LAND VALUATION USING AN INNOVATIVE MODEL COMBINING MACHINE LEARNING AND SPATIAL CONTEXT

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

FERIDE TANRIKULU

Dr. Timothy Haithcoat, Thesis Supervisor

May 2023

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled:

LAND VALUATION USING AN INNOVATIVE MODEL COMBINING MACHINE LEARNING AND SPATIAL CONTEXT

presented by Feride Tanrikulu,

a candidate for the degree of Master of Data Science and Analytics,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Timothy L. Haithcoat

Professor Grant J Scott

Professor Matthew Foulkes

Professor Ilker Ersoy

This thesis is dedicated to my parents, who have always been my biggest supporters and who have made countless sacrifices to help me pursue my academic goals.

Feride Tanrikulu
University of Missouri – Columbia
May 2023

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to everyone who has contributed to this thesis, without whom this work would not have been possible.

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Timothy Haithcoat, for his unwavering support, insightful feedback, and patient guidance throughout my graduate studies. His mentorship and expertise have been invaluable to the successful completion of this thesis.

I would also like to thank my thesis committee members, Dr. Grant Scott, Dr. Matthew Foulkes, and Dr. Ilker Ersoy for their constructive criticism, and generous sharing of their time and expertise.

I am also deeply grateful to the Republic of Turkey for providing me with a scholarship to support my graduate studies and research. Their financial support has enabled me to focus on my studies and research and has been instrumental in the successful completion of this thesis. Thank you for investing in my future and for giving me the opportunity to pursue my academic goals.

Most importantly, I owe an enormous debt of gratitude to my family. My parents, sister, and brother have always been my biggest supporters, providing me with unconditional love and support. To my husband, thank you for believing in me even when I doubted myself. Your constant love and guidance have been my rock, and this thesis is a testament to our shared commitment to each other's dreams.

I am also thankful to my friends for their constant encouragement, understanding, and patience during this challenging journey. Their unwavering support and belief in me have been a constant source of motivation.

Thank you to all who have helped me along the way. Your contributions and support have been invaluable, and I could not have done this without you.

CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS AND SYMBOLS

ANN	Artificial Neural Networks
CART	Classification and Regression Tree
DT	Decision Tree
EVI	Enhanced Vegetation Index
GBM	Gradient Boosting Machine
GIS	Geographical Information System
GWR	Geographically Weighted Regression
HPM	Hedonic Pricing Model
KNN	K Nearest Neighbour
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLP-NN	Multilayer Perception Neutral Networks
MLR	Multiple Linear Regression
MRA	Multiple Regression Analysis
MSE	Mean Squared Error
MSI	Multi-spectral Instrument
NDBI	Normalized Difference Built-Up Index
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
OLS	Ordinary least Squares

PCA	Principal Component Analysis
RF	Random Forest
RMAE	Root Mean Absolute Error
RMSE	Root Mean Square Error
SHM	Spatial Hedonic Models
STC	State Tax Commission of Missouri
SVM	Support Vector Machine
SWIR	Short-Wave Infrared Bands
VARI	Visible Atmospherically Resistant Index
VIF	Variance Inflation Factor

LAND VALUATION USING AN INNOVATIVE MODEL COMBINING MACHINE LEARNING AND SPATIAL CONTEXT

Feride Tanrikulu

Dr. Timothy L. Haithcoat, Thesis Supervisor

ABSTRACT

Valuation predictions are used by buyers, sellers, regulators, and authorities to assess the fairness of the value being asked. Urbanization demands a modern and efficient land valuation system since the conventional approach is costly, slow, and relatively subjective towards locational factors. This necessitates the development of alternative methods that are faster, user-friendly, and digitally based. These approaches should use geographic information systems and strong analytical tools to produce reliable and accurate valuations. Location information in the form of spatial data is crucial because the price can vary significantly based on the neighborhood and context of where the parcel is located. In this thesis, a model has been proposed that combines machine learning and spatial context. It integrates raster information derived from remote sensing as well as vector information from geospatial analytics to predict land values, in the City of Springfield. These are used to investigate whether a joint model can improve the value estimation. The study also identifies the factors that are most influential in driving these models. A geodatabase was created by calculating proximity and accessibility to key locations as well as integrating socio-economic variables, and by adding statistics related to green space density and vegetation index utilizing Sentinel-2 -satellite data. The model has been trained using Greene County government data as truth appraisal land values through supervised machine learning models and the impact of each data type on price prediction was explored. Two types of modeling were conducted. Initially, only spatial context data were used to assess their predictive capability. Subsequently, socio-economic variables were added to the dataset to compare the performance of the models. The results showed that there was a slight difference in performance between the random forest and gradient boosting algorithm as well as using distance measures data derived from GIS and adding socioeconomic variables to them. Furthermore, spatial autocorrelation analysis was conducted to investigate how the distribution of similar attributes related to the location of the land affects its value. This analysis also aimed to identify the disparities that exist in terms of socio-economic structure and to measure their magnitude.

CHAPTER 1

INTRODUCTION

Valuation is a general term that can refer to any process of determining the value of an asset, while the appraisal is a more detailed and focused specific type of valuation that is conducted by a professional appraiser for a specific purpose such as tax assessment, insurance, urban development, or mortgage lending. Particularly considering ever-rising urbanization that leads to updating, the valuation of the land is one of the most essential constituents of any national economy and cadaster system (Pinter et al., 2020). However, there is a lack of objective spatiality considerations to be factored in regarding the appraisal process, which is able to provide more accurate and uniform values that are not influenced by inequities (Wei et al., 2022). This research aims to provide an unbiased, as accurate as possible, and consistent land valuation across the City of Springfield, Missouri so that undeveloped provinces utilize it as well as to identify and evaluate a joint model integrating machine learning and spatial context.

The land appraisal method is a widely studied topic within the fields of economics related to real estate, geography, and urban land (Pai & Wang, 2020). According to the literature, Hedonic Price Model (HPM) based on the demand theory and regression models measuring the relationship between the transaction price and attributes has been widely used (Osland et al., 2022). This traditional method is costly, time-consuming, susceptible to corruption, difficult to confirm, likely to be biased when it comes to location, and may only take into consideration a limited number of property features, resulting in potentially inaccurate appraisals (Aladwan & Ahamad, 2019); therefore, spatial variables and advanced technologies in the era of big data has incorporated into the traditional method on land and property valuation. Hence, one of the integrated improvements has been machine learning (ML) algorithms boosting valuation accuracy by providing a more objective and efficient way to determine the value of the land. Likewise, the assembling of remote sensing images and spatial modeling can remarkably improve the accuracy of valuation by determining the spatial characteristics and environment of the land. Lastly, Geographical Information

System (GIS) spatial analysis has been considered the fundamental tool as a supplementary analysis to manage spatial data to improve the precision of estimation value. (Brimble et al., 2020).

Appraisal methods are becoming more diverse and increasingly making progress toward more intelligent approaches. Scholars in the appraisal field tend to prefer machine learning and deep learning algorithms such as artificial neural networks (ANN) (G. Zhou et al., 2018), classification and regression tree (CART), support vector machine (SVM), and random forest (RF) (Antipov & Pokryshevskaya, 2012) due to their strong ability to generalize, fit accurately, and handle nonlinear mapping (Pai & Wang, 2020). There are currently limitations to using machine learning models in the appraisal, particularly when it comes to accounting for spatial factors. It is common to include longitude and latitude coordinates as input features in these models to analyze the impact of location on land prices, but it is difficult to accurately understand how these coordinates contribute to the model's predictions. There is a need to find a way to better incorporate location into machine learning models (Brimble et al., 2020). Meanwhile, spatial modeling approaches have also some limitations in terms of combining big geospatial data in vector form with remote sensing data in raster form, including the loss of spatial information converting data to values. Ultimately, spatial factors are a key aspect that sets land valuation data apart from other non-geographic data and are important to consider in real estate appraisal. While spatial modeling can improve the accuracy of appraisal compared to traditional methods, it is still not as effective as machine learning approaches. Nonetheless, it is criticized that ML is insufficient in terms of spatial considerations. (Wei et al., 2022). Consequently, to improve the accuracy and avoid inequity in real estate appraisal, it is important to utilize a range of regression methods and develop a more effective and reliable incorporated appraisal model. One way to achieve this is to combine the strengths of spatial regression and machine learning in a joint model, which can further enhance assessment accuracy.

Given the lack of research regarding a joint appraisal model combining spatial elements and machine learning, this study will aim to provide uniform and accurate land estimation values utilizing an innovative geo-machine learning approach. Furthermore, the predicted valuation will be compared to the actual valuations provided by the Office of the Greene County Assessor to assess differences, which will be further analyzed to better understand where potential biases or inequities within the city. According to the assessor manual by the State Tax Commission of Missouri (STC) (STC, 2017), to begin the valuation process, it is important to establish explicit comprehension of the appraisal question and set the boundaries and parameters for the task, which helps to prevent misunderstandings. Therefore, the objectives of this study, as outlined in the manual, are to identify the intended user, determine the intended use of the appraisal and the purpose of the appraisal, establish the date of the appraisal, identify the attributes of the land being appraised, and identify any assumptions or hypothetical conditions. These steps are important in order to direct and manage the appraisal process, determine the value of the land, and ensure that all necessary information is considered.

The ethics rule in the appraisal profession is designed to maintain trust and ensure that appraisers uphold certain standards of conduct, including avoiding fraud, criminal behavior, advocacy, and discrimination, and not engaging in misleading advertising or accepting compensation based on the outcome of an assignment (STC, 2017). Discriminatory appraisal practices in the past have widened the gap in land values based on race, ethnicity, and socioeconomic status, leading to inequity in access to resources and opportunities affecting the well-being of individuals and communities. Appraisers must use unbiased methods to ensure fair property values and avoid perpetuating existing disparities. Accordingly, this study will contribute to the body of knowledge with a novel approach by assessing land values objectively and more precisely as well as avoiding inequity in which the era of big data and computer science is rapidly and constantly changing. This will help address the current shortage of research and provide a sample model to be able to be applied in undeveloped cities for the administration of real estate policies and setting property taxes.

After establishing a foundation, as explained by STC, in order to build an assessment process, the appraiser collects both general (macro-level) data providing context about the subject property's market area including governmental, economic, and demographic/social characteristics and specific (micro-level) data pertaining to the subject property and comparable properties covering physical and financial details necessary for the valuation process. Moreover, it states that the appraiser should determine a discrete value for the property's land, when possible, as this is useful for the income or sales comparison approaches and critical for the cost approach. Meanwhile, the cost approach, explained by STC, is a widely used method for estimating the market value of land, particularly for tax purposes, by comparing it to similar properties. It is efficient and widely accepted by assessors because it can be applied to all types of property and the necessary data, which is primarily physical in nature, is readily available. Hence, the cost approach will be considered for this study (STC, 2017).

The value of land, defined in STC, is influenced by a range of factors, including location, accessibility, planning, zoning, utilities, and the economic and social environment, as well as topography, shape, industrial capacity, and accessibility to transportation and other relevant factors, and a standard appraisal system should consider all of these factors. Also, all factors are grouped under four headings which are physical, economic, governmental, and social factors. In this study, some of these factors will be used for land assessment as possible, like those used in the literature as well. The location (Droj et al., 2010; Osland et al., 2022; Pai & Wang, 2020), which is considered the most important physical factor in STC, and accessibility to amenities (Brimble et al., 2020; Hong et al., 2020) are physical factors. Planning and zoning (Brimble et al., 2020) are essential governmental factors, while socioeconomic factors (Brimble et al., 2020; Osland et al., 2022) may also affect land value and should be considered. Moreover, other factors about property attributes can be considered; the number of bedrooms and bathrooms (Aladwan & Ahamad, 2019; Baldominos et al., 2018; Pai & Wang, 2020), construction features (Pai & Wang, 2020; Wei et al., 2012), the size of each floor (Aladwan & Ahamad, 2019; Wei et al., 2012), age of the house (Aladwan & Ahamad, 2019; Baldominos et al., 2018; Droj et al., 2010), air conditioning (Aladwan & Ahamad, 2019), storage, garage, renovations, and curb appeal.

This study aims to develop an accurate and unbiased prediction model for land values using machine learning algorithms and spatial distance metrics derived from GIS data in the City of Springfield, Missouri. The study solely focuses on land valuation, excluding any improvements or structures built upon the land, or property valuation as a whole. The potential audience for this research includes property owners, real estate developers, financial institutions, government agencies, and legal professionals, who may have a vested interest in understanding land values for purposes such as selling, refinancing, development, collateral, land use planning, and legal disputes.

CHAPTER 2

BACKGROUND

Traditional property valuation relies on appraiser opinions and is subject to biases, leading to inaccurate valuations. Advanced valuation methods such as those using mathematical models and big data aim to improve the accuracy of property valuations and overcome the limitations of traditional methods, making model-based appraisals more cost-effective and reliable. Model-based appraisals are cheaper than traditional appraisals, but the accuracy depends on the amount of reference data available (Diaz and Hansz, 1997).

The HPM substantially based on regression models, which can be accepted as a conventional theory, has been widely used for appraising or estimating the value of the property (Sipan et al., 2012). Hedonic pricing indicates that the price of a product is the sum of what buyers are willing to pay for each of its unique features (Goodman, 1998). Demand theory by Lancaster and the supply-demand equilibrium model by Rosen form the underlying theoretical framework for HPM (Wen et al., 2005). According to (Lancaster, 1966), product demand depends on a product's characteristics rather than on itself, and goods are preferred as a combination of these characteristics. Rosen (1974) developed the characteristics demand theory to the hedonic pricing model by proposing the value of a good can be interpreted as the aggregate of internal and external utility-bearing attributes prices. HPM measures the relationship between the transaction price and attributes to emphasize the effects of certain characteristics on the price relevant to the housing structure and environmental amenities for property assessment, utilizing regression analysis to estimate the overall market price of the property (Gatheru and Nyika, 2015).

Studies have shown that the hedonic price model is a trustworthy and precise way to determine the value of real estate properties. It was found that the HPM gave a more accurate estimate of property value compared to conventional appraisal techniques, like the sales comparison method (Choi & Yi, 2021). The hedonic price model is advantageous due to its flexible and intuitive modeling that allows for easy observation of variable impacts. Its strong multi-parameter processing capability makes it favorable for scholars, particularly in the era of big data with a large number of housing price transaction cases (Liebelt et al., 2018; Selim, 2009).

The traditional hedonic pricing model employs the ordinary least squares (OLS) linear regression approach, which is preferred for its ease of simple and straightforward calculation (Wei et al., 2012). The method uses a mathematical equation to estimate the coefficients for each independent variable, considering only physical attributes, which are then used to predict the market price of a property. OLS is a widely used method for estimating the value of a property, but it assumes that the relationship between the dependent and independent variables is linear. However, it has limitations such as poor consistency and accuracy in assuming a linear relationship between features, which can be resulting in inaccurate estimations of property value if the relationship is not linear (Wei et al., 2012).

In the literature, linear regression models generally do not account for the spatial dependencies between properties in a given area in the literature. This is why spatial analysis, such as geostatistical methods, and incorporating spatial variables into the modeling process are essential for accurately predicting house prices (J. Zhou et al., 2018). Due to the limitations of traditional linear regression, researchers have tried to improve upon traditional methods using more advanced techniques by integrating the spatial elements that affect the real property, through spatial modeling including two approaches: spatial interpolation (Helbich & Kuntz, 2014) and spatial econometrics models (Wen et al., 2011). The former allows for estimating the price surface of an entire region with just a few sampling points, while the latter focuses on the impact of spatial stability and heterogeneity on housing prices.

OLS has improved to spatial hedonic models (SHM) being added locational attributes to mitigate the issue of spatial autocorrelation and to the Geographically Weighted Regression (GWR) model incorporating the weight factors of location, considering the influence of coordinates as well as other attributes (Sipan et al., 2012). Reed offers an intriguing method of incorporating spatial characteristics using demographic information into an OLS predictive model, arguing that it is often overlooked in many models. Principal Component Analysis (PCA) was used to reduce correlated demographic features and found a linear relationship between these features and sale price. This method has the potential to enhance house price prediction models by merging spatial information with census data (Reed, 2016). Similarly, in this thesis, socio-demographic data is utilized for the same purpose.

GWR is a statistical method that accounts for spatial heterogeneity, which is crucial in avoiding issues related to spatial non-stationarity. The model estimates coefficients for each independent variable, similar to linear regression, but takes into account the proximity of data points through the use of a kernel function. This proximity determines the weight assigned to each data point (Fotheringham et al., 2002).

Spatial autocorrelation is an important concept to consider in spatial analysis such as GWR. This concept is famously summarized by Tobler's First Law of Geography, which states that everything in a geographic area is related, but the strength of the relationship decreases with distance (Tobler, 1970). Spatial autocorrelation refers to the correlation between data points that are close in proximity. Some effects may be influenced only by nearby factors, while others may be influenced by factors across the entire city. Understanding spatial autocorrelation is important to ensure accurate modeling and analysis of geographic data. In a study conducted by (Brunsdon et al., 1999), the GWR model was evaluated against the OLS model predicting property value, and results showed that the GWR model performed better than the OLS model proved to be superior by resolving the issue of spatial autocorrelation.

Overall, spatial modeling has improved the accuracy of property assessments compared to traditional linear regression models, but they still rely on traditional statistical methods. Thus, spatial modeling and machine learning can be combined to improve valuation accuracy, but it's important to also include GIS spatial analysis in supplementary analysis, regardless of the method used. (Wei et al., 2022).

GIS technology is a powerful tool not only limited to displaying results and obtaining locational data but can also be used to develop prediction models by handling large data sets and conducting various analyses that have become essential for property appraisal. It can help researchers and practitioners make sense of large amounts of unstructured data and convert it into useful quantitative information. For instance, GIS can be used to analyze satellite imagery to measure vegetation indices, calculate land cover areas, estimate travel times and distances, and calculate accessibility indices (Wei et al., 2022). GIS's spatial analysis function allows researchers to identify patterns and relationships between real estate and infrastructure factors, providing insights into the underlying drivers of real estate values. Studies have demonstrated the value of GIS technology for real estate analysis, emphasizing its effectiveness in analyzing big data (Aladwan & Ahamad, 2019).

