931003817	7148100478	165747504527
UTYDt	4950804253378	22952636103	25624413120	20908868515	7756015247	164537406917
VacDt	4866688709299	21785026497	23811013407	15720995734	8330386964	154569252379
FHSUDt	4860721000529	13911261736	17441256806	14468747040	6997741632	115034668431
AggDt	4706889262682	17751395828	21996230269	15289276768	6163808578	125738963789
SternDt	4660985975614	13634701305	15725572741	10695163344	5059792287	101499150253
POSDt	4593657552680	18536051007	24597589599	16653903805	8443963632	142453214993

Appendix 5: HASI Script

```
#William A. Wallace
#3/31/2023
#Housing Attribute and Spatial Index Tool
#RF V0.04

##General Notes##
#This script is designed to automate the visualization of hedonic price modeling completed utilizing Random Forest Regression Analysis in ArcGIS Pro
   #infc - Input Feature Class Containing All of the Data Used for the Random Forest Regression
   #vmin - the minimum value under consideration for the analysis
   #vmax - The maximum value under consideration for the analysis
   #spatialfc - The Independent Feature Classes Brought in For analysis that represent points, lines and polygons of relevant spatial features
   #outfc - output feature class that will be the resultant of this script containing a visualization of the results of random regression analysis
   #depvar - the dependent variable that is to be used to answer the question typical for this analysis Total Appraised Value of a Property
   #indepvarcat - the categorical independent variables utilized for analysis
   #indepvarnum - the numerical independent variables utilized for analysis

#Import Necessary Libraries
import arcpy
from arcpy import env
import pandas as pd
import numpy as np
import os

#Establish a Scratch Workspace for Analysis
aprx = arcpy.mp.ArcGISProject("CURRENT")
gb = arcpy.env.scratchGDB
arcpy.AddMessage(arcpy.env.scratchGDB)

#Asking for Input Parameters from User
gdout = arcpy.GetParameterAsText(0)
infc = arcpy.GetParameterAsText(1)
vmin = arcpy.GetParameterAsText(2)
vmax = arcpy.GetParameterAsText(3)
depvar = arcpy.GetParameterAsText(4)
spatialfc = arcpy.GetParameterAsText(5)
indepvarcat = arcpy.GetParameterAsText(6)
indepvarnum = arcpy.GetParameterAsText(7)

arcpy.AddMessage('Visualizing Random Forest Regression Analysis Output for Properties within the Range of: $' + vmin + ' - $' + vmax)
```

```
#If the input data is a polygon, convert to centroid point, if already point, copy to working folder
infc_c = os.path.join(gdout + "\\" + "working")
desc = arcpy.Describe(infc)
if desc.shapeType == "Polygon":
    arcpy.FeatureToPoint_management(infc, infc_c)
else:
    arcpy.FeatureClassToFeatureClass_management(infc, infc_c)

#Giving a message to the user about the selected dependent and independent variables
arcpy.AddMessage('Dependent Variable: ' + depvar)
arcpy.AddMessage('Categorical Independent Variables: ' + indepvarcat)
arcpy.AddMessage('Numerical Independent Variables: ' + indepvarnum)
arcpy.AddMessage('Selected Spatial Variables: ' + spatialfc)

#Count number of null values in the dependent variable spot
count = 0
with arcpy.da.SearchCursor(infc_c, depvar) as cursor:
    for row in cursor:
        if row[0] == 0:
            count += 1
count = str(count)
arcpy.AddMessage('Number of entries deleted in the copy shapefile because of Null Value : ' + count)

#Delete the rows within the copied featureclass that contain a null value in the dependent variable
column
with arcpy.da.UpdateCursor(infc_c, depvar) as cursor:
    for row in cursor:
        if row[0] == 0:
            cursor.deleteRow()

#Calculate Latitude/Longitude Coordinates of the points for the spatial analysis
if desc.shapeType == "Polygon":
    arcpy.AddXY_management(infc_c)
else:
    pass

#Clean unneccessary z and m fields
arcpy.DeleteField_management(infc_c, ["POINT_Z", "POINT_M"])

#Begin working with spatial data by first seperating the multi point input into different string names
spatialfc = spatialfc.split(";")

#Run the near analysis in order to calculate the distance betweeen each residential parcel and the