Besides, remote sensing technology has become an integral component of spatial analysis and GIS, as it enables researchers to evaluate external environmental conditions. The practical use of remote sensing technology mainly focuses on assessing the vegetation in urban green spaces by using the normalized difference vegetation index (NDVI) method. This method is considered an essential index for capturing information on environmental factors such as vegetation and impervious surfaces (Mora-Garcia et al., 2022). Moreover, other indices such as enhanced vegetation index (EVI), visible atmospherically resistant index (VARI), and normalized difference built-up index (NDBI) could provide valuable insights into vegetation. In this study, we will explore the applicability of NDVI, EVI, VARI, and NDBI for evaluating vegetation in urban green spaces and their potential for improving the accuracy of the land valuation model.

Some studies demonstrated raster data derived from remote sensing images with other geospatial data improves the performance of property appraisal by providing valuable insights into the neighborhood environment, which is closely linked to the value of a property (Bency et al., 2017; Yu et al., 2007). The impact of urban green vegetation on residential housing prices in Beijing using a unique housing transaction dataset and the hedonic price model was examined by Mei et al. (2018). Results show that urban green vegetation positively affects residential housing prices within 135m, with a potential increase of 7.95-10.59%, highlighting urban green spaces are a valuable amenity for residents.

It is widely acknowledged that no prediction model can accurately predict the output value for every observation when it comes to property values. Nevertheless, researchers have proposed methodologies that produce better predictions than linear regression. Utilizing machine learning techniques including neural networks has exhibited encouraging outcomes in this regard (Embaye et al., 2021).

Machine learning is a field of computer science that has evolved from artificial intelligence and computational learning theory without being explicitly programmed. The main aim of machine learning is to train algorithms' parameters to learn hidden patterns in data, which can then be used for prediction or classification tasks (Simon et al., 2016). Generally, machine learning can be divided into two categories: supervised and unsupervised learning. Supervised learning involves the use of target data, where the algorithm learns from a set of independent variables with an associated dependent response variable. The algorithm is trained on this data, with the goal of being able to predict the response variable for previously unseen data. Unsupervised learning, on the other hand, involves the use of unlabeled data, where the algorithm is given data without a response variable and is tasked with finding patterns based on common attributes. In this case, the algorithm is required to find patterns based on common attributes in the data with minimal human intervention (Hastie et al., 2009). While both types of machine learning have their uses, we will focus on supervised learning in this study, as our prediction task requires outputs of a response variable.

Many scholars have claimed the use of advanced machine learning models has been found to be more effective in predicting prices as compared to the current practice of utilizing multiple linear regression, indicating that it may be feasible to decrease the difference between predicted and observed prices in various contexts (Yacim & Boshoff, 2014; Yilmazer & Kocaman, 2020).

SVM is a powerful machine learning algorithm that can be used to forecast housing prices based on its characteristics, such as location, size, age, and amenities. By analyzing historical data of similar properties, SVM can learn to recognize patterns and relationships between the features and the property values. SVM can also be used to identify the most important features that contribute the most to the property's value which can then be used to predict the value of a new property based on its features. Unlike traditional methods that can suffer from curse of dimensionality due to large sample sizes, SVM is a small-sample learning method that relies only on a few support vectors to make decisions (Chen et al., 2017).

Researchers have explored that SVM has a strong ability to forecast short-term changes in prices and has successfully used SVM to predict changes in the Shanghai real estate market thanks to small sample learning despite the lack of comparable time series (Xie & Gang, 2007). Additionally, in a classification problem of determining whether house prices would go up or down, Banerjee & Dutta (2017) compared SVM, ANN, and the RF technique. The results show that while the RF technique was the most accurate, it suffered from overfitting. In contrast, SVM was the most consistent and reliable.

A decision tree (DT) is a model that helps make predictions and decisions by breaking down a problem into smaller, simpler parts. It is essentially a tree-like structure where each branch represents a decision, or a possible outcome based on certain conditions or variables. Decision trees have three key ingredients that make them successful. They are known for their simplicity, interpretability, and ability to handle both categorical and continuous variables (Song & Lu, 2015).

An investigation into the use of DT in forecasting Singapore real estate value was carried out by (Fan et al., 2006). They found that decision trees are more advantageous than traditional methods like LR because they can analyze both linear and nonlinear relationships between input and output variables, and are more interpretable, allowing users to identify the most influential attributes. In addition, the study revealed that homebuyers have varying priorities depending on the type of apartment they are interested in. Buyers of smaller apartments tend to prioritize the fundamental aspects of a home such as the age of the building. Meanwhile, buyers of executive flats place more importance on quality and service features such as the amenities offered and the surrounding.

A DT, however, can have disadvantages when used for property assessment, as highlighted in their study. While they excel at categorizing continuous variables into various categories, they may have trouble predicting a variable's precise value (Fan et al., 2006). Furthermore, James et al. (2013) stated that decision trees could be unstable because even a small change in the data could lead to big differences in the predictions.

RF is a popular machine-learning algorithm created by Breiman (2001). It works by combining multiple decision trees to make predictions based on the patterns observed in the input data. Each decision tree is built independently using a subset of randomly chosen predictors from the same dataset. The trees have independent "leaves" representing different decisions. After growing the desired number of trees, the predictions are averaged to get the final result. What makes Random Forest unique is that it randomly selects a specific number of trees from the decision tree results and takes their average. This technique is useful for predicting values based on multiple predictors and can produce more accurate results than using a single decision tree. Studies have demonstrated that in estimating property values, the RF technique outperforms the other algorithms mentioned in the Introduction chapter (Čeh et al., 2018; Dellstad, 2018; Ho et al., 2020; Ja’afar et al., 2021)

Various studies have explored the use of RF in predicting housing prices for different cities around the world. Hong et al. (2020) conducted a study to compare the effectiveness of the Random Forest method and OLS-based approach in predicting housing prices in Seoul, South Korea. They discovered that RF outperformed conventional HPM in several ways, such as its ability to handle both linear and nonlinear relationships without explicit user specifications and its proficiency in managing categorical variables. Furthermore, RF had higher accuracy than the OLS-based predictor, with 72% of predictions as opposed to 17%

in Multiple Linear Regression (MLR). Nevertheless, RF's multiple decision trees and use of a random sample of predictors make it more challenging to interpret the output, unlike regression models, which can be more easily explained through all predictors.

Masías et al. (2016) compared NN, SVM, and RF with conventional OLS in predicting residential housing prices in Santiago, Chile, finding that RF was superior to the other methods in terms of accuracy. (Yilmazer & Kocaman, 2020) used MLR and RF to analyze a portion of the commercial real estate in Ankara, finding that RF was slightly better than MLR. However, they highlighted that RF still has some limitations when it comes to selecting the number of the right variables for the model. Despite this, the RF has shown great potential in predicting housing prices, as demonstrated in previous studies.

(Hu et al., 2019) investigated forecasting ability through supervised learning algorithms for apartment rental prices in Shenzhen, China. The authors used RF, SVM, and multilayer perceptron neural networks (MLP-NN). The outcomes demonstrated the superior forecast performance of the RF algorithms. Similarly, Neloy et al. (2019) evaluated the performance of several algorithms including MLP-NN, RF, SVM, DT, ridge, and elastic net, finding that RF to predict rental prices in Dhaka, finding RF algorithm had better predictive accuracy compared to the other algorithms.

Gradient boosting is a machine-learning technique that builds a group of decision trees in an iterative process. Each new tree tries to correct the errors of the previous ones by fitting new trees to the residual errors. This approach can handle different types of data, including categorical and continuous variables using different types of decision trees. Besides, it can handle missing data, which is helpful when working with real-world datasets that often have incomplete information. However, careful tuning of hyperparameters is necessary to achieve optimal performance, and the algorithm can be computationally expensive (Gu & Xu, 2017). It is also possible for gradient boosting to be affected by outliers. This is because the model focuses on correcting errors from previous trees, which can cause it to prioritize accurately predicting outliers in the data. In other words, the model may try too hard to fit the outliers, which can lead to less accurate predictions for the rest of the data(Li & Bradic, 2018).

A study by Kagie & Wezel (2007) looked at how boosted decision trees performed in predicting the Dutch housing market. They compared the results to using MLR and found that boosted decision trees improved accuracy by more than 40%. However, like RF, the large number of trees used in the model makes it less interpretable. This means it may be difficult to understand how the model arrived at its predictions.

Ho et al. (2020) used SVM, RF, and Gradient Boosting Machine (GBM) to estimate property prices in Hong Kong. Their findings revealed that RF and GBM were more accurate in predicting housing prices than SVM. On the other hand, the authors also noted that machine learning algorithms have limitations in convenient feature selection, interpretability, and computational time when compared to conventional methods such as the hedonic pricing model. Despite these limitations, RF has several benefits in property valuation. It can effectively tackle missing values and categorical variables with levels, is not sensitive to outliers, and can eliminate overfitting problems (Mora-Garcia et al., 2022).

Technological limitations in the past restricted researchers from collecting adequate data. However, the advent of big data on the internet has led to the increased usage of ML over other traditional methods in various fields, including real estate valuation (Kauko, 2003). ANN algorithms are now commonly employed to forecast fluctuations in regional real estate prices (Daradi et al., 2018; Lim et al., 2016) and to evaluate individual properties (Ottomanelli et al., 2014; Selim, 2009).

An ANN typically consists of an input layer with independent variables, one or more hidden layers, and the output layer with the dependent variable. However, the complex prediction functions generated by the hidden layers, sometimes referred to as a "black box", are often difficult to comprehend, and the model may be overfitted with high-dimensional data (Guidotti et al., 2018). To address these challenges, researchers have developed two strategies - dimensionality reduction by mining correlations between various data and integrating advanced algorithms like genetic algorithms. Despite the promising results, ANN models are often criticized for their lack of interpretability, which could be a problem in cases where a clear understanding of the prediction process is necessary, such as in land appraisal and tax calculations (Peterson & Flanagan, 2009).

Limsombunchai et al., (2004) carried out a research study in Christchurch, New Zealand, where they explored the use of artificial neural networks for predicting property prices, and considered various factors, such as descriptive features of the house, amenities around it, and its geographical location. The results showed that neural networks have significant advantages over traditional HPM. One of the main advantages is their flexibility and nonlinear properties, allowing them to learn any problem without the need for specifying the details of the structure or parametric form in advance. Unlike MLR, the neural network can determine the appropriate functional form on its own, which can lead to more accurate predictions. The study found that in some cases, the neural network had a prediction accuracy improvement of nearly 50% compared to MLR. Even though previous studies have pointed out that neural networks are black boxes and come to varied conclusions, the results promoted their potential for using property appraisal.

Nguyen & Cripps (2001) evaluates the performance of ANN and multiple regression analysis (MRA) in predicting single-family property prices. ANN outperforms MRA when a moderate to large data sample size is used, which ranges from 13% to 39% of the total data sample. The results suggest that previous studies comparing MRA and ANN in predicting housing values may have produced varied results due to differences in data sample sizes. Studies with small sample sizes could have led to varied results, whereas moderate to large sample sizes may have found that ANN performs better than MRA. Therefore, using a moderate to large sample size is crucial in accurately comparing the predictive performance of MRA and ANN in predicting housing values.

Remote sensing technology has become increasingly crucial in property valuation to evaluate external environmental conditions owing to presenting the aerial view of a property and its surrounding. The practical use of remote sensing technology mainly focuses on assessing the vegetation in urban green spaces by using the normalized difference vegetation index (NDVI) method. This method is considered an essential index for capturing information on environmental factors such as vegetation and impervious surfaces (Mora-Garcia et al., 2022). In addition to NDVI, there are other remote sensing indices that can provide valuable insights into vegetation, impervious surfaces, and built-up areas. For example, the enhanced vegetation index (EVI) is used to monitor changes in vegetation conditions and can detect changes in both high and low vegetation density areas (Huete et al., 2002)Next, the visible atmospherically resistant index (VARI) can be used to estimate vegetation density and can be useful in areas with mixed vegetation cover, taking into account the effects of atmospheric scattering and absorption on vegetation reflectance (Gitelson et al., 2002). Additionally, the normalized difference built-up index (NDBI) is a commonly used index for mapping urban built-up areas and can be used to distinguish between impervious surfaces and non-impervious surfaces, such as vegetation and bare soil (Zha et al., 2003).

In our study, we believe that incorporating these additional indices in addition to NDVI can provide a more comprehensive understanding of the environmental conditions surrounding a property. By utilizing these indices, we can gain insights into the amount of vegetation cover, impervious surfaces, and built-up areas in the vicinity of a property, which can provide valuable information for land valuation. Furthermore, by exploring the applicability of these indices in our modeling approach, we can potentially improve the accuracy and reliability of our land valuation model.

Some studies demonstrated raster data derived from remote sensing images with other geospatial data improves the performance of property appraisal by providing valuable insights into the neighborhood environment, which is closely linked to the value of a property (Bency et al., 2017; Yu et al., 2007). The impact of urban green vegetation on residential housing prices in Beijing using a unique housing transaction dataset and the hedonic price model was examined by Mei et al. (2018). Results show that urban green vegetation positively affects residential housing prices within 135m, with a potential increase of 7.95-10.59%, highlighting urban green spaces are a valuable amenity for residents.

Spatial modeling techniques have been developed to estimate land value by incorporating spatial autocorrelation into traditional regression models. However, it still faces limitations similar to classic regression models when compared to advanced machine learning algorithms, which offer more capabilities for data analysis and prediction. Despite this, a major limitation of using machine learning in the valuation literature is that it often neglects spatial components of the data, which can limit their effectiveness. To overcome this limitation and improve the accuracy of a land appraisal, this study aims to develop a more accurate and reasonable land valuation model by incorporating spatial context into data for machine learning algorithms.

The primary objective of this study is to develop a more accurate and reliable land valuation model that can support informed decision-making processes. Specifically, this study aims to demonstrate the importance of incorporating spatial relations into land value estimation. Moreover, the study will conduct a comparative analysis of land features based on their spatial distribution and corresponding values throughout the city. The study will employ GIS analysis techniques to extract spatial context data, including distance measures and other related factors. The aim is to understand how the location of the land influences its value and identify the most critical spatial features that impact land valuation. Additionally, this study aims to explore the role of urban green spaces in influencing land value by utilizing remote sensing tools and measures. Ultimately, the findings of this study are expected to provide insights that can improve land appraisal accuracy and contribute to the advancement of land valuation methods that can better reflect the true value of land in urban environments.

CHAPTER 3

METHODOLOGY

3.1 Data

This chapter outlines the data acquisition process and the geographic location from which the data was obtained. The decision was made to create a customized data set for this project to gather more spatial information using GIS since none of the datasets seen in the literature include spatial context comprehensive with descriptive and socio-demographic attributes. Springfield, Missouri in the United States was chosen as the center of our data collection. The chapter provides an overview of the data carpentry and geo-analysis steps taken to prepare the data for modeling.

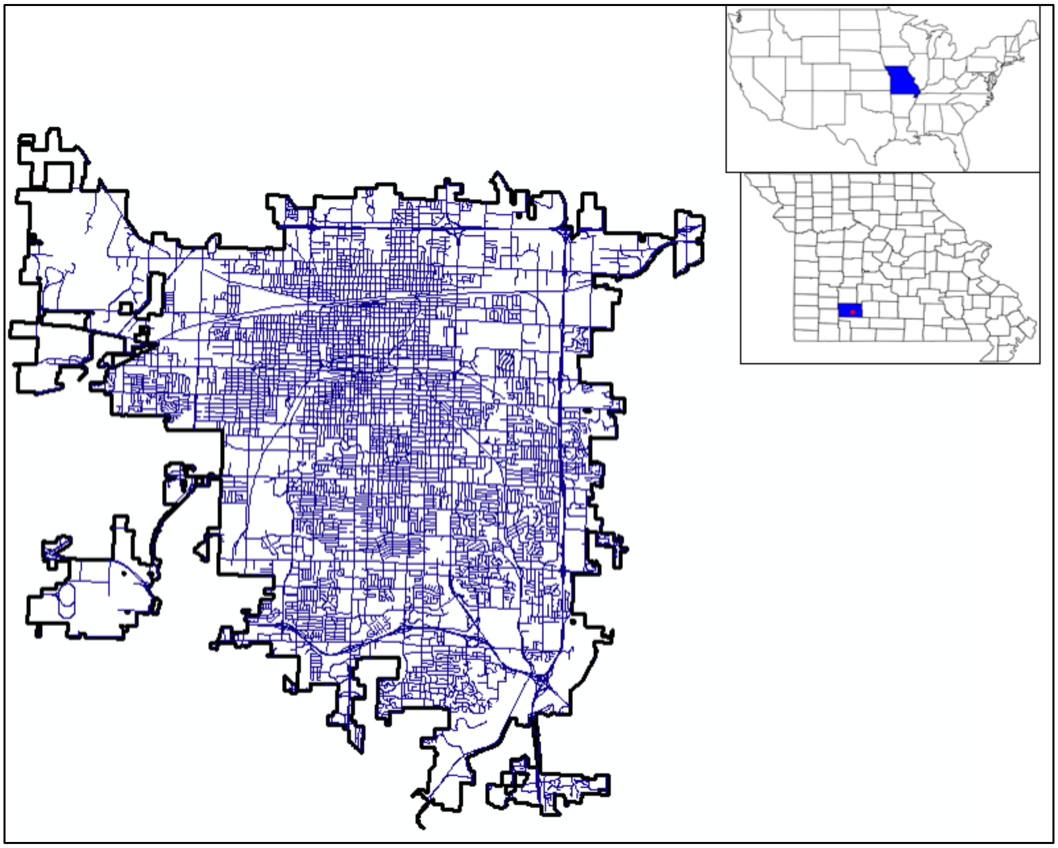


Figure 3. 1 Location of Springfield, Missouri selected as the study area in the United States with roads vector.

Springfield's growing population, diverse economy, and stable real estate market make it significant for property valuation research. Accurate property valuations are in high demand in the city, making a predictive model for land value important for stakeholders like real estate professionals, investors, and local government officials.

The data used to collect information about the land and its location in this study were sourced from four distinct sources. Firstly, vector data for most features, excluding hospital data, were obtained from the open-source GIS data belonging City of Springfield, Missouri1. This included the following data layers: city limits, roads, police, fire, schools, parks, greenway trails, bike routes, sinkholes, flood plain. Secondly, point data for hospitals was acquired from the Homeland Security Infrastructure Program (HSIP), which is an infrastructure geospatial data inventory managed by the Homeland Infrastructure Foundation-Level Data (HIFLD)[2] in partnership with the National Geospatial-Intelligence Agency. Thirdly, the socioeconomic data used in this analysis were collected from the American Community Survey (ACS), which is conducted by the United States Census Bureau, and provides a wide range of demographic, social, economic, and housing characteristics for communities across the country[3]. Variables of interest pulled from the larger dataset of Greene County Block Group numbered 833 and included population, household income, age, race, education, economic conditions, etc. These were then analyzed to create a final selected list which is described below. Next, the Sentinel-2 raster data, downloaded from the Copernicus SciHub in 2022[4], providing high-resolution

[1] The data used in this study were obtained from the City of Springfield’s open data systems, which were accessed in January 2023. The data were modified for use from their original source, www.springfieldmo.gov, the official website of the City of Springfield, Missouri. The City of Springfield makes no claims as to the content, accuracy, timeliness, or completeness of any of the data provided at this site. The data provided at this site is subject to change at any time. It is understood that the data provided at this site is being used at one’s own risk.

[2] The Homeland Infrastructure Foundation-Level Data (HIFLD) website, located at https://hifld-geoplatform.hub.arcgis.com, contains information that is marked U//FOUO and is intended for the exclusive use of Government and Contractor personnel with need-to-know HIFLD information. As such, this information is specifically prohibited from posting on unrestricted websites or other unrestricted applications.

[3] U.S. Census Bureau, American Community Survey (2021), Greene County Block Group. This analysis is not endorsed or sponsored by the United States Census Bureau, and any opinions, findings, and conclusions expressed here are those of the author(s) and do not necessarily reflect the views of the Census Bureau.

[4] https://scihub.copernicus.eu, the Copernicus Open Access Hub provides free and open access to Sentinel-2 data.

imagery of the study area was utilized to obtain additional land information. Finally, the appraisal value of the land was utilized, denominated in US dollars, as a proxy for land value. This parameter serves as the dependent variable in the models and is sourced from the Office of the Greene County Assessor[5], ensuring its accuracy and reliability. Thus, by combining these diverse data sources, we were able to create a comprehensive geodatabase that allowed us to analyze various aspects of the study area.

3.2 Data Carpentry and Exploratory Data Analysis

Geodatabases are specialized types of databases designed to store and manage geographic information in a structured manner. They are typically used in software such as ArcGIS, QGIS, etc., and provide a complete information model for managing and representing geographic data. Within a geodatabase, there are three primary types of datasets: feature classes (points, lines, or polygons), raster databases, and tables. Personal geo-databases are a commonly used format within ArcGIS and offer numerous advantages such as allowing data to be shared easily, the ability to compress files to save space, a longer limit on field names, and the provision of null data. Therefore, we have created a geodatabase utilizing ArcGIS to manage the data from different sources as well as create auxiliary spatial information derived from spatial analytics to be applied as explanatory variables.

Vector data, including point, line, and polygon data, are used in GIS to represent real-world features such as roads, buildings, and land area. They are preferred for their ability to represent these features in great detail and accuracy, making them useful in spatial analysis tasks such as proximity analysis, buffering, and overlay analysis. On the other hand, raster data are used to represent continuous data such as land use/land cover and vegetation. They are useful in representing and analyzing data over large areas, which is difficult to achieve with vector data.

[5] https://greenecountymo.gov/assessor/

This study used vector and raster data to derive parcel-level metrics tied to a point database that represents the centroid of each Single-Family Residential parcel in the Springfield database for subsequent machine-learning applications. These data were transformed into metrics that were tied to the parcel or evaluated for the parcel area, and then to the parcel point database. For example, the polygon data represented land use types, such as residential or commercial areas. Line data, such as roads were built into a network database in order to conduct network analysis to calculate accessibility to the amenities for each parcel. Point data, such as addresses, were used to represent the centroid of each parcel. Lastly, raster data were used to calculate the density of green areas. This was achieved by deriving the density of green areas from each pixel in the raster data and then performing zonal functions to compile this information for each parcel which were then linked to the points in the geodatabase. This allowed us to analyze the green space coverage in each parcel and its surrounding areas, which we felt might be an important parameter for land valuation calculations.

The decision to use a point database for subsequent ML applications and modeling was based on the need to have a unique identifier for each parcel. By having a unique identifier for each parcel, it becomes easier to link the parcel-level metrics derived from vector and raster data to other datasets such as socio-economic data for modeling purposes. By integrating socio-economic information data with the parcel-level metrics, we were able to conduct a comprehensive analysis of the study area, which can have various implications for policy and decision-making. The use of these diverse data types allowed us to capture various aspects of the study area, including its physical features, land use, and socio-economic characteristics, which can help us understand and address complex challenges in the area.

The data carpentry process started by cleaning the data downloaded from the City of Springfield. This involved removing data points that were located outside of Missouri and narrowing down the data to include only residential areas with single-family zoning. To achieve this, we used both address points and zoning polygon layers. However, during the data cleaning process, we discovered that some of the polygons or parcels contained multiple address points. Therefore, we performed manual verification on each polygon to ensure that it was associated with only one address point representing single-family residential land, left with 43,445 out of the initial 52,087 points. By taking the time to review each polygon individually, we were able to eliminate any potential errors or inaccuracies that may have been introduced during the data-cleaning process (Figure. 3.2).

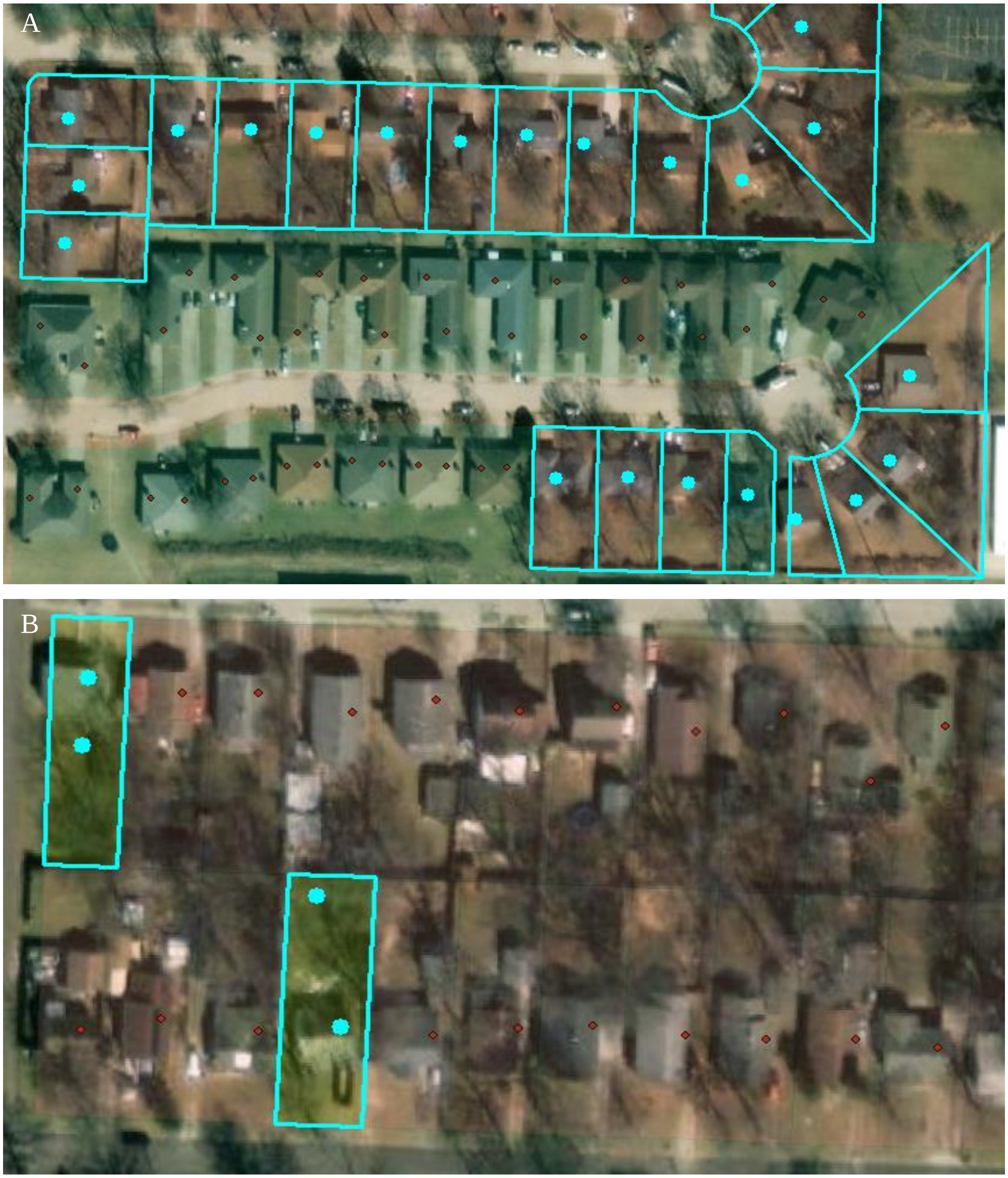


Figure 3. 2 Manual verification process of each polygon is linked to only one address point for single-family residential land. (A) the identification of appropriate polygons, (B) identified single-family residential polygons (represented by red colors), misidentification (represented by green colors)

3.2.1 Distance Contextual Measures:

The term "location features" in this paper pertains to the attributes of the surrounding area and nearby structures. According to the literature, factors such as easy access to amenities like convenience shops, parks, and hospitals can increase the value of a house. Additionally, the time it takes to travel to important locations like the city center may be a more important factor in determining the value than just the distance from these locations (Guliker et al., 2022). Thus, geospatial analysis was conducted by calculating the proximity of various features nearby such as police stations, fire stations, bike routes, trails, parks, and schools, with the different educational institutions being categorized into elementary, middle, high, and early childhood. The Near tools in ArcMap 10.8.2 were utilized to layer these features and calculate their distances within 52800 feet. This information could provide valuable insights into the potential positive impact of these nearby features on land value. However, it's also important to consider the potential negative impact of certain factors on land value. For example, if a nearby highway were to close due to noise pollution, it could have a negative impact on the value of nearby parcels of land. Therefore, the distances of these highways should also be calculated in the geospatial analysis.

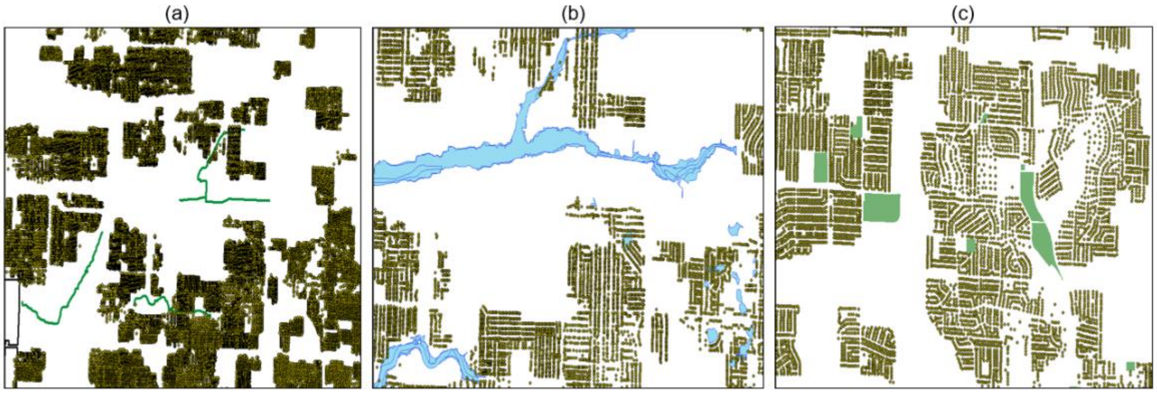


Figure 3. 3 Geospatial layer features estimated based on distances (A) Greenway trails, (B) Floodplain, (C) Parks

3.2.2 Network Analysis:

Following this, road network analysis extension was conducted to evaluate the connectivity of each point to schools, police and fire stations, and hospitals, calculating travel time for both walking and driving. By conducting this road network analysis, we were able to gain a better understanding of how accessible each point was important features, which could help to inform land valuations and potential development opportunities. In the resulting data table, the travel time to hospitals, fire stations, and police stations remained as driving time. However, for the travel time to elementary schools, it was calculated as walking time. This distinction was made because walking time is a more relevant factor for proximity to schools while driving time is more relevant for proximity to emergency services.

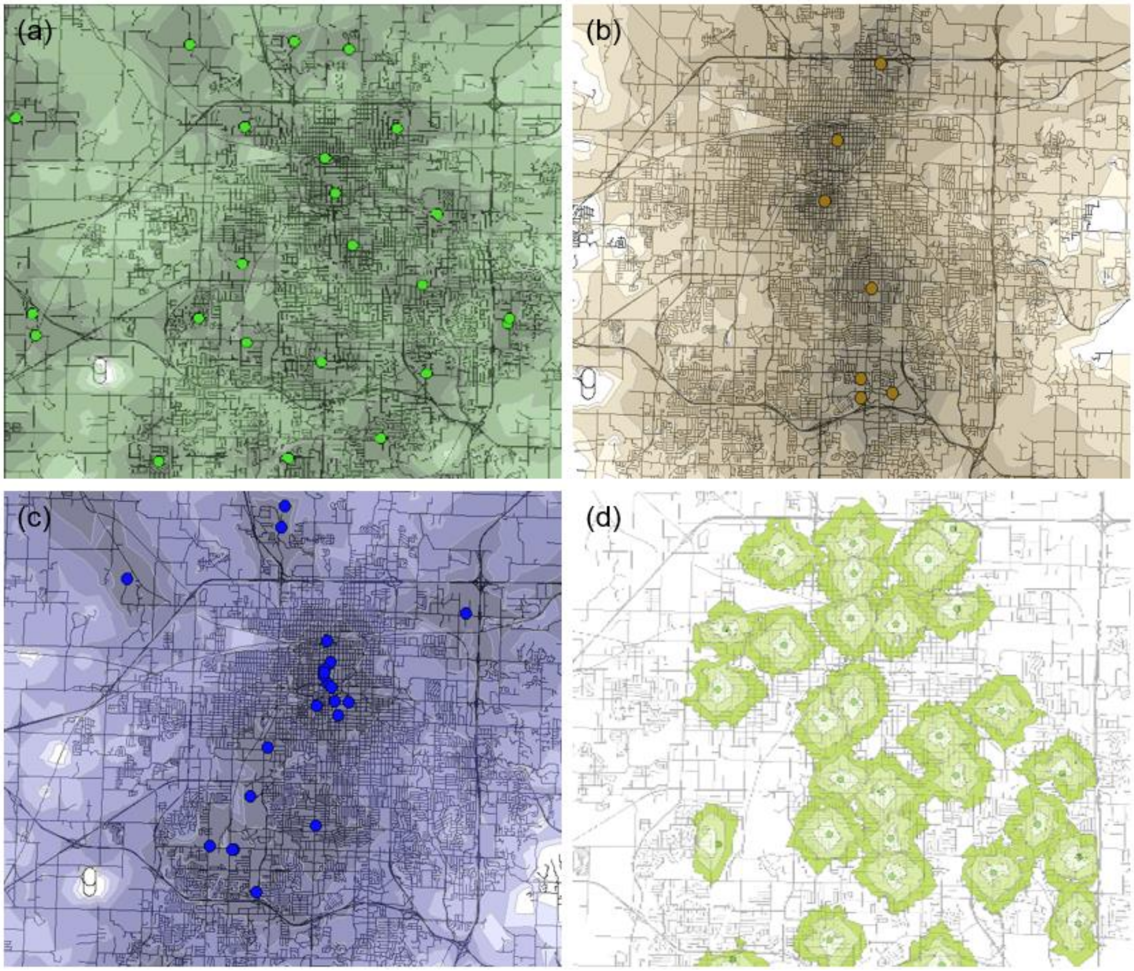


Figure 3. 4 Road network extension and map of distances (service area) with the centers (a) fire stations, (b) hospital, (c) police stations, and (d) elementary school.

3.2.3 Raster Imagery Processing:

Sentinel-2, a series of earth observations satellites, is equipped with a custom Multi-spectral Instrument (MSI) sensor, which enables the provision of remote sensing imageries with high spatial, spectral, and temporal resolutions. The 13 bands provided by Sentinel-2 cover visible, near-infrared, and shortwave infrared information, which can be used for various purposes such as atmospheric and geophysical parameter correction, vegetation detection, and land classification. Moreover, Sentinel-2 imagery can be applied in different fields including land use and land cover mapping, crop monitoring, disaster management, and environmental monitoring. The high level of detail provided by Sentinel-2 images facilitates accurate analysis of land features and changes over time. The free and open data policy of Sentinel-2 makes it a cost-effective and accessible tool for both researchers and professionals.