different spatial variables
for fc in spatialfc:
    arcpy.GenerateNearTable_analysis(infc_c, fc, fc + 'DT')
```

```
#Generate a list that has name of the tables produced in the previous step
appendstr = "Dt"
distTables = [sub + appendstr for sub in spatialfc]

#Alter the names of the columns in dist table
for table in distTables:
   arcpy.AlterField_management(table, "NEAR_DIST", table, table)

#Use a permenant Join Field to append distance from locations to working table
for table in distTables:
   arcpy.JoinField_management(infc_c + ".shp", "ORIG_FID", table, "IN_FID", table)

#Delete Distance Tables
for table in distTables:
   arcpy.Delete_management(table)

#Work with the list of categorical independent variables to establish a list of useful text strings
indepvarcat = indepvarcat.split(";")

#Work with the list of numerical independent variables to establish a list of useful text strings
indepvarnum = indepvarnum.split(";")

#Create a List of Categorical Independent Variables that can be read into the random forest tool
lencat = len(indepvarcat)
catU = [None] * lencat
lencat = lencat - 1
index  = 0
while index <= lencat:
   catU[index] = [indepvarcat[index], "true"]
   index = index + 1

#Create a List of Numeric Independent Variables that can be read into the random forest tool
lennum = len(indepvarnum)
numU = [None] * lennum
lennum = lennum - 1
index = 0
while index <= lennum:
   numU[index] = [indepvarnum[index], "false"]
   index = index + 1

#Combine the two Catergorical Independent Variable List and the Numeric Independent Variable List so
that each can be analyzed in the random forest Regression
indepvar = catU + numU

#Create an analysis table for the RFR Tool
vminint = int(vmin)
vmaxint = int(vmax)
with arcpy.da.UpdateCursor(infc_c, "TAV") as cur:
    for row in cur:
        if (row[0] >= vminint and row[0] <= vmaxint):
            pass
        else:
            cur.deleteRow()

#Random Forest for Housing Attribute Features
arcpy.AddMessage("Running Random Forest Regression on Housing Attribute Data")
prediction_type = "TRAIN"
in_features = infc_c
variable_predict = depvar
treat_variable_as_categorical = None
explanatory_variables = indepvar
distance_features = None
explanatory_rasters = None
features_to_predict = infc_c
output_features = "outputatt.shp"
output_raster = None
explanatory_variable_matching = indepvar
explanatory_distance_matching = spatialfc
explanatory_rasters_matching = None
output_trained_features_att = os.path.join(gdout + "\\" + "training_outputatt")
output_importance_table_att = os.path.join(gdout + "\\" + "output_tableATT")
use_raster_values = False
number_of_trees = 100
minimum_leaf_size = 5
maximum_level = None
sample_size = 100
random_sample = 10
percentage_for_training = 10
output_classification_table = None
output_validation_table_att = os.path.join(gdout + "\\" + "validation_table_att")
compensate_sparse_categories = "FALSE"
number_validation_runs = 10
calculate_uncertainty = "TRUE"

arcpy.stats.Forest(prediction_type, in_features, variable_predict,
    treat_variable_as_categorical, explanatory_variables, distance_features,
    explanatory_rasters, features_to_predict, output_features, output_raster,
    explanatory_variable_matching, explanatory_distance_matching,
    explanatory_rasters_matching, output_trained_features_att, output_importance_table_att,
    use_raster_values, number_of_trees, minimum_leaf_size, maximum_level,
    sample_size, random_sample, percentage_for_training, output_classification_table,
    output_validation_table_att, compensate_sparse_categories, number_validation_runs,
calculate_uncertainty)
arcpy.AddMessage("Hosusing Attribute Random Forest Regression Completed")

#Random Forest for Spatial Variables
```

```
arcpy.AddMessage("Running Random Forest Regression on Spatial Variables")
prediction_type = "TRAIN"
in_features = infc_c
variable_predict = depvar
treat_variable_as_categorical = None
explanatory_variables = distTables
distance_features = None
explanatory_rasters = None
features_to_predict = infc_c
output_features = "outputspat.shp"