The Sentinel-2 satellite imagery used in this study was acquired through the official Copernicus Open Access Hub website as mentioned previous section called data. To select the imagery, criteria such as cloud cover, acquisition date, and geographic location were taken into account. Images with low cloud cover and minimal atmospheric interference were selected for analysis. Two dates, one in June and one in November 2022, were selected for the analysis. This allowed for the comparison of vegetation and land cover changes over time, as well as the detection of any seasonal variations. The June image captured the beginning of the growing season, while the November image captured the end of the growing season. This enabled avoiding any lack of vegetation index due to the season. Consequently, in this study, Sentinel-2 satellite imagery acquired in June and November 2022 was used to analyze vegetation index and land cover information with the assistance of ArcGIS and Python.

At the end of the analysis, the difference in statistics value between vegetation indexes for June and November was used as a basis for comparison. By calculating the difference in vegetation index values between the two dates, the study was able to determine the magnitude of vegetation changes that occurred over the growing season. This information was critical for understanding how changes in vegetation cover could impact land valuation.

When determining vegetation abundance and monitoring changes in plant health, NDVI is used to measure how green the vegetation is. According to conventional methods, NDVI is defined as a ratio between the red and near-infrared (NIR) values which are expressed as:

$$NDVI = \frac{NIR - Red}{NIR + Red} \qquad (1)$$

This study utilized the 10-meter resolution imagery obtained from Sentinel-2 to generate the Normalized Difference Vegetation Index (NDVI) for two distinct seasons, namely June and November (Fig 3.5).

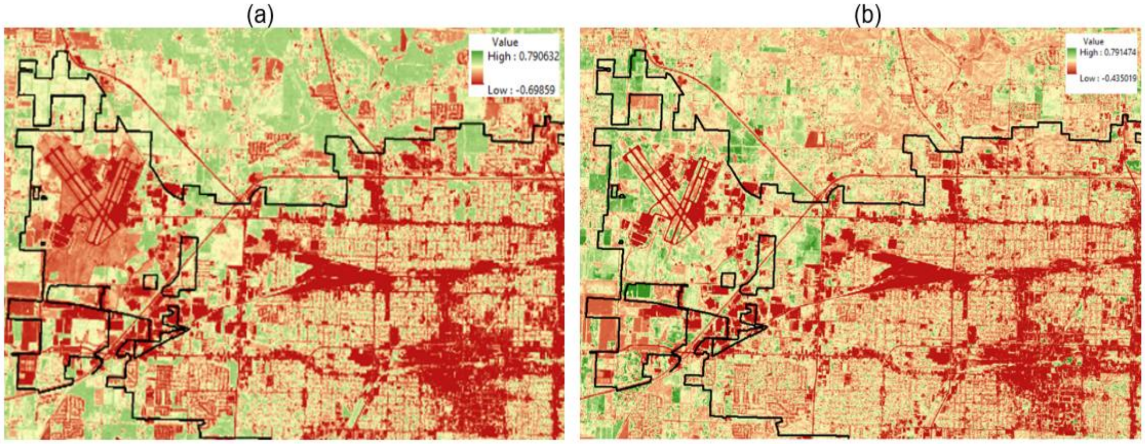


Figure 3. 5 Normalized Difference Vegetation Index from 10-meter resolution imageries derived from Sentinel-2 at (a) NDVI June, (b) NDVI November

Additionally, the Enhanced Vegetation Index (EVI) is a modification of the Normalized Difference Vegetation Index (NDVI), designed to correct noises and increase vegetation monitoring accuracy (Figure 3. 6). It uses blue, red, and near-infrared bands to minimize background noise and atmospheric interference and calculates the ratio between the red and near-infrared bands. The formula for calculating EVI is:

$$EVI = = 2.5 * ((NIR - RED) /(\mathrm{NIR} + 6 * \mathrm{RED} - 7.5 * \mathrm{Blue} + 1)) \qquad (2)$$

which is designed for Sentinel-2 raster with 10-m resolution bands and used in this study.

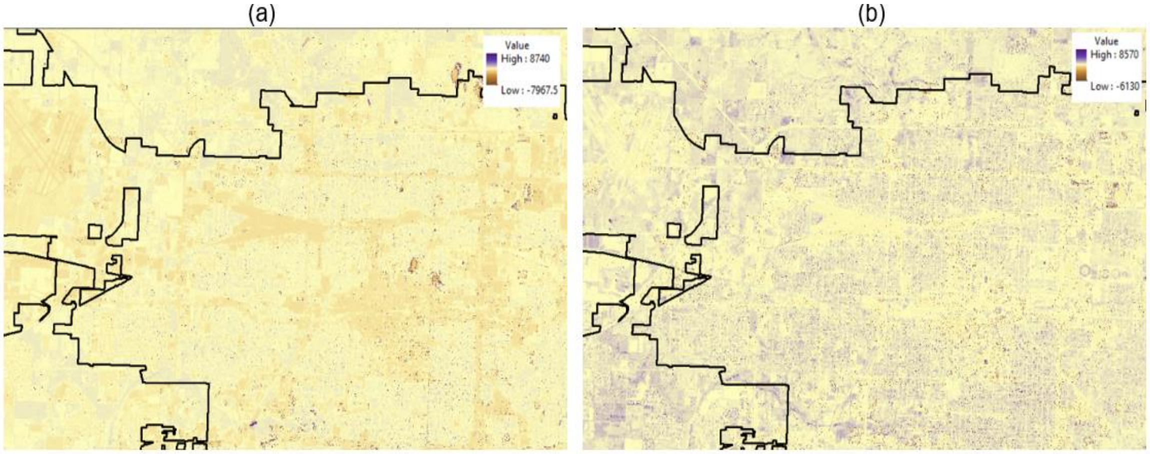


Figure 3. 7 Enhanced Vegetation Index of the 10-meter resolution imageries from the Sentinel-2 at (a) EVI June and (b) EVI November

Subsequently, another vegetation index investigated was the Visible Atmospherically Resistant Index (VARI), a method used to measure vegetation using visible light, especially in areas with atmospheric interference (Figure 3.7). It is resistant to atmospheric conditions like dust or haze and is widely used in remote sensing for vegetation analysis. On the other hand, NDVI is more effective in areas with high vegetation coverage, while VARI is more suitable in areas with sparse vegetation or where other factors are affecting the reflectance values. VARI was calculated using 10-meter resolution bands to estimate vegetation using the formula which is:

$$\mathrm{VARI} = \frac{(\mathrm{Green} - \mathrm{Red})}{(\mathrm{Green} + \mathrm{Red} - \mathrm{Blue})} \qquad (3)$$

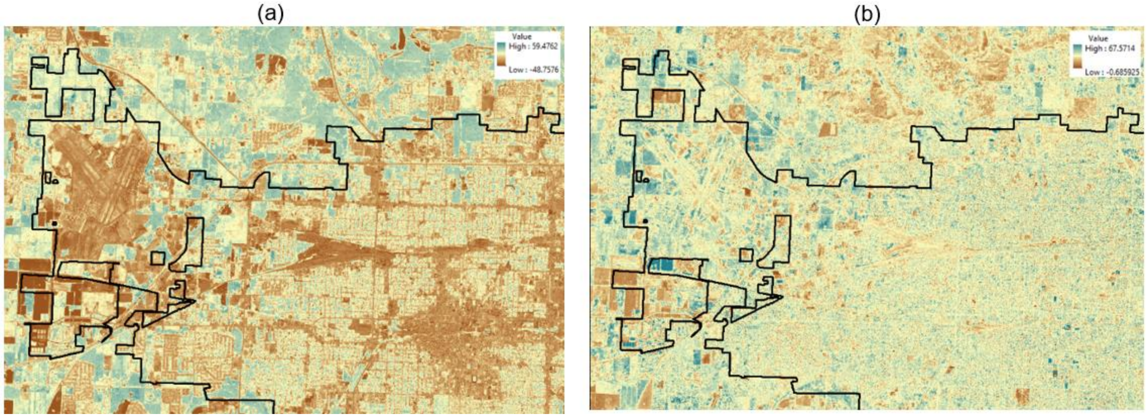


Figure 3. 8 Visible Atmospherically Resistant Index (VARI) measurement of vegetation using visible light within atmospheric interference area. (a)VARI June, and (b) VARI November

The Normalized Difference Built-Up Index (NDBI) is a remote sensing measure for identifying built-up areas like cities and towns. By comparing near-infrared and short-wave infrared bands (SWIR), the NDBI measures the number of built-up materials present in an area. The index provides values ranging from -1 to 1, where positive values indicate built-up areas and negative values indicate non-built-up areas. The NDBI formula:

$$NDBI = \frac{(SWIR - NIR)}{(SWIR + NIR)} \qquad (4)$$

is ratio-based to compensate for variations in terrain and atmospheric conditions. Due to the unavailability of the Short-Wave Infrared (SWIR) band for 10-meter resolution in Sentinel-2 imagery, in this study, the 10-meter NIR and 20-meter SWIR bands were used to generate NDBI for June and November (Figure 3. 8). In addition, another NDBI was created by utilizing only 20-meter resolution bands to compare any variations. We wanted to compare the results we got from the different combinations of light to see if there were any differences that affect the response variable.

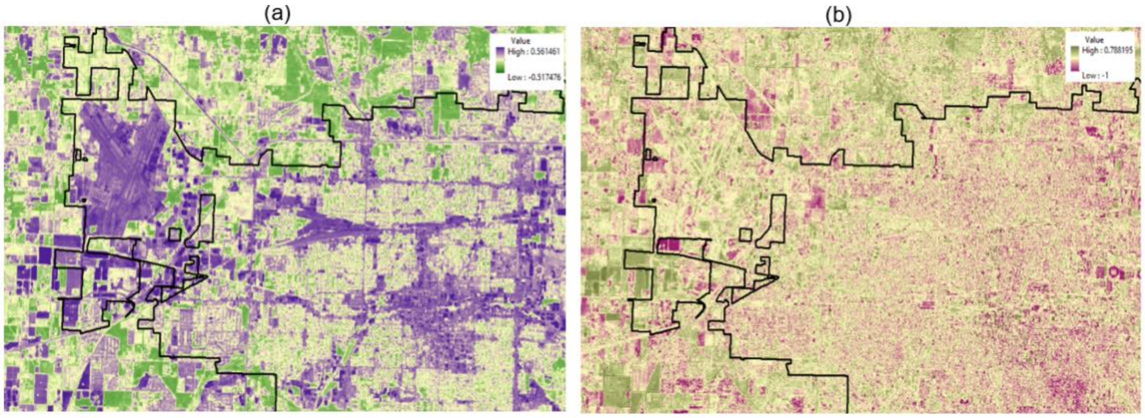


Figure 3. 9 The 10-meter NIR and 20-meter SWIR bands were used to generate NDBI for (a) June and (b) November.

Ultimately, the zonal statistic method was utilized to incorporate statistical values of the generated indexes into the data set.

To gather additional information through remote sensing, a land cover map was generated with seven distinct classes: high and low-density forest, built-up areas, grass, bare ground, and water. The SVM classifier is a powerful tool in remote sensing analysis that can accurately classify complex land cover types (Jog & Dixit, 2016; Soni et al., 2021). Initially, a signature file was manually created by drawing polygons as representative sample areas for each land cover type, which served as the basis for training the SVM Classifier in ArcMap software. Once the training was completed, the raster data was classified using SVM Classify tools. A zonal statistic table was also generated to incorporate the relevant data into the dataset, providing useful information about the characteristics of each land cover type in relation to the surrounding areas. By calculating statistics such as mean, median, and standard deviation for each zone or polygon, it is possible to gain insights into the spatial distribution and variability of the land cover types, and their relationship to other environmental factors. Figure 3.9 provides insight into the comparison between the created land cover map and the world imagery of the study area.

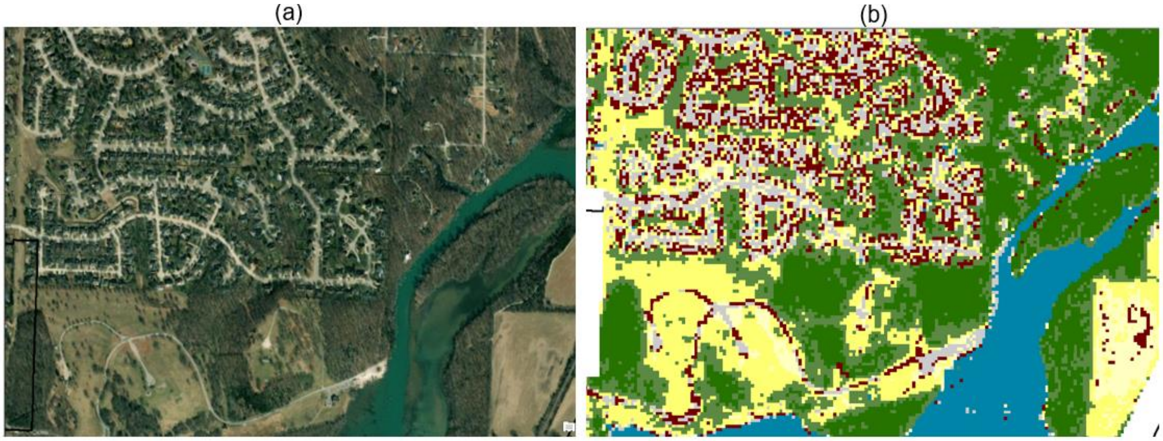


Figure 3. 10 Satellite World imagery vs SVM Land Cover Classification.

In this study, zonal geospatial analytics were used to calculate variety and majority statistics which were then added to the table. These metrics were deemed suitable for the purpose of our analysis providing a more comprehensive understanding of the distribution and diversity of land cover types. The majority can provide valuable information on the dominant land cover class within a particular area or pixel, while variety provides information on the heterogeneity or diversity of land cover types within a given region.

3.2.4 Analytical Database Assembly:

After conducting the vector and raster analyses, a large dimensional data table was created, consisting of 43,445 points. The next step in the analysis process involved adding land appraisal value, which is the target value with their calculated areas, obtained from the Office of the Greene County Assessor. During the data exploration phase, a total of 91,648 land appraisal values were identified for the year 2022. However, after implementing a spatial join via address with the large data table, the number of values was reduced to 15,062. The discrepancy in numbers occurred because not all of the 43,445 points in the dimensional data table were associated with land appraisal values. This may have been due to a variety of reasons, such as missing or incomplete data on the part of the Office of the Greene County Assessor, or differences in data collection methods between the two sources. Additionally, the spatial join via address may have excluded some data points that did not have a corresponding address in the land appraisal dataset. Figure 3.11 shows the distribution of land appraisal values for the two scenarios, namely before and after conducting the spatial join.

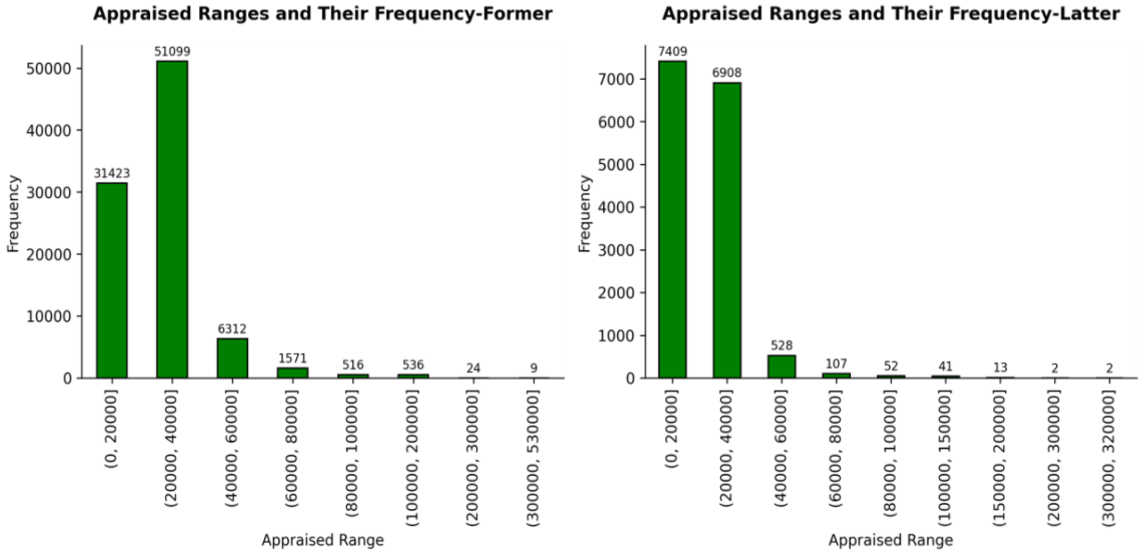


Figure 3. 11 The distribution of land appraisal values (a) before and (b) after spatial join

To ensure accurate results in our analysis, we performed outlier analysis on both the land value and calculated areas. Our aim was to identify any extreme values that could potentially distort our model. From the results of this analysis, we identified land values greater than $200,000 as outliers and removed them from our dataset. We also removed calculated areas with values greater or less than six standard deviations from all quantitative characteristics, resulting in the removal of 65 outliers.

It's important to note that the land value was initially calculated as an acre, but for consistency, the land value has been recalculated for a quarter acre. After applying these cleaning processes, we were left with 14,995 values in our dataset. These steps were taken to ensure the integrity of our analysis and to provide accurate insights from our data.

To address issues with heteroscedasticity and improve the fit of the data, the dependent variable, land value, was converted into natural logarithms. This transformation also made it easier to interpret the coefficients, as they now indicate the percentage change in the dependent variable for every one-unit increase in the explanatory variable.

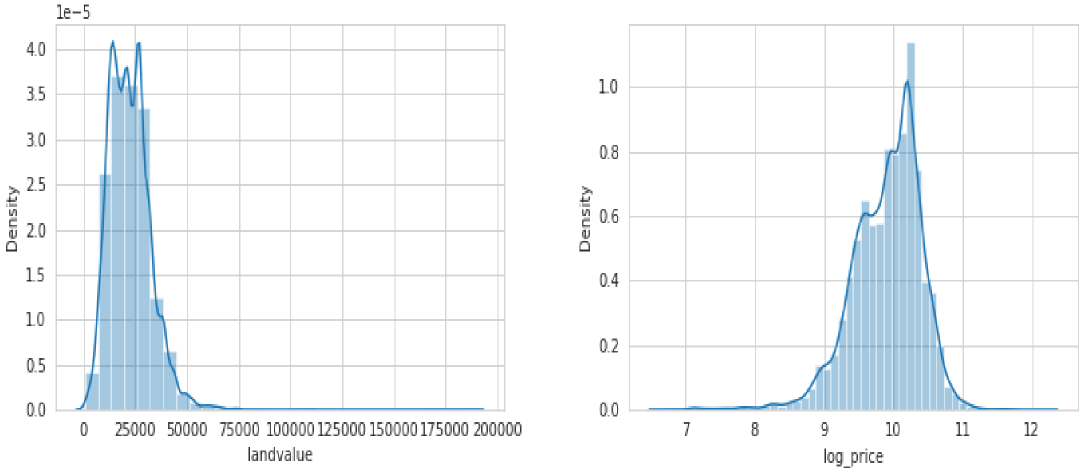


Figure 3. 12 The distribution (a) positively skewed and (b) the natural logarithm of land appraisal values in the dataset.

During the feature engineering stage, the relationships between features were examined, and highly correlated features were removed to avoid an unstable model, especially for linear regression. The variance inflation factor (VIF) was utilized to assess collinearity among features, with a VIF value above 10 indicating potential collinearity issues. Additionally, features with low variance or a high proportion of missing values were excluded.

The dataset represents neighborhoods in two dimensions, with GIS-derived spatial distances context data and socio-demographic features by the census. After applying the Pearson correlation and VIF analysis, the initial 52 GIS variables were reduced to 16, which will serve as independent features for initial modeling. These initial variables include distance and travel time to early childhood, elementary, middle, and high school, fire and police stations, hospitals, public healthcare, urgent care, and nursing home; distance to sinkholes, flood plains, and highways; mean and standard deviations of four different vegetation indexes for June and November, and values for land cover classes. Some of them were removed due to a correlation between similar features. Thus, there were 11 distance-based features: distance to flood plain, sinkholes, greenway trails, bike routes, parks, highways, and early childhood schools; travel time to elementary schools, hospitals, police, and fire stations. As well, 5 raster-based vegetative indices were included: the

(June-November) difference between means of NDVI, NDBI, VARI, and EVI, as well as the majority value for land cover.

Subsequently, a new table was created using manual encoding from these variables. The aim was to facilitate further analysis of the dataset by assigning each variable a value within the range of 1 to 3, based on specific criteria that enabled the grouping of variables.

Floodplains were assigned a value of 1 if they had a proximity of a flood plain is 0, while those with a size less than or greater than 50 meters were assigned values of 2 and 3, respectively. Similar criteria were employed for sinkhole areas and distance from the highway as their proximity to the land is likely to cause lower land value. However, variables that positively affect land value, such as distance and travel time, were assigned 3 for lower distance and time. The vegetation and built-up indexes were assigned values based on their mean statistics, while the forest class was assigned a score of 3 for land cover. The scores of all the features were summarized in a new column, providing a comprehensive overview of the neighborhood. Using this new dataset, spatial autocorrelation, and group analysis were performed to analyze the distribution of the best conditions and high values.

3.2.5 Socioeconomic Data Analysis:

After applying the Variance Inflation Factor (VIF) to census data, variables exhibiting multicollinearity were eliminated to enhance the accuracy of the predictive model for the second dataset. From the variables that were found to be significant, the total population within a block area was removed to reduce multicollinearity with the total household income variable. Additionally, the VARI variable was removed to address the same correlation issue with NDVI, as illustrated in the Figure 3.13.

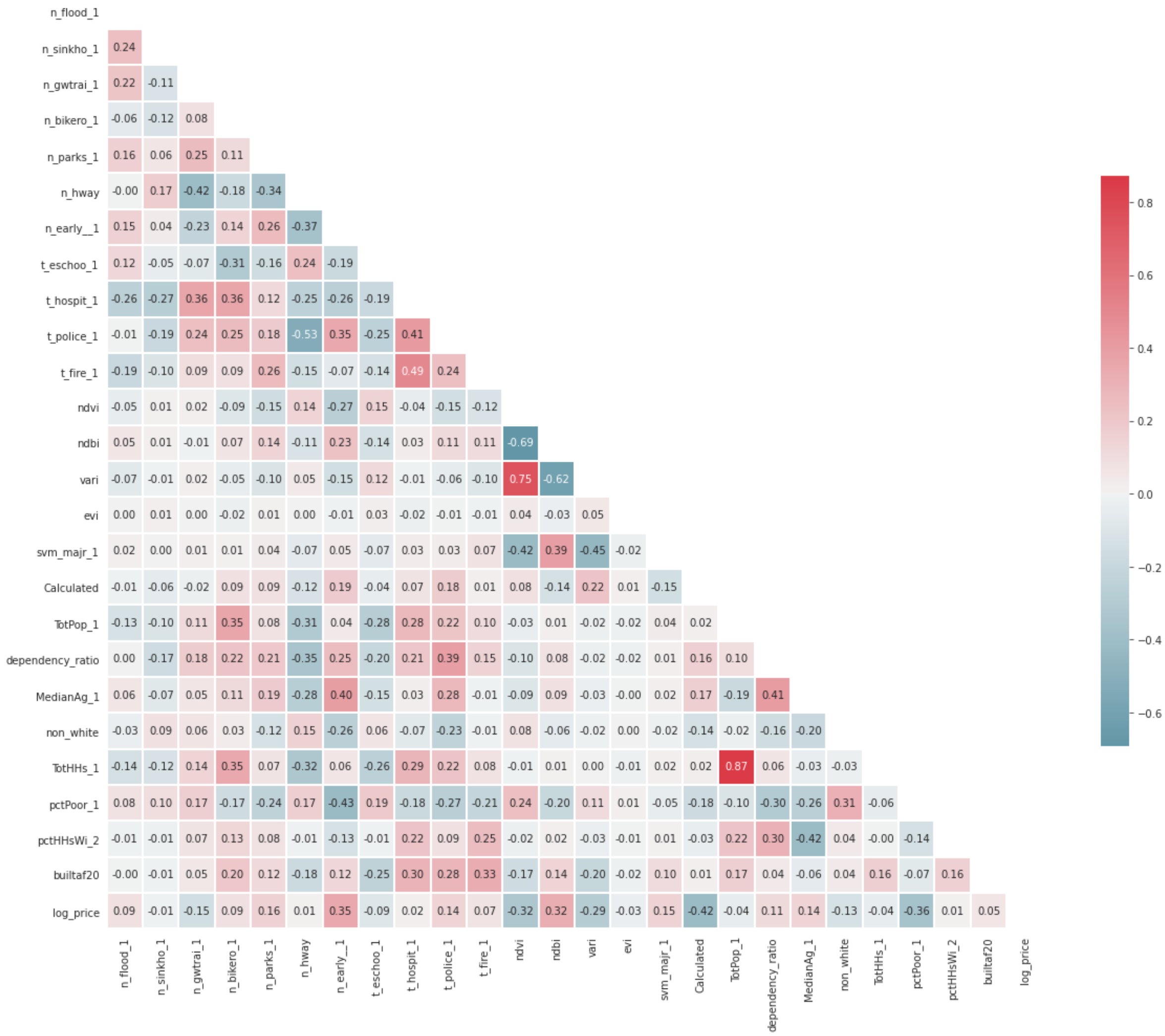


Figure 3. 13 Correlation plot of the data combined with socio-economic variables.

Meanwhile, the median age was retained, while the dependency age ratio was derived from the combination of the under-18 and over-65 age groups, representing a more efficient approach to analyzing their effects. Moreover, a new variable was generated by aggregating the populations of non-white racial groups, namely black, Indian, Asian, and Indian since the white population was found to constitute the majority (Figure 3.14) and was therefore less significant in determining the outcome variable. Lastly, the variables of total household income, the percentage of individuals living in poverty, and the percentage of households with children in the block group were preserved in the dataset and subsequently combined with the previous dataset comprising spatial features for the second analysis, with the objective of assessing their significance.

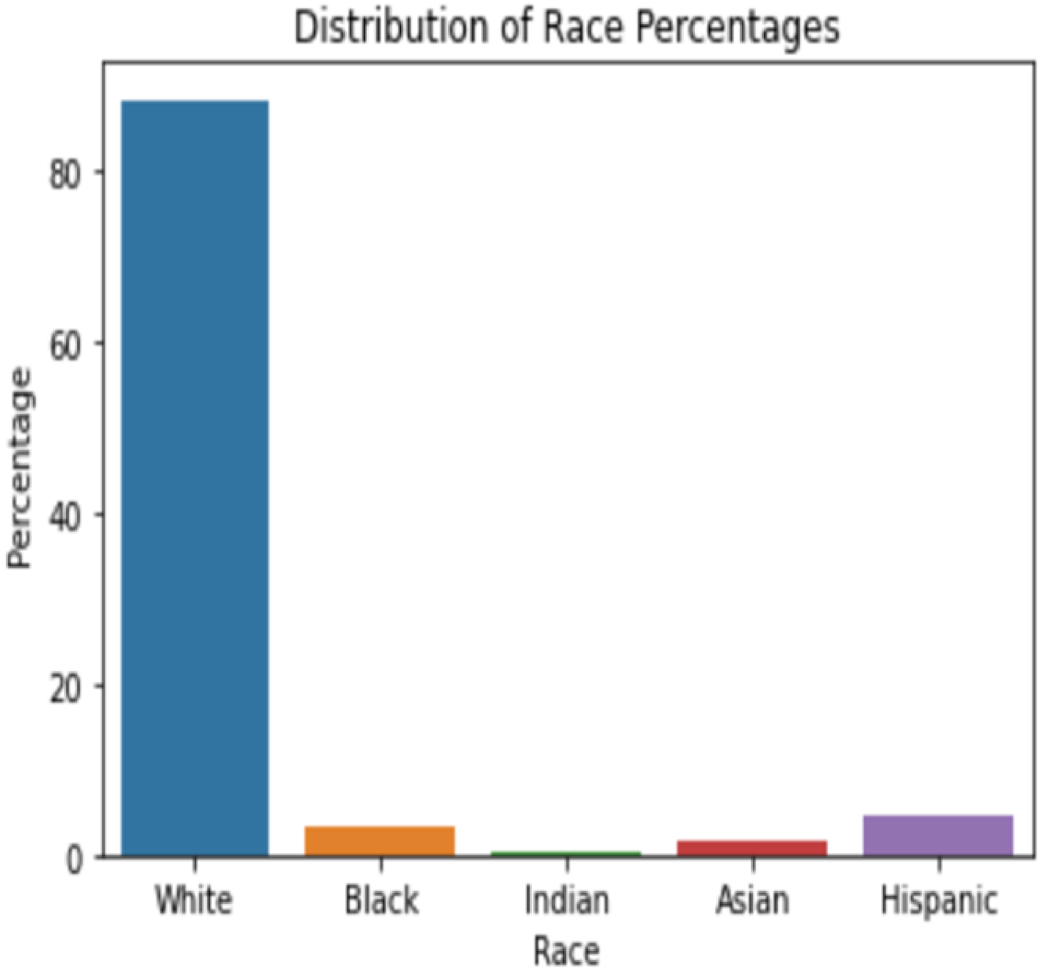


Figure 3. 14 Exploration bar plot representing the distribution of the race in America.

CHAPTER 4

RESULTS

This research engaged in each step of the machine learning workflow, including data preparation, feature engineering, hyperparameter tuning, model evaluation, and selection, and result interpretation (Figure 4.1).

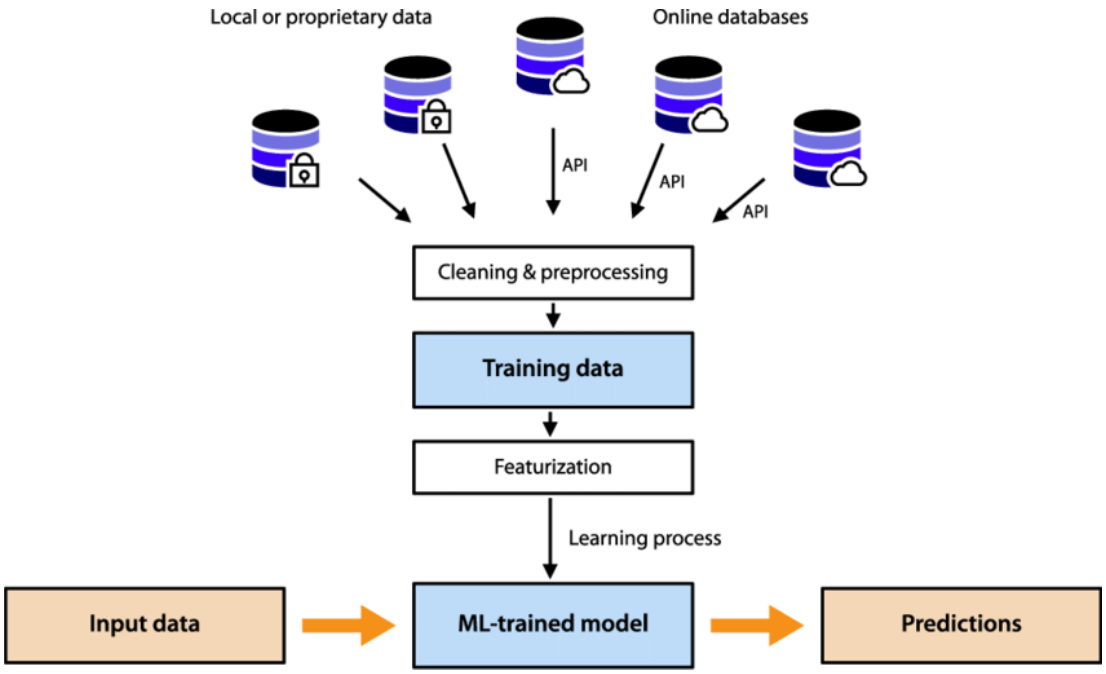


Figure 4. 1 Machine Learning Workflow used in this study. Taking from Chibani and Coudert, (2020).

In this study, predicting land value was treated as a regression task. To create models for this task, four algorithms were utilized: DT regression, RF regression, and GBM regression algorithm as well as linear regression. While linear regression is a traditional regression algorithm, the other algorithms are first commonly considered for classification. Nonetheless, in cases where the relationship between the target and independent variables is not linear and/or is complex, these algorithms can be effective alternatives to use as regression algorithms.

Classification trees divide the dataset based on homogeneity, whereas regression trees fit the target variable using all independent variables. In regression trees, the data of each independent variable is divided at several points, and the error between predicted and actual values is squared at each point to arrive at a Sum of Squared Errors (SSE). This SSE is then compared across all variables, and the point or variable with the lowest SSE becomes the split point, and the process continues recursively. The algorithm attempts to learn a mapping from the input features to the continuous target variable by constructing a decision tree-based model that approximates the underlying relationship between the features and the target. This model can then be used to predict the target variable for new data points. In summary, to use classification algorithms for regression problems, the target variable can be discretized and mapped to numerical values, allowing for the construction of a model that takes into account the relationship between the independent variables and the target variable. This resulting model can be utilized to make predictions in the form of numeric values for new instances.

Random forest and gradient boosting are both ensemble methods for machine learning that use decision trees. Random forest trains multiple decision trees on random subsets of the training data and features, and the model's output is the mean or median of the individual trees' predictions. In contrast, gradient boosting trains decision trees sequentially to correct the errors of the previous trees, and the model's output is the sum of the predictions from each tree.

Python programming language version 3.7.11 was employed, and the pandas (1.3.2) and numpy (1.19.5) libraries were utilized for data processing. For implementing the machine learning algorithms, the scikit-learn (0.23.2) package was employed with the following models: Linear Regression, DT, RF, and GBM. Graphs were created using the matplotlib (3.4.2) and seaborn (0.11.2) libraries. The model interpretation was carried out through the use of scikit-learn (0.23.2).

4.1 Preparing the Data & Feature Engineering:

Data for the study was sourced from various sources, with certain data derived through geographic information systems and feature engineering techniques. Exploratory analysis of the data was conducted beforehand to facilitate the preprocessing phase. This phase involved identifying outliers, cleaning the data, and carrying out other procedures as outlined in Section 3.2. Furthermore, the dependent variable, namely land values, was subjected to a natural logarithmic transformation. This transformation served to mitigate issues of heteroscedasticity and enhance the goodness-of-fit of the data, in accordance with previous studies (Guliker et al., 2022; Wei et al., 2022).

4.2 Hyperparameter Tuning Optimization and Model Selection

4.2.1 Training Set Up:

The dataset was divided into training (80%) and test (20%) sets to develop and evaluate machine learning models. The performance of the models was evaluated using several error metrics, including mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), and goodness-of-fit (Adjusted R-squared).

Unlike LR, other machine learning algorithms used in this study have model parameters that can be optimized. Each algorithm was trained using different combinations of error metrics and goodness-of-fit scores. Therefore, hyperparameter optimization was performed using cross-validation on the training set to improve the model's accuracy and minimize prediction errors (Figure 4.2). Non-overlapping k-fold cross-validation with 20 folds was used on the training set to assess the trained model's performance after adjusting the hyperparameters. During the cross-validation process, a subset of data was used to train the model, while another subset was used to estimate its performance. The process was repeated k times which is 20, with each fold excluded from training each time, and the average and standard deviation of the scores were used as metrics to identify the best hyperparameters and classify the algorithms according to their performance. To search for hyperparameters, three different algorithms were used: grid search, random search, and Bayesian search.