output_raster = None
explanatory_variable_matching = distTables
explanatory_distance_matching = None
explanatory_rasters_matching = None
output_trained_features_spat = os.path.join(gdout + "\\" + "training_outputspatt")
output_importance_table_spat = os.path.join(gdout + "\\" + "output_tableSpat")
use_raster_values = False
number_of_trees = 100
minimum_leaf_size = 5
maximum_level = None
sample_size = 100
random_sample = 10
percentage_for_training = 10
output_classification_table = None
output_validation_table_spat = os.path.join(gdout + "\\" + "validation_table_spat")
compensate_sparse_categories = "FALSE"
number_validation_runs = 10
calculate_uncertainty = "TRUE"

arcpy.stats.Forest(prediction_type, in_features, variable_predict,
   treat_variable_as_categorical, explanatory_variables, distance_features,
   explanatory_rasters, features_to_predict, output_features, output_raster,
   explanatory_variable_matching, explanatory_distance_matching,
   explanatory_rasters_matching, output_trained_features_spat, output_importance_table_spat,
   use_raster_values, number_of_trees, minimum_leaf_size, maximum_level,
   sample_size, random_sample, percentage_for_training, output_classification_table,
   output_validation_table_spat, compensate_sparse_categories, number_validation_runs,
calculate_uncertainty)
arcpy.AddMessage("Random Forest Regression on Spatial Variables Completed")

#Extract Values in the Working Shapefile that fall within the value range into a dataframe
depvarstr = '"{}"'.format(depvar)
columns_data = [f.name for f in arcpy.ListFields(infc_c)]
df_range = pd.DataFrame(data=arcpy.da.SearchCursor(infc_c,columns_data,
      '{} >= {:s} And {} <= {:s}'.format(depvarstr,vmin,depvarstr,vmax)),
      columns = columns_data)

#Establish a Dataframe of Utilized Attribute Values within the range
```

```
list_HK = ["FID", depvar]
list_att = (indepvarnum + indepvarcat)
list_att1 = (list_HK + list_att)
df_range_att = df_range[list_att]

#Establish a Dataframe of utilized Spatial variables withinn the range
list_spat = (distTables)
list_spat1 = (list_HK + distTables)
df_range_spat = df_range[list_spat]

#Begin Working with Outputs
#Create a dataframe form the Attribute Output variable importance and validation tables
columns = [f.name for f in arcpy.ListFields(output_importance_table_att)]
df_att_VI = pd.DataFrame(data=arcpy.da.SearchCursor(output_importance_table_att, columns),
columns = columns)
columns = [f.name for f in arcpy.ListFields(output_validation_table_att)]
df_att_Val = pd.DataFrame(data=arcpy.da.SearchCursor(output_validation_table_att, columns),
columns = columns)

#Create a dataframe that contains the min/max normalized mean varaible improtance value for
attributs across all trials
df_att_VI_mean = df_att_VI.mean(axis = 0)
df_att_VI_mean = df_att_VI_mean.drop('OBJECTID')
df_att_VI_F = ((df_att_VI_mean - df_att_VI_mean.min())/(df_att_VI_mean.max() -
df_att_VI_mean.min()))

#Create a dataframe that contains the means for Attributes within range
df_range_att_mean = df_range_att.mean()

#Subtract the means from each individual entry in the range then take the absolute value
df_range_att_adj = df_range_att[list_att].subtract(df_range_att_mean, axis = 1)
df_range_att_adj = df_range_att_adj.abs()

#Normalize the absolute values of the difference between value and means for attribute variables
df_att_norm = ((df_range_att_adj - df_range_att_adj.min())/(df_range_att_adj.max()-
df_range_att_adj.min()))
df_att_norm_inv = (1 - df_att_norm)
columns = df_att_norm_inv.columns
df_att_norm_inv.columns = [x.upper() for x in columns]
df_att_fin = df_att_norm_inv.multiply(df_att_VI_F, axis = 1)

#Create a dataframe from the Spatial Output variable importance and validation tables
columns = [f.name for f in arcpy.ListFields(output_importance_table_spat)]
df_spat_VI = pd.DataFrame(data=arcpy.da.SearchCursor(output_importance_table_spat, columns),
columns = columns)
columns = [f.name for f in arcpy.ListFields(output_validation_table_spat)]
df_spat_Val = pd.DataFrame(data=arcpy.da.SearchCursor(output_validation_table_spat, columns),