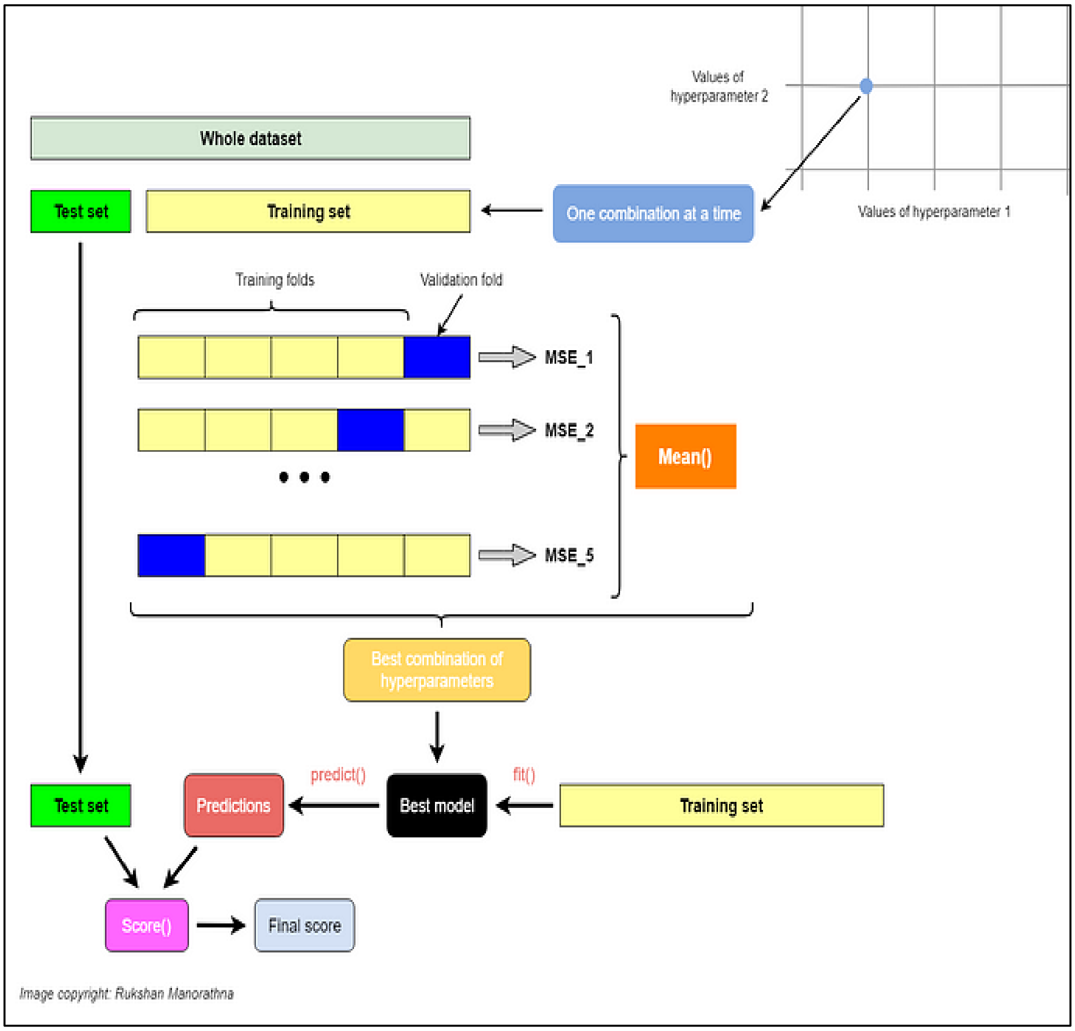


Image copyright: Rukshan Manorathna

Figure 4. 2 Workflows in the training, optimization, evaluation, and model selection phases. Taking from Rukshan Pramoditha, (2020).

4.2 Model Training and Optimization

To train the machine learning models, the dataset and training pipeline were initially prepared. Two separate datasets were implemented, one consisting of GIS-derived spatial distance context data, and the other combined with socioeconomic block group level data, as explained in Chapter 3.2. To optimize the performance of each algorithm, hyperparameters were tuned using grid search. The results for models are presented in Tables 4.1 and 4.2 show the best set of hyperparameters for each algorithm

Table 4. 1 Errors, goodness-of-fit measures, and tuned hyperparameters for the algorithms on distance measure datasets (The land value is expressed in dollars, but this value represents the error and standard deviation in terms of the logarithm of the land values for each quarter acre of land.)

Model	MAE	MSE	RMSE	Adj.-R^2	Std.	Tuned Hyperparameters
Linear Regression	0.2598	0.1131	0.3364	0.48	0.34	-
Decision Tree	0.824	0.0351	0.1873	0.84	0.19	max_depth:50 max_features:16 min_samples_leaf:5
Random Forest	0.0583	0.0155	0.1246	0.93	0.12	max _features:12 min_samples_split:12 n_estimators:400
Gradient Boosting	0.0753	0.0170	0.1305	0.92	0.13	learning_rate = 0.1 n_estimators = 1500 subsample = 1.0 max_depth = 4

Table 4. 2 Errors, goodness-of-fit measures, and tuned hyperparameters for the algorithms on distance measures and socioeconomic data set. (The land value is expressed in dollars, but this value represents the error and standard deviation in terms of the logarithm of the land values for each quarter acre of land.)

Model	MAE	MSE	RMSE	Adj.-R^2	Std.	Tuned Hyperparameters
Linear Regression	0.2483	0.1036	0.3218	0.52	0.32	-
Decision Tree	0.0718	0.0239	0.1546	0.89	0.16	max_depth:25 max_features:12 min_samples_leaf:5
Random Forest	0.0520	0.0128	0.1130	0.94	0.11	max _features:12 min_samples_split:12 n_estimators:400
Gradient Boosting	0.0665	0.0149	0.1221	0.93	0.12	learning_rate = 0.1 n_estimators = 1000 subsample = 1.0 max_depth=4

The tables present the performance metrics and hyperparameters of four different regression models on two datasets containing distance measures and socioeconomic variables for the second one. The lower the values of MAE, MSE, and RMSE, the better the model's predictive performance. The higher the value of the Adjusted R-squared, the better the model's explanatory power. A low standard deviation also indicates that the model's predictions are consistent across different parts of the data.

The results suggest that Random Forest has the best performance among the four algorithms, with the lowest MAE, MSE, RMSE, and standard deviation and the highest Adj.-R2. It is followed by Gradient Boosting, which has slightly higher error metrics but still performs well. The linear regression model has the poorest performance with the lowest Adj. R-squared value and highest Std. deviation. Figure 4.3 shows the comparison of the model's performance.

The hyperparameters for each algorithm are also listed, which can be useful for optimizing the performance of the algorithms on similar datasets. For example, the Random Forest algorithm has a max_features of 12, min_samples_split of 12, and n_estimators of 400, which can be used as a starting point for tuning the algorithm for similar problems.

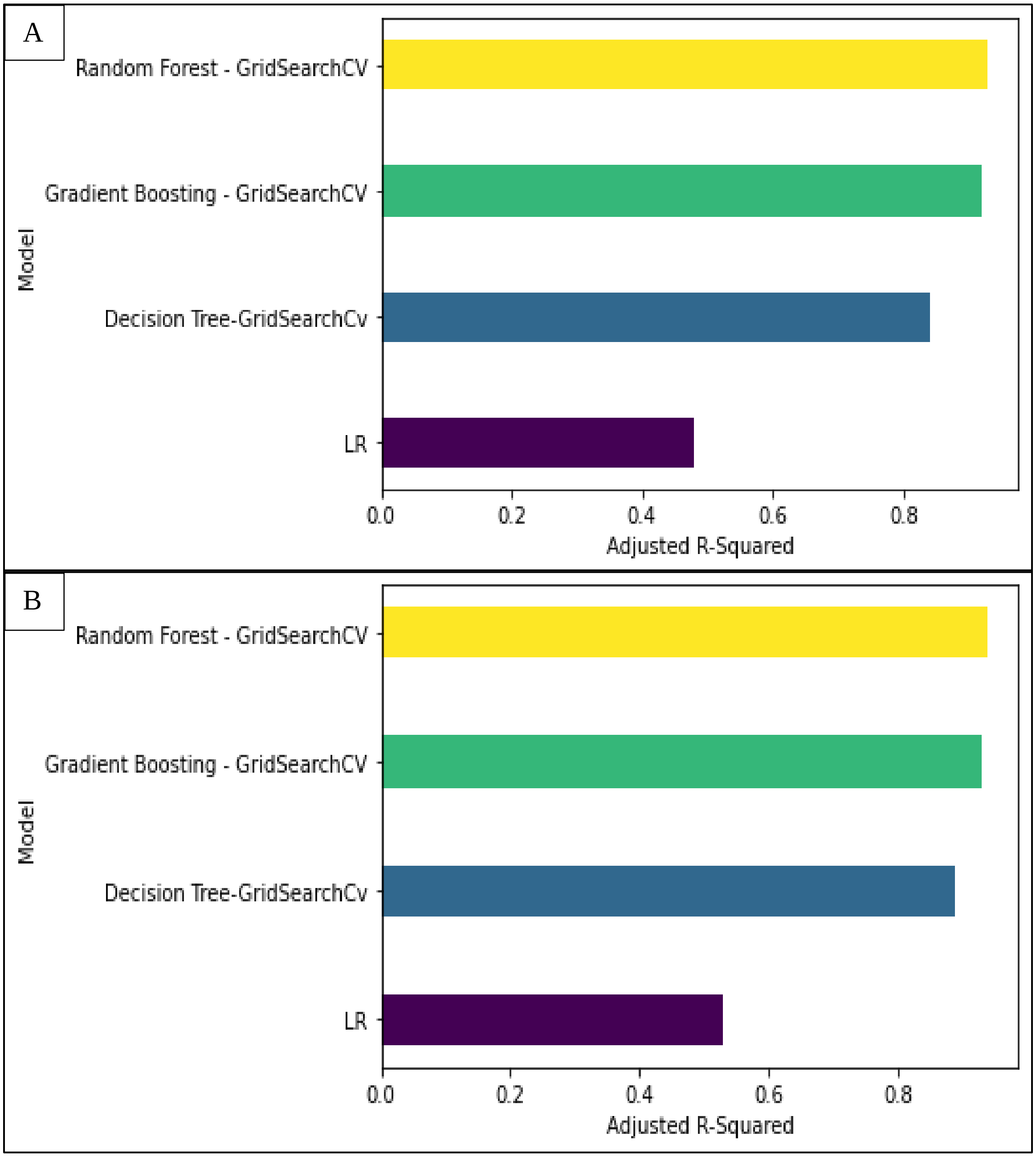


Figure 4. 3 Model performance comparison on training data for (a) distance measures dataset, (b) distance measures with the socioeconomic dataset.

4.3 Model Evaluation and Selection

During the model evaluation phase, we used the test datasets, algorithms, and hyper-parameters from the previous phase to analyze our model's performance. We employed performance metrics and visualization techniques to evaluate the prediction errors and select the best algorithm for predicting the dependent variable.

The plot illustrating the correlation between the actual and predicted values allows for an analysis of the residuals and overfitting, whereas the point cloud for the set appears to be more scattered. It can be observed from the results that the RF and GBR algorithms display a higher degree of overfitting in Figure (4.4).

To assess the ultimate performance of the models, they were retrained using the complete dataset including both the training and test sets for each two datasets, with the optimal hyperparameters obtained from the previous step. The resulting performances of the models are tabulated in Tables 4.3 and 4.4.

Table 4. 3 Results on distance measures (training + test, 100%). (The land valve is expressed in dollars, but this value represents the error and standard deviation in terms of the logarithm of the land values for each quarter acre of land.)

Model	**MAE**	**MSE**	**RMSE**	**Adj.-R^2**	**Std.**
Decision Tree	0.0311	0.0091	0.0955	0.95	0.21
Random Forest	0.0373	0.0029	0.0538	0.99	0.12
Gradient Boosting	0.0317	0.0048	0.0695	0.98	0.16

Table 4. 4 Results on distance measures and socioeconomic data set (training + test, 100%). (The land value is expressed in dollars, but this value represents the error and standard deviation in terms of the logarithm of the land values for each quarter acre of land.)

Model	**MAE**	**MSE**	**RMSE**	**Adj.-R^2**	**Std.**
Decision Tree	0.0425	0.0089	0.0940	0.96	0.21
Random Forest	0.0298	0.0046	0.0675	0.98	0.13
Gradient Boosting	0.0390	0.0035	0.0589	0.98	0.15

Based on the given complete dataset (training + test, 100%), several algorithms were trained and evaluated based on their performance in terms of MAE, MSE, RMSE, and Adjusted-R^2. To select the most efficient algorithm to predict land prices in the given dataset, various aspects should be considered. First, the difference in performance between the algorithms should be evaluated, varying between 0.84 and 0.98 (Adj. R-squared score). For both datasets, RF has the best performance with 0.99 and 0.98 Adj. R-squared respectively and the lowest standard deviation, while others also show better performance on the complete dataset.

Second, the algorithm with the lowest prediction variability in the cross-validation process should be chosen, and in this case, the RF algorithm has the lowest variability, and standard deviation and error metrics are the lowest.

It's worth noting that the DT model also performed well in both datasets. However, RF is generally preferred over Decision Trees because it's more resistant to overfitting and can provide better generalization performance. Overall, based on the given results, the Random Forest model with hyperparameter optimization using Grid search is the best model for the distance measure dataset and is the best model for the distance measures with the socioeconomic dataset.

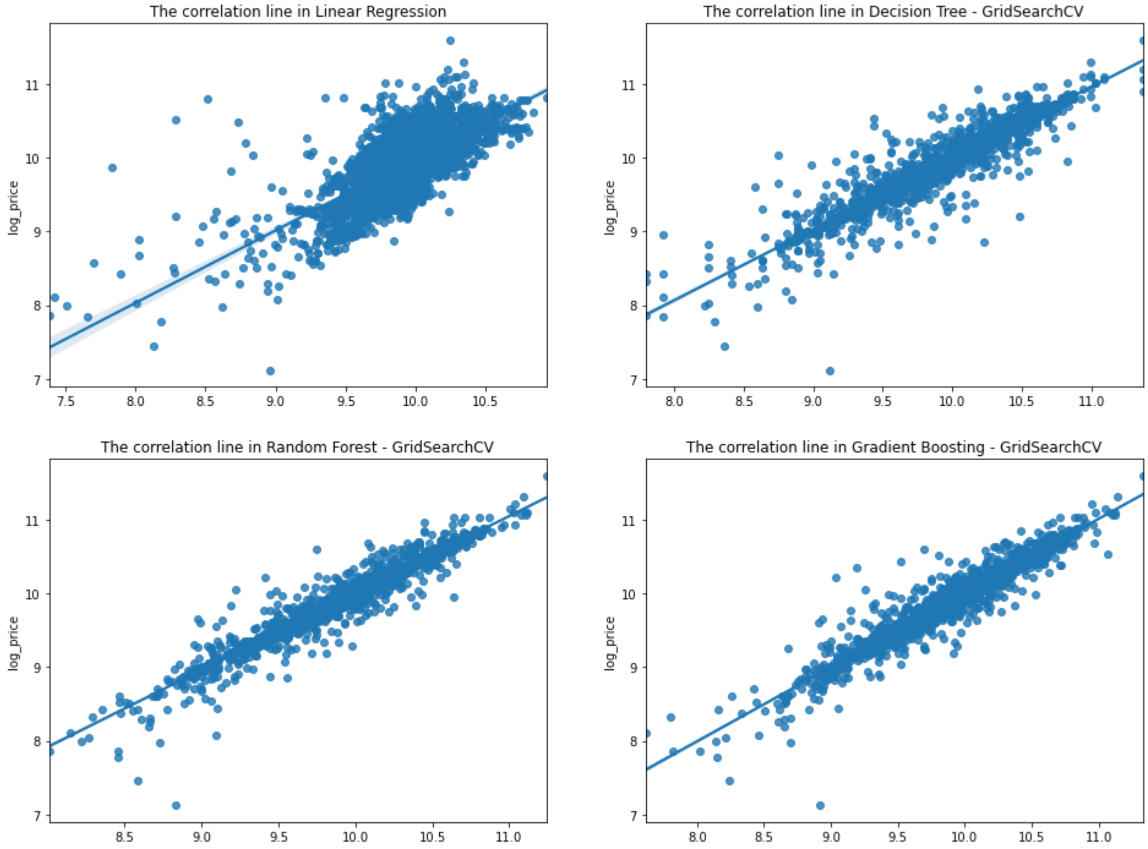


Figure 4. 4 Plot for correlation between the predicted and the truth as measured by the provided Land Valuation numbers from Greene County Assessor's Office.

4.4 Model Interpretation

In the model interpretation phase, we identified the most important features after we trained the model with all available data and extracted performance metrics for the final deployment.

Most machine learning algorithms can assess the relative importance of input features and determine which ones are more significant in predicting the outcome variable. Figure 4.5 illustrates the significant factors that have a greater impact on the forecasting of land value when utilizing two algorithms, namely RF and GBR, with distance measure datasets, while Figure 4.6 shows the significant factors on the prediction using RF and GBR algorithms for the distance measures with socioeconomic datasets.

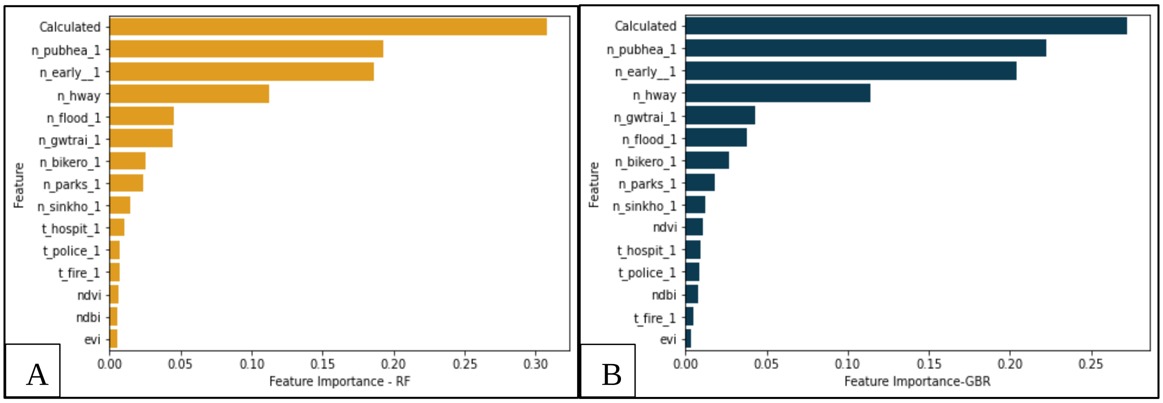


Figure 4. 5 Relative importance of the most relevant features according to the (a) Random Forest (RF); and (b) the Gradient Boosting Regressor algorithm (GBR) using distance measures dataset.

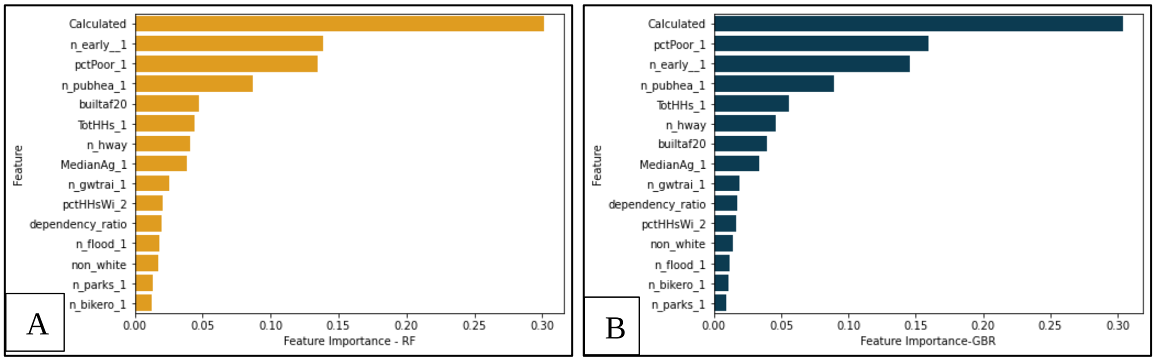


Figure 4. 6 Relative importance of the most relevant features according to the (a) Random Forest (RF); and (b) the Gradient Boosting Regressor algorithm (GBR) using the combined distance measures and socioeconomic dataset.

The calculated area is the most significant feature of both algorithms. Besides, the most important features of the land are the distance to public health and early childhood school, the total population in block groups, and the distance to the highway (n_pubhea_1, n_early_1, TotPop_1, n_hway). Regarding distance to public health and early childhood school, it can be said that they are important factors because they are rare in the city center as seen in Figure 4.7.

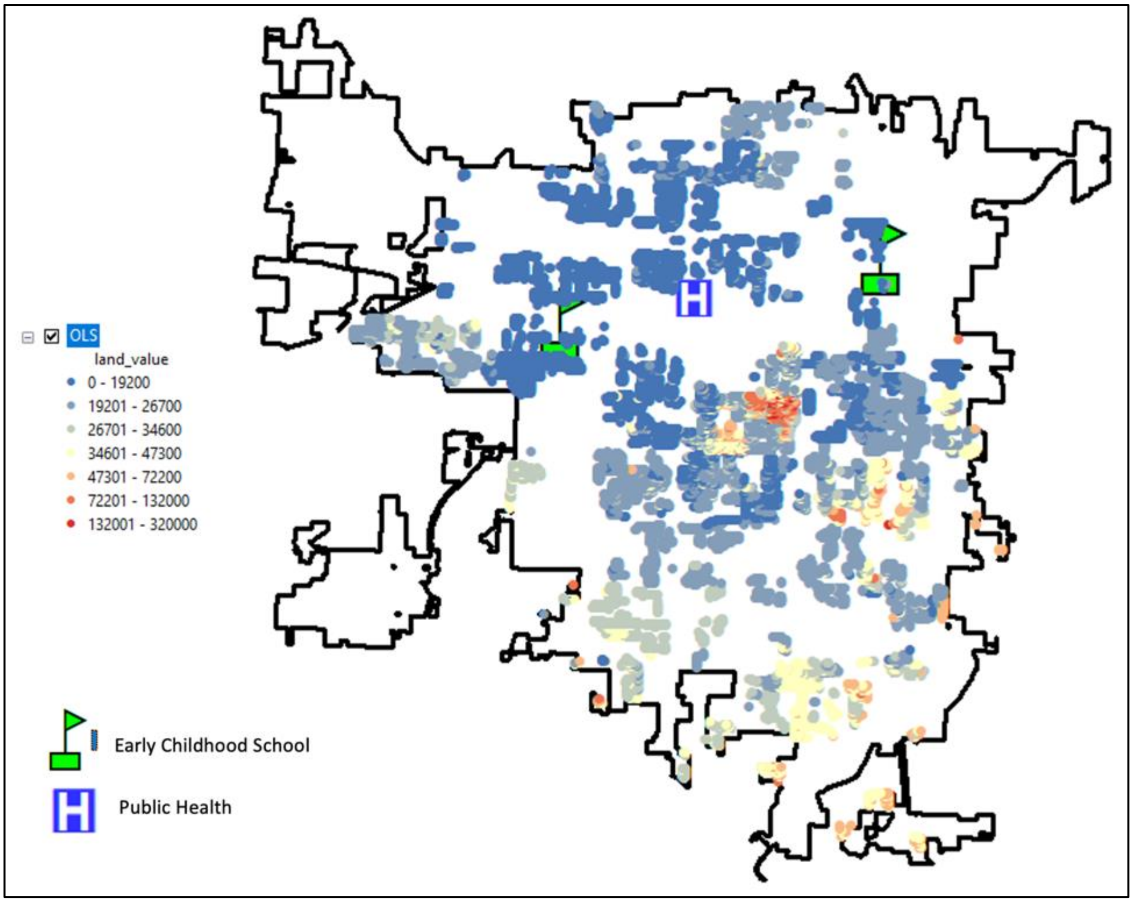


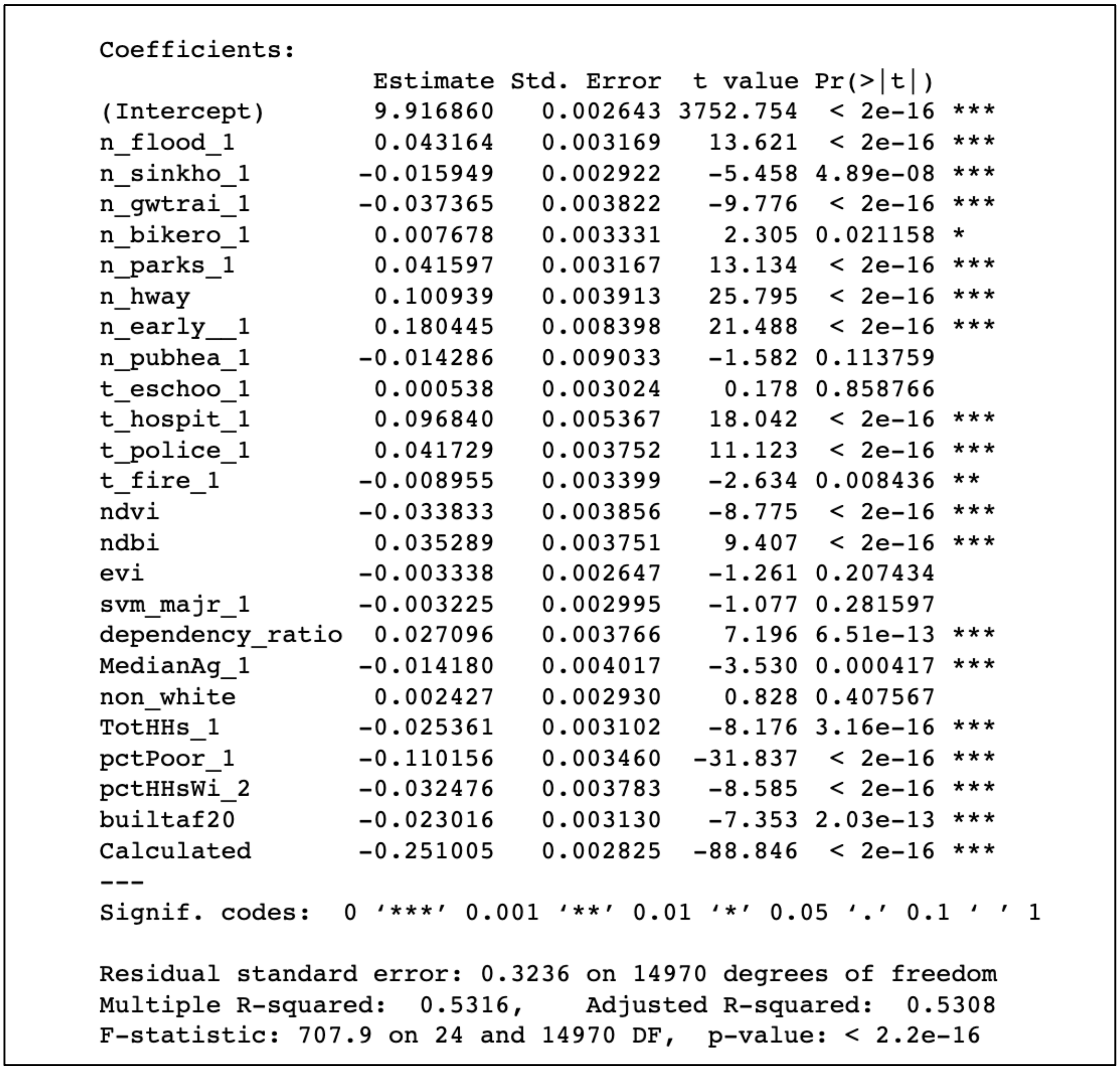
Figure 4. 7 Land value distribution plot located with a public health center and two early childhood schools.

In the distance measure with the socioeconomic dataset, the calculated area is the most significant feature and early childhood school is included again. The percentage of the poor population is also demonstrated as important feature (pctPoor_1).

Moreover, linear regression also gives the important feature result. As seen in Figure 4.8, the linear regression model shows an adjusted R-squared value of 0.4867, indicating that approximately 48.67% of the variation in price is explained by the model. The F-statistic of 837.3 with a p-value less than 2.2e-16 indicates that the model is statistically significant.

The coefficients represent the relationship between each independent variable and the dependent variable, with p-values indicating the statistical significance of each coefficient. Variables with p-values less than 0.05 are considered statistically significant. In this model, we see that distance to flood, sinkhole, groundwater train, parks, highway, early childcare, hospital, police, fire station, ndvi, ndbi, svm_majr_1, and Calculated variable all have statistically significant relationships with log_price, while the distance to bike route, public health, and school do not.

It is important to note that the importance of features may vary across different models. While distance to public health and early childhood school appear to be important features in RF and GBM, they may not be as important in a linear regression model.

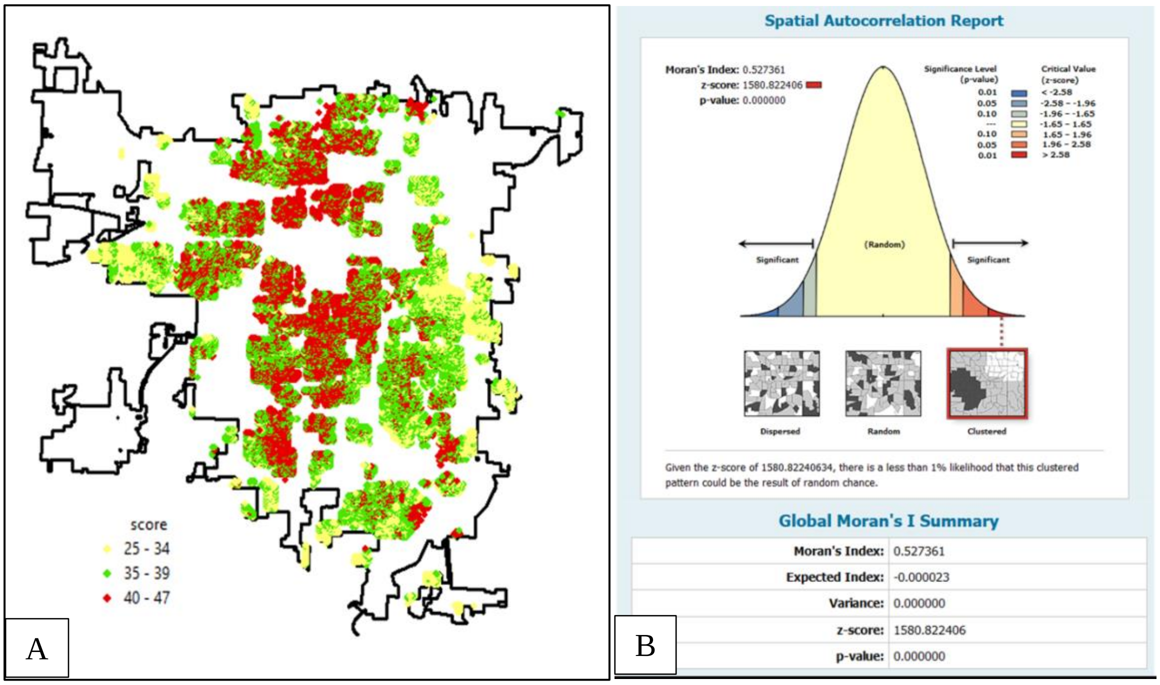


```
Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)    
(Intercept)        9.916860   0.002643 3752.754  < 2e-16 ***
n_flood_1          0.043164   0.003169   13.621  < 2e-16 ***
n_sinkho_1        -0.015949   0.002922   -5.458 4.89e-08 ***
n_gwtrai_1        -0.037365   0.003822   -9.776  < 2e-16 ***
n_bikero_1         0.007678   0.003331    2.305 0.021158 *  
n_parks_1          0.041597   0.003167   13.134  < 2e-16 ***
n_hway             0.100939   0.003913   25.795  < 2e-16 ***
n_early__1         0.180445   0.008398   21.488  < 2e-16 ***
n_pubhea_1        -0.014286   0.009033   -1.582 0.113759    
t_eschoo_1         0.000538   0.003024    0.178 0.858766    
t_hospit_1         0.096840   0.005367   18.042  < 2e-16 ***
t_police_1         0.041729   0.003752   11.123  < 2e-16 ***
t_fire_1          -0.008955   0.003399   -2.634 0.008436 ** 
ndvi              -0.033833   0.003856   -8.775  < 2e-16 ***
ndbi               0.035289   0.003751    9.407  < 2e-16 ***
evi               -0.003338   0.002647   -1.261 0.207434    
svm_majr_1        -0.003225   0.002995   -1.077 0.281597    
dependency_ratio   0.027096   0.003766    7.196 6.51e-13 ***
MedianAg_1        -0.014180   0.004017   -3.530 0.000417 ***
non_white          0.002427   0.002930    0.828 0.407567    
TotHHs_1          -0.025361   0.003102   -8.176 3.16e-16 ***
pctPoor_1         -0.110156   0.003460  -31.837  < 2e-16 ***
pctHHsWi_2        -0.032476   0.003783   -8.585  < 2e-16 ***
builtaf20         -0.023016   0.003130   -7.353 2.03e-13 ***
Calculated        -0.251005   0.002825  -88.846  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3236 on 14970 degrees of freedom
Multiple R-squared:  0.5316,    Adjusted R-squared:  0.5308 
F-statistic: 707.9 on 24 and 14970 DF,  p-value: < 2.2e-16
```

Figure 4. 8 Summary of linear regression models to show p-values and important features of the model.

After obtaining results on the performance of machine learning models and the significance of relative features, we proceeded to conduct a spatial analysis in the second part of the study. The dataset used in Chapter 3 was coded to reflect the degree of features based on location. We assigned a score to each point based on these coded values. Figure 4.8 displays the distribution of these scores.



Spatial Autocorrelation Report

Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	▨	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of 1580.82240634, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Global Moran's I Summary

Moran's Index:	0.527361
Expected Index:	-0.000023
Variance:	0.000000
z-score:	1580.822406
p-value:	0.000000

Figure 4. 9 Analyzing features using coded values called score on (a) distribution map; (b) spatial autocorrelation report.

Spatial autocorrelation refers to the degree to which nearby observations in a dataset are similar to each other. The report in Figure 4.8 provides an analysis of spatial autocorrelation using Moran's Index for score variables, which measures the spatial clustering or dispersion of a variable in a geographic area. Moran's Index value of 0.53 indicates a positive spatial autocorrelation, meaning that similar values of the variable tend to be clustered together in space. The z-score is a measure of how many standard deviations the observed Moran's Index is from the expected Moran's Index under the null hypothesis of spatial randomness. Also, the p-value is the probability of obtaining a Moran's Index as extreme as the one observed, assuming the null hypothesis of spatial randomness. The z-score of 1580.82 and the p-value of 0.00 suggest that the observed spatial pattern is significant and not the result of random chance. Therefore, we can conclude that the observed pattern is unlikely to be due to random variation, and there is a strong spatial dependence among the score variables.

Moreover, spatial autocorrelation is also calculated for land prices. In this case, Moran's Index is 0.59, which indicates that there is a positive spatial autocorrelation, meaning that similar values tend to cluster together. The z-score is 425.56, which is very large, indicating a very low probability that the observed pattern could be due to chance, while the p-value is 0.00, which is less than 0.01, indicating strong evidence against the null hypothesis and in favor of spatial autocorrelation.

CHAPTER 5

DISCUSSION

In this study, the performance of various machine learning algorithms was compared to predict land values using two separate datasets, one incorporating distance measures context and the other including both distance measures and socioeconomic data. We found that all of the machine-learning algorithms outperformed the linear model. Additionally, the Random Forest and Gradient Boosting Algorithm performed better than the Decision Tree algorithm.

RF and GBM, outperformed the traditional linear model in predicting land values. We attribute this success to the ensemble learning techniques used by these algorithms, which combine multiple weak models to produce a strong model that can handle complex relationships between input variables and the output variable. Furthermore, RF and GBM can address the issue of overfitting, which is a common problem with DT, thus enhancing the accuracy and generalizability of the model. Notably, these algorithms are capable of capturing non-linear relationships between input variables and the output variable, which is crucial when predicting land values that may be influenced by a complex set of factors. In summary, our study highlights the efficacy of RF and GBM in predicting land values and demonstrates the potential for machine learning techniques to improve the accuracy and reliability of land valuation models.

One such factor is the choice of hyperparameters used in the model, such as the number of trees in the Random Forest algorithm or the learning rate in the Gradient Boosting Algorithm. Fine-tuning these hyperparameters can lead to better model performance and should be considered in future research.

Furthermore, the choice of features included in the dataset can also impact the performance of the machine learning algorithms. While this study focused on distance measures and socioeconomic data, other variables such as crime rates, air quality, and proximity to public transportation could also be considered as potential predictors of land values. It is important to carefully select relevant features to include in the model as adding irrelevant or redundant features can decrease the model's accuracy and interpretability.

In addition, it is important to note that the performance of machine learning algorithms in predicting land values may vary depending on the size and complexity of the dataset. This study focused on a specific urban area with relatively simple geographic and socioeconomic features. It would be interesting to apply similar machine-learning approaches to larger and more diverse urban areas to determine if the same algorithms and features hold true.

This study revealed that incorporating census data, in this case, did not yield a big improvement in model performance compared to using only the distance measures context data. This suggests that location remains a crucial determinant of land prices and that spatial information is the most significant variable influencing land value. Our findings are consistent with previous research that has highlighted the importance of considering spatial factors in land valuation.

Although it was anticipated that the density of green areas would have an impact on land value, our analysis using a dataset derived from remote sensing to create a vegetation index and land cover information showed that these variables were not significant in determining land value. However, this does not necessarily mean that green areas are not a factor in determining land value. The reason for this might be that the resolution of the image used in this study was limited to 10 m, which might not have been sufficient to capture the relevant information.

Another reason might be the using only mean values for the vegetation indexes. One way to address this issue would be to include additional statistics such as median, standard deviation, and range for vegetation indexes in the dataset. This can provide a more comprehensive understanding of the green areas in each neighborhood, allowing the model to better capture the relationship between green areas and housing prices. Additionally, it might be useful to investigate the relationship between different statistics of vegetation indexes and housing prices to determine which statistics are most informative for the given problem.

This study investigated further revealed a spatial correlation in land value and in the scored dataset, which was a coded dataset reflecting the degree of features based on location. This correlation suggests that spatially proximate features are interrelated, and the same holds true for land value. As a result, we can conclude that land value is highly dependent on location and spatial features. These findings highlight the importance of considering spatial factors in land valuation to account for the underlying interdependencies among features that can influence land value.

Based on the results of our modeling approach, we did not find significant inequities in land values across different areas within the city. However, it is important to note that our study was conducted in a specific city with a relatively homogeneous population and socioeconomic status. Therefore, it would be interesting to apply our approach in a larger and more diverse city where there may be greater disparities in land values across different areas.

By studying a larger and more diverse city, we may be able to identify potential disparities in land values that are not present in our current study. For example, certain neighborhoods or areas within the city may have historically faced disinvestment or neglect, leading to lower land values compared to other areas. Alternatively, certain areas may be experiencing gentrification, resulting in rapidly increasing land values that could lead to displacement and further exacerbate existing inequities.

Therefore, further research is needed to fully understand the extent of inequities in land values across different areas and to identify potential factors that may be driving these disparities. Our study provides a useful starting point for investigating these issues, and we believe that our approach can be applied in other cities to identify and address potential inequities in land values.

CHAPTER 6

CONCLUSIONS

This thesis has proposed a novel approach for forecasting land valuation that integrates distance-measured context data into machine learning across the City of Springfield, Missouri as a case study. Our results suggest that incorporating distance measure context data is critical for accurate land value prediction and that census data may not significantly improve model performance. We found that all of the machine learning algorithms outperformed the linear model with Random Forest and Gradient Boosting Algorithm performing best.

The results of the spatial autocorrelation analysis highlight the importance of considering location and spatial features in the valuation of land. The high degree of spatial autocorrelation in land values suggests that the value of a given parcel of land is strongly influenced by the values of surrounding parcels, indicating that a spatially aware approach to land valuation is crucial. Accounting for spatial dependence and incorporating spatial features in the valuation process can lead to more accurate and precise estimates of land value.

Furthermore, the proposed geodatabase provides a comprehensive framework for collecting and integrating various data types, including proximity and accessibility to key locations, socio-economic variables, green space density, and vegetation index. However, the use of Sentinel-2 satellite data for vegetation index performed inadequately for predicting land values.

Overall, our study provides a foundation for future research on land valuation that incorporates distance measure context and machine learning algorithms. This approach can enable fast, accurate, and unbiased land valuation, which is crucial for urbanization, taxation, insurance, mortgage lending, and other applications. In particular, our approach can help to reduce inequities in land valuation, thereby promoting social justice and economic growth.

Future research can explore additional features that can influence land value, such as the proximity to public transportation, and the quality of schools. Moreover, research can investigate the application of our approach in other urban areas with better-resolution raster data. Additionally, we can also consider using other features that may affect the land prices such as the crime rate, and the quality of the neighborhood. It is important to carefully select the relevant features to include in the model as adding irrelevant or redundant features can actually decrease the model's accuracy and interpretability.

1

REFERENCES

Aladwan, Z., & Ahamad, M. S. S. (2019). Hedonic Pricing Model for Real Property Valuation via GIS - A Review. *Civil and Environmental Engineering Reports*, *29*(3), 34–47. https://doi.org/10.2478/ceer-2019-0022

Antipov, E., & Pokryshevskaya, E. B. (2012). Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics. *Expert Systems with Applications*, *39*, 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. *Applied Sciences (Switzerland)*, *8*(11). https://doi.org/10.3390/app8112321

Banerjee, D., & Dutta, S. (2017). Predicting the housing price direction using machine learning techniques. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2998–3000.

Bency, A., Rallapalli, S., Ganti, R., Srivatsa, M., & Manjunath, B. (2017). *Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery*. 320–329. https://doi.org/10.1109/WACV.2017.42

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brimble, P., Mcsharry, P., Bachofer, F., Bower, J., & Braun, A. (2020). *Using Machine Learning And Remote Sensing To Value Property In Rwanda*.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (1999). Some Notes on Parametric Significance Tests for Geographically Weighted Regression. *Journal of Regional Science*, *39*, 497–524.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information*, *7*(5). https://doi.org/10.3390/ijgi7050168

Chen, H., J., Ong, C.-F., Zheng, L., & Hsu, S.-C. (2017). Forecasting Spatial Dynamics of the Housing Market Using Support Vector Machine. *International Journal of Strategic Property Management*, *21*, 273–283. https://doi.org/10.3846/1648715X.2016.1259190

Chibani, S., & Coudert, F.-X. (2020). Machine learning approaches for the prediction of materials properties. *APL Materials*, *8*, 80701. https://doi.org/10.1063/5.0018384

Choi, S., & Yi, M. (2021). Computational Valuation Model of Housing Price Using Pseudo Self Comparison Method. *Sustainability, 13*, 11489. https://doi.org/10.3390/su132011489

Daradi, S., Yusof, U., & Ab Kader, N. I. (2018). Prediction of Housing Price Index in Malaysia Using Optimized Artificial Neural Network. *Advanced Science Letters*, *24*, 1307–1311. https://doi.org/10.1166/asl.2018.10738

Dellstad, M. (2018). *Comparing Three Machine Learning Algorithms in the Task of Appraising Commercial Real Estate*.

Diaz, J., & Hansz, J. A. (1997). How Valuers Use The Value Opinions Of Others. In *Journal of Property Valuation and Investment* (Vol. 15, Issue 3, pp. 256–260). https://doi.org/10.1108/14635789710184970

Droj, G., Droj, L., & Mancia, A. (2010). *Nominal Assets Valuation by GIS*.

Embaye, W. T., Zereyesus, Y. A., & Chen, B. (2021). Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches. *PLOS ONE, 16*(2), 1–20. https://doi.org/10.1371/journal.pone.0244953

Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A Decision Tree Approach. *Urban Studies*, *43*(12), 2301–2315. https://doi.org/10.1080/00420980600990928

Fotheringham, A., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. *John Wiley & Sons*, *13*.

Gitelson, A., Kaufman, Y., Stark, R., & Rundquist, D. (2002). Novel Algorithms for Remote Estimation of Vegetation Fraction. *Remote Sensing of Environment*, *80*, 76–87. https://doi.org/10.1016/S0034-4257(01)00289-9

Goodman, A. C. (1998). Hedonic Prices, Price Indices and Housing Markets. *Journal of Urban Economics*, *44*(2), 291–298. https://doi.org/https://doi.org/10.1006/juec.1997.2071

Gu, G., & Xu, B. (2017). Housing Market Hedonic Price Study Based on Boosting Regression Tree. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *21*(6), 1040–1047. https://doi.org/10.20965/jaciii.2017.p1040

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, *51*(5). https://doi.org/10.1145/3236009

Guliker, E., Folmer, E., & van Sinderen, M. (2022). Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach. *ISPRS International Journal of Geo-Information*, *11*(2). https://doi.org/10.3390/ijgi11020125

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*.

Helbich, M., & Kuntz, M. (2014). Geostatistical Mapping of Real Estate Prices: An Empirical Comparison of Kriging and Cokriging. *International Journal of Geographical Information Science*, *28*. https://doi.org/10.1080/13658816.2014.906041

Ho, W., Tang, B.-S., & Wong, S. (2020). Predicting Property Prices with Machine Learning Algorithms. *Journal of Property Research*, *38*, 1–23. https://doi.org/10.1080/09599916.2020.1832558

Hong, J., Choi, H., & Kim, W. S. (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. https://doi.org/10.3846/ijspm.2020.11544

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (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. https://doi.org/https://doi.org/10.1016/j.landusepol.2018.12.030

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1), 195–213. https://doi.org/https://doi.org/10.1016/S0034-4257(02)00096-2

Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). Machine Learning for Property Price Prediction and Price Valuation: A Systematic Literature Review. *Planning Malaysia*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Tree-Based Methods. In *An Introduction to Statistical Learning: with Applications in R* (pp. 303–335). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7_8

Jog, S., & Dixit, M. (2016). Supervised classification of satellite images. *Conference on Advances in Signal Processing, CASP 2016*, 93–98. https://doi.org/10.1109/CASP.2016.7746144

Kagie, M., & Wezel, M. (2007). Hedonic Price Models and Indices Based on Boosting Applied to The Dutch Housing Market. *Erasmus University Rotterdam*, *Econometric Institute, Econometric Institute Report*, *15*. https://doi.org/10.1002/isaf.287

Kauko, T. (2003). On Current Neural Network Applications Involving Spatial Modelling of Property Prices. *Journal of Housing and the Built Environment*, *18*, 159–181. https://doi.org/10.1023/A:1023977111302

Li, A. H., & Bradic, J. (2018). Boosting in the Presence of Outliers: Adaptive Classification With Nonconvex Loss Functions. *Journal of the American Statistical Association*, *113*(522), 660–674. https://doi.org/10.1080/01621459.2016.1273116

Liebelt, V., Bartke, S., & Schwarz, N. (2018). Hedonic pricing analysis of the influence of urban green spaces onto residential prices: the case of Leipzig, Germany. *European Planning Studies*, *26*, 133–157.

Lim, W., Wang, L., Wang, Y. L., & Chang, Q. (2016). *Housing Price Prediction Using Neural Networks*. 518–522. https://doi.org/10.1109/FSKD.2016.7603227

Limsombunchai, V., Gan, C., & Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences*, *1*. https://doi.org/10.3844/ajassp.2004.193.201

Masías, V. H., Valle, M., Crespo, F., Crespo, R., Vargas Schüler, A., & Laengle, S. (2016). *Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile.*

Mei, Y., Zhao, X., Lin, L., & Gao, L. (2018). Capitalization of Urban Green Vegetation in a Housing Market with Poor Environmental Quality: Evidence from Beijing. *Journal of Urban Planning and Development*, *144*. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000458

Mora-Garcia, R. T., Cespedes-Lopez, M. F., & Perez-Sanchez, V. R. (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. *Land*, *11*(11). https://doi.org/10.3390/land11112100

Neloy, A., Haque, H., & Islam, M. (2019). *Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring*. 350–356. https://doi.org/10.1145/3318299.3318377

Nguyen, N., & Cripps, A. (2001). Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *Journal of Real Estate Research*, *22*, 313–336.

Osland, L., Östh, J., & Nordvik, V. (2022). House Price Valuation of Environmental Amenities: An Application of GIS-Derived Data. *Regional Science Policy and Practice*, *14*(4), 939–959. https://doi.org/10.1111/rsp3.12382

Ottomanelli, M., Chiarazzo, V., Marinelli, M., & Caggiani, L. (2014). A Neural Network based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location. *Transportation Research Procedia*, *3*, 810–817. https://doi.org/10.1016/j.trpro.2014.10.067

Pai, P. F., & Wang, W. C. (2020). Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. *Applied Sciences (Switzerland)*, *10*(17). https://doi.org/10.3390/app10175832

Peterson, S., & Flanagan, A. (2009). Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. *Journal of Real Estate Research*, *31*(2), 147–164. https://doi.org/10.1080/10835547.2009.12091245

Pinter, G., Mosavi, A., & Felde, I. (2020). Artificial Intelligence for Modeling Real Estate Price Using Call Detail Records and Hybrid Machine Learning Approach. *Entropy*, *22*(12), 1–14. https://doi.org/10.3390/e22121421

Reed, R. (2016). The Relationship Between House Prices and Demographic Variables: An Australian Case Study. *International Journal of Housing Markets and Analysis*, *9*, 520–537. https://doi.org/10.1108/IJHMA-02-2016-0013

Rukshan Pramoditha. (2020, December 18). *k-fold cross-validation explained in plain English*. For Evaluating a Model’s Performance and Hyperparameter Tuning.

Selim, H. (2009). Determinants of House Prices in Turkey: Hedonic Regression versus Artificial Neural Network. *Expert Syst. Appl.*, *36*, 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044

Simon, A., Deo, M., Selvam, V., & Babu, R. (2016). An Overview of Machine Learning and Its Applications. *International Journal of Electrical Sciences & Engineering*, *Volume*, 22–24.

Song, Y. Y., & Lu, Y. (2015). Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Soni, P. K., Rajpal, N., Mehta, R., & Mishra, V. K. (2021). Urban land cover and land use classification using multispectral sentinal-2 imagery. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-021-10991-0

STC. (2017). *State Tax Commission of Missouri Assessor Manual: Specialty Property Guidelines*. https://stc.mo.gov/wp-content/uploads/sites/5/2017/12/Chapter-6-with-Supplement-2017.pdf

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*(sup1), 234–240. https://doi.org/10.2307/143141

Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2012). GIS-Based Mass Appraisal Model for Equity And Uniformity Of Rating Assessment. In *International Journal of Real Estate Studies* (Vol. 7, Issue 2).

Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2022). The Research Development of Hedonic Price Model-Based Real Estate Appraisal in the Era of Big Data. In *Land* (Vol. 11, Issue 3). MDPI. https://doi.org/10.3390/land11030334

Wen, H., Lu, J., & Fu, Y. (2005). Product differentiation and hedonic prices: an empirical analysis. *Proceedings of ICSSSM '05. 2005 International Conference on Services Systems and Services Management, 2005.*, *2*, 1259-1263 Vol. 2. https://doi.org/10.1109/ICSSSM.2005.1500199

Wen, H., Zhang, Z.-L., & Zhang, L. (2011). An Empirical Analysis on Spatial Effects of The Housing Price Based on Spatial Econometric Models: Evidence From Hangzhou City. *Xitong Gongcheng Lilun Yu Shijian/System Engineering Theory and Practice*, *31*, 1661–1667.

Xie, X., & Gang, H. (2007). *A Comparison of Shanghai Housing Price Index Forecasting*. *3*, 221–225. https://doi.org/10.1109/ICNC.2007.14

Yacim, J. A., & Boshoff, D. G. B. (2014). Mass Appraisal of Properties Appropriateness of Models. *Virual Multidisciplnary Conference QUAESTI*, 182–193.

Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy*, *99*, 104889. https://doi.org/10.1016/j.landusepol.2020.104889

Yu, D., Wei, Y., & Wu, C. (2007). Modeling Spatial Dimensions of Housing Prices in Milwaukee, WI. *Environment and Planning B: Planning and Design*, *34*, 1085–1102. https://doi.org/10.1068/b32119

Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing - INT J REMOTE SENS*, *24*, 583–594. https://doi.org/10.1080/01431160304987

Zhou, G., Ji, Y., Chen, X., & Zhang, F. (2018). Artificial Neural Networks and the Mass Appraisal of Real Estate. *International Journal of Online Engineering (IJOE)*, *14*, 180. https://doi.org/10.3991/ijoe.v14i03.8420

Zhou, J., Zhang, H., & Gu, Y. (2018). Affordable Levels of House Prices Using Fuzzy Linear Regression Analysis: The Case of Shanghai. *Soft Computing*, *22*. https://doi.org/10.1007/s00500-018-3090-4

LAND ASSESSMENT USING AN INNOVATIVE MODEL COMBINING MACHINE LEARNING AND SPATIAL CONTEXT.

by

Feride Tanrikulu

University of Missouri