columns = columns)
```

```
#Create a dataframe that contains the min/max normalized mean variable importance values for spatial
varialbes across all trials
df_spat_VI_mean = df_spat_VI.mean(axis = 0)
df_spat_VI_mean = df_spat_VI_mean.drop('OBJECTID')
df_spat_VI_F = ((df_spat_VI_mean - df_spat_VI_mean.min())/(df_spat_VI_mean.max() -
df_spat_VI_mean.min()))
arcpy.AddMessage(df_spat_VI_F)
#Create a dataframe that contains the means for spatial variables within range
df_range_spat_mean = df_range_spat.mean()

#Subtract the means from each individual entry in the range then take the absolute value
df_range_spat_adj = df_range_spat[list_spat].subtract(df_range_spat_mean, axis = 1)
df_range_spat_adj = df_range_spat_adj.abs()

#Normalize the absolute values of the differnce between value and mean for spatial variables
df_spat_norm = ((df_range_spat_adj - df_range_spat_adj.min())/(df_range_spat_adj.max()-
df_range_spat_adj.min()))
df_spat_norm_inv = (1 - df_spat_norm)
columns = df_spat_norm_inv.columns
df_spat_norm_inv.columns = [x.upper() for x in columns]
df_spat_fin = df_spat_norm_inv.multiply(df_spat_VI_F, axis = 1)

#Sum all of the Columns in the final attribute index
df_att_index = pd.DataFrame(columns = ["FID","AttIndex"])
df_att_index["FID"] = df_range["OBJECTID"]
df_att_index["AttIndex"] = df_att_fin.sum(axis = 1)

#Sum all of the Columns in the final spatial index
df_spat_index = pd.DataFrame(columns = ["FID","SpatIndex"])
df_spat_index["FID"] = df_range["OBJECTID"]
df_spat_index["SpatIndex"] = df_spat_fin.sum(axis = 1)

#Delete the vailadation tables, output attribution tables and the training output feature classes
#arcpy.Delete_management([output_importance_table_spat,output_trained_features_spat,output_vali
dation_table_spat,output_importance_table_att,output_trained_features_att,output_validation_table_
att])

#Create a copy of the Input Feature Class that will contain the output index features
arcpy.FeatureClassToFeatureClass_conversion(infc, gdout, "AttributeIndex.shp")
arcpy.FeatureClassToFeatureClass_conversion(infc, gdout, "SpatialIndex.shp")

#Create a copy of the OBJECTID Field in the Attribute and Spatial Index Field that is an integer
arcpy.AddField_management(os.path.join(gdout + "\\" + "AttributeIndex.shp"), "FID", "FLOAT")
arcpy.AddField_management(os.path.join(gdout + "\\" + "SpatialIndex.shp"), "FID", "FLOAT")
arcpy.CalculateField_management(os.path.join(gdout + "\\" + "AttributeIndex.shp"), "FID",
"!OBJECTID!", "PYTHON_9.3")
```

```
arcpy.CalculateField_management(os.path.join(gdout + "\\" + "SpatialIndex.shp"), "FID", "!OBJECTID!",
"Python_9.3")

#Convert Output Dataframes into CSV files in the scratch GDB
att_out = os.path.join(gb + "\\" + "att_out.csv")
df_att_index.to_csv(att_out)
spat_out = os.path.join(gb + "\\" + "spat_out.csv")
df_spat_index.to_csv(spat_out)
arcpy.TableToTable_conversion(att_out, gdout, "att_index")
arcpy.TableToTable_conversion(spat_out, gdout, "spat_index")

#Join by attribute the copies of the input features to the newly created csv files
att_index_loc = os.path.join(gdout + "\\" + "AttributeIndex.shp")
spat_index_loc = os.path.join(gdout + "\\" + "SpatialIndex.shp")
att_table_loc = os.path.join(gdout + "\\" + "att_index")
spat_table_loc = os.path.join(gdout + "\\" + "spat_index")
arcpy.JoinField_management(att_index_loc, "FID", att_table_loc, "FID", "AttIndex")
arcpy.JoinField_management(spat_index_loc, "FID", spat_table_loc, "FID", "SpatIndex")

#Delete the temporarily created FID fields in the output shapefiles
arcpy.DeleteField_management(att_index_loc, "FID")
arcpy.DeleteField_management(spat_index_loc, "FID")

#Delete Index Tables from the user defiend geodatabase
arcpy.Delete_management(att_table_loc)
arcpy.Delete_management(spat_table_loc)
```

Thesis: Developing the housing attribute and spatial index tool to identify characteristic neighborhoods using variable importance factors calculated utilizing random forest regression modeling in ArcGIS PRO

Author: William A. Wallace

Signature:

Date: 04/30/2